OXFORD

Briefings in Bioinformatics, 2025, **26(2)**, bbaf136
https://doi.org/10.1093/bib/bbaf136
Review

Deep learning in single-cell and spatial transcriptomics data analysis: advances and challenges from a data science perspective

Shuang Ge[1,2], Shuqing Sun[1], Huan Xu[3], Qiang Cheng[4,5,*], Zhixiang Ren[2,*]

[1]Shenzhen International Graduate School, Tsinghua University, 2279 Lishui Road, Nanshan District, Shenzhen 518055, Guangdong, China
[2]Pengcheng Laboratory, 6001 Shahe West Road, Nanshan District, Shenzhen 518055, Guangdong, China
[3]School of Public Health, Anhui University of Science and Technology, 15 Fengxia Road, Changfeng County, Hefei 231131, Anhui, China
[4]Department of Computer Science, University of Kentucky, 329 Rose Street, Lexington 40506, Kentucky, USA
[5]Institute for Biomedical Informatics, University of Kentucky, 800 Rose Street, Lexington 40506, Kentucky, USA
*Corresponding authors. Zhixiang Ren, Pengcheng Laboratory, 6001 Shahe West Road, Nanshan District, Shenzhen 518055, Guangdong, China. E-mail: jason.zhixiang.ren@outlook.com; Qiang Cheng, E-mail: Qiang.Cheng@uky.edu.

Abstract

The development of single-cell and spatial transcriptomics has revolutionized our capacity to investigate cellular properties, functions, and interactions in both cellular and spatial contexts. Despite this progress, the analysis of single-cell and spatial omics data remains challenging. First, single-cell sequencing data are high-dimensional and sparse, and are often contaminated by noise and uncertainty, obscuring the underlying biological signal. Second, these data often encompass multiple modalities, including gene expression, epigenetic modifications, metabolite levels, and spatial locations. Integrating these diverse data modalities is crucial for enhancing prediction accuracy and biological interpretability. Third, while the scale of single-cell sequencing has expanded to millions of cells, high-quality annotated datasets are still limited. Fourth, the complex correlations of biological tissues make it difficult to accurately reconstruct cellular states and spatial contexts. Traditional feature engineering approaches struggle with the complexity of biological networks, while deep learning, with its ability to handle high-dimensional data and automatically identify meaningful patterns, has shown great promise in overcoming these challenges. Besides systematically reviewing the strengths and weaknesses of advanced deep learning methods, we have curated 21 datasets from nine benchmarks to evaluate the performance of 58 computational methods. Our analysis reveals that model performance can vary significantly across different benchmark datasets and evaluation metrics, providing a useful perspective for selecting the most appropriate approach based on a specific application scenario. We highlight three key areas for future development, offering valuable insights into how deep learning can be effectively applied to transcriptomic data analysis in biological, medical, and clinical settings.

Keywords: single-cell; spatial transcriptomics; deep learning

Introduction

The advancement of single-cell and spatial transcriptomics techniques has facilitated in-depth investigations of cellular characteristics, functions, and interactions, considering both cellular activity and spatial context within tissues. Single-cell RNA sequencing (scRNA-seq) quantifies gene expression at the cellular level, thereby elucidating cellular composition, gene expression patterns, and molecular characteristics [1, 2]. Recognized as the Method of the Year by Nature Methods in 2013 [3], scRNA-seq has significantly advanced research into complex biological questions, including mechanisms of disease resistance [4, 5], tissue heterogeneity [6, 7], targeted therapies [8], and embryonic development [9]. However, tissue dissociation disrupts spatial cell distribution and intercellular interactions, thereby constraining our understanding of the intricate processes occurring within multicellular organisms.

Spatial transcriptomics (ST) generates spatially resolved transcriptomic data to create detailed tissue maps at the subcellular level. This technique represents a significant advance in the field of transcriptomics, transitioning from cellular resolution to spatially sub-cellular resolution. In recognition of its importance in biomedical research, Nature Methods named spatially resolved transcriptomics as the Method of the Year in 2020 [10].

Single-cell and ST are crucial for studying the microenvironment at cellular and spatial resolutions, respectively. However, the complexity of biological tissues and the limitations of current sequencing techniques present significant analytical challenges. Traditional analysis techniques for bulk RNA sequencing were not designed to capture single-cell heterogeneity and spatial context, limiting their applicability to single-cell and ST [2]. As single-cell data increase in volume and complexity, traditional statistical methods struggle with high-dimensional and sparse data, facing challenges such as nonlinear relationships and high computational complexity.

Deep learning (DL), a powerful tool for modeling large-scale, high-dimensional complex data, has demonstrated its versatility

Received: December 16, 2024. **Revised:** February 17, 2025. **Accepted:** March 5, 2025
© The Author(s) 2025. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Table 1. Summary of previous reviews on the topic of single-cell and spatial transcriptomics.

Reference	Topic	Main points
Molho *et al.* 2024 [27]	Deep learning in single-cell analysis	Reviews seven key tasks in single-cell analysis—highlighting recent advancements, comparing classical and DL methods, and summarizing tools, datasets, and challenges.
Zahedi *et al.* 2024 [28]	Deep learning in spatially resolved transcriptfomics: a comprehensive technical view	Reviews DL methods for ST, highlighting areas for improvement such as capture complex biological details, data normalization, and gene expression counts modeling, while also providing a directory of accessible ST databases to guide future research.
Erfanian *et al.* 2023 [16]	Deep learning applications in single-cell genomics and transcriptomics data analysis	Examines DL applications in genomics, transcriptomics, spatial transcriptomics, and multi-omics integration, evaluating whether DL techniques offer advantages or face unique challenges in the single-cell omics domain.
Heydari *et al.* 2023 [29]	Deep learning in spatial transcriptomics: learning from the next next-generation sequencing	Provides an overview of state-of-the-art tools for spatially resolved transcriptomics analysis, focusing on DL-based approaches.
Bao *et al.* 2022 [17]	Deep learning-based advances and applications for scRNA-seq data analysis	Recent advances in DL-based methods for scRNA-seq data analysis were summarized, along with their applications, tools, and future challenges in analysis and interpretation.
Flores *et al.* 2022 [18]	Deep learning tackles single-cell analysis—a survey of deep learning for scRNA-seq analysis	Reviews 25 DL algorithms, analyzing their applicability in single-cell RNA-seq processing by providing a unified mathematical representation, comparing training strategies and loss functions, and linking these functions to specific data processing objectives.
Brendel *et al.* 2022 [19]	Application of deep learning on single-cell RNA sequencing data analysis: a review	Examines recent DL techniques in scRNA-seq data analysis, highlights their advantages over conventional tools, and discusses challenges and potential improvements in current approaches.
Ma *et al.* 2022 [30]	Deep learning shapes single-cell data analysis	Considers the progress, limitations, best practices and outlook of adapting DL methods for analyzing single-cell data.

Note. The table summarizes the author and published year, title, and the most important points covered by these reviews.

across numerous scientific domains, including small molecule modeling [11, 12], protein structure prediction [13, 14], drug development [15], etc. Recent reviews [16–19] of DL applications in single-cell data have introduced methods such as multilayer perceptrons [20], autoencoder (AE) [21], generative adversarial network (GAN) [22], convolutional neural network (CNN) [23], and graph neural network (GNN) [24]. These reviews explored both traditional and DL methods across various stages of the scRNA-seq analysis pipeline (Table 1). But they do not summarize current technological advances and challenges from a data science perspective. Moreover, existing data analysis techniques may not always be effective for addressing novel problems as the number of data modalities continues to grow [25, 26]. Here, modalities refer to different types of biological data, such as transcriptomic regulatory data, including chromatin accessibility, DNA methylation, gene regulatory networks, etc.

This review aims to discuss four major data science challenges, elucidate their origins, explore potential solutions, and provide insights into the underlying mechanisms of the methodologies. In terms of data sparsity, we examine issues such as the curse of dimensionality, noise, and uncertainty. Concerning data diversity, we categorize the integration of single-cell and ST data into two primary types: multimodal integration and multi-source integration. When dealing with data scarcity, we focus on missing data annotations and missing modalities. Additionally, from a data correlation perspective, we analyze methods for modeling spatiotemporal dependencies and incorporating prior knowledge. We highlight DL techniques and compare them with traditional machine learning approaches. We emphasize the advantages of DL, especially when integrated with statistical frameworks. Each algorithm is discussed alongside its mathematical foundations, focusing on both similarities and differences. Finally, we outline future directions in three key areas: the application of novel artificial intelligence (AI) methodologies, the development of fair and robust benchmark datasets with biologically interpretable evaluation metrics, and the exploration of DL applications in practical scenarios. This review provides a comprehensive overview of DL applications in single-cell and ST data analysis from a data science perspective. It offers insights that could inspire innovative solutions to emerging challenges in biological and medical research. The overall structure of the article is shown in Fig. 1.

Transcriptomic data

Bulk RNA sequencing provides gene expression profiles from various RNA samples, such as those from tissues, organs, or complete individuals. However, it averages the gene expression data across these samples, which limits its ability to accurately capture cellular heterogeneity. Consequently, it becomes challenging to discern whether the observed differences are due to changes in cellular composition or variations in gene expression (Fig. 2). The scRNA-seq addresses this limitation by profiling gene expression at the single-cell level. Additionally, ST integrates sequencing data with spatial context, providing a more comprehensive understanding of tissue construction and function.

Single-cell transcriptomes

In 2009, scRNA-seq technology emerged, making it possible to study the transcriptomes of individual cells [31]. Single-cell sequencing technology requires four main steps: (1) isolation of single cells; (2) reverse transcription; (3) cDNA amplification; and (4) sequencing library preparation and sequencing.

Isolation of single cells. Isolation of single cells refers to the process of separating individual cells from a complex tissue or cell population. Accurate and reliable capture is essential for

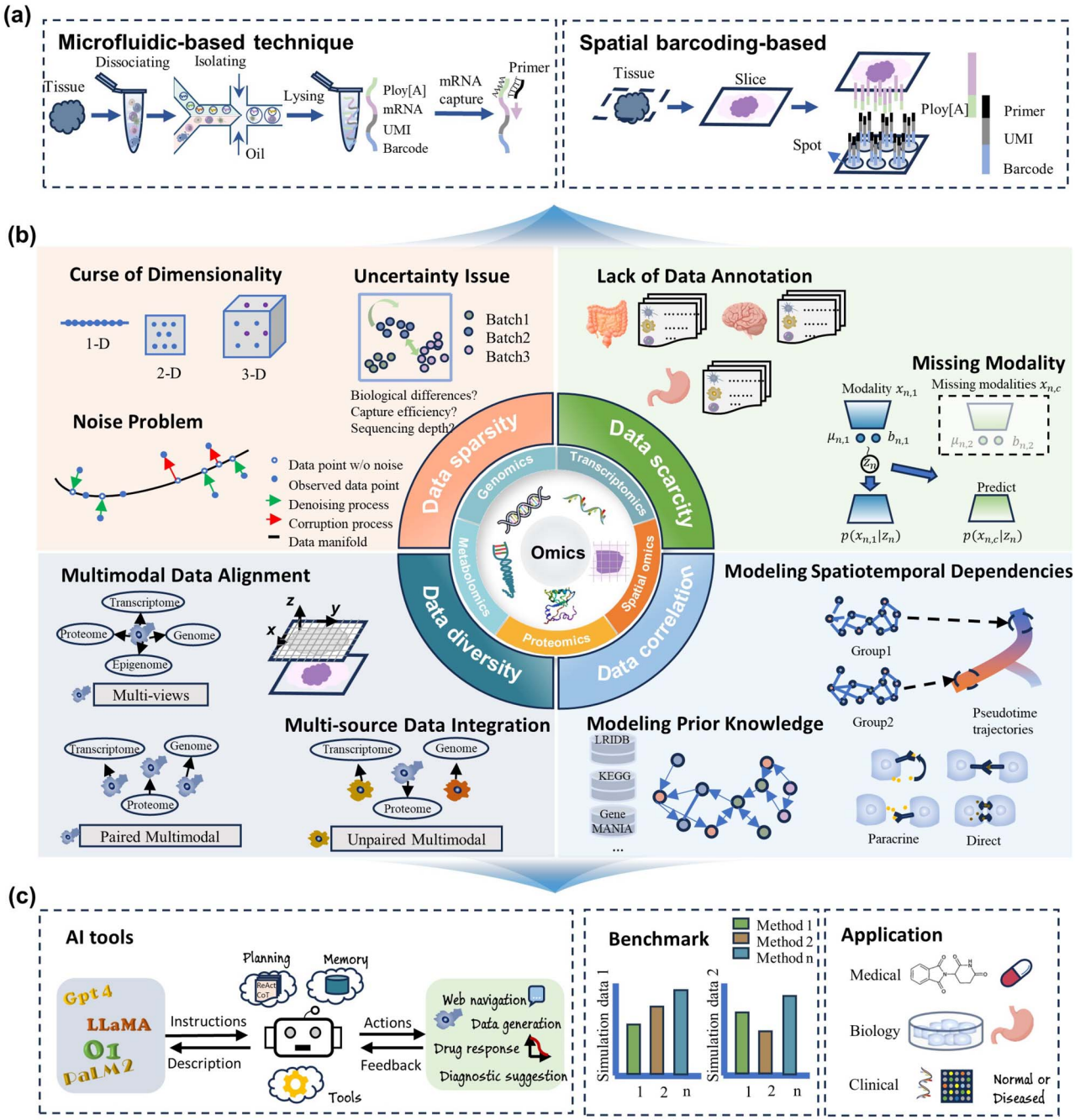

Figure 1. The overall structure of the article is organized into three main sections. (a) An overview of key sequencing technologies in single-cell and ST; (b) a discussion of four significant scientific and technical challenges within the field from a data science perspective: data sparsity, data diversity, data scarcity, and data correlation; (c) an exploration of potential future perspectives that includes innovative AI methodologies, benchmark datasets and evaluation metrics, as well as applications of DL in practical scenarios. Some components of this figure are drawn by Figdraw.

single-cell sequencing. The dissociation methods mainly include mechanical dissociation, enzymatic dissociation, and chemical dissociation. Target cells are selected from single-cell suspensions based on specific characteristics such as size, fluorescence, or surface labeling.

Fluorescence-activated cell sorting (FACS) [32] is a widely used high-throughput technique that labels cells with fluorescent dyes or antibodies targeting specific molecules on or within the cell. In flow cytometry, a laser excites a fluorescence marker, which emits a signal that is measured to quantify the molecular content. However, this method is less effective for cells with low marker expression and struggles to distinguish a subset of cells with similar fluorescence markers. Typical sequencing methods include Smart-seq [33], VASA-seq [34], and FLASH-seq [35].

Magnetic-activated cell sorting (MACS) [36] is another high-throughput isolation technique that separates and enriches specific cell types by binding magnetic beads to target cell proteins. The beads are conjugated with antibodies or other ligands, but unlike FACS, MACS isolates cells based on surface protein expression rather than gene expression, sorting them into positive and negative populations. MACS is primarily employed for the initial enrichment of cells and does not facilitate precise single-cell sorting as FACS does.

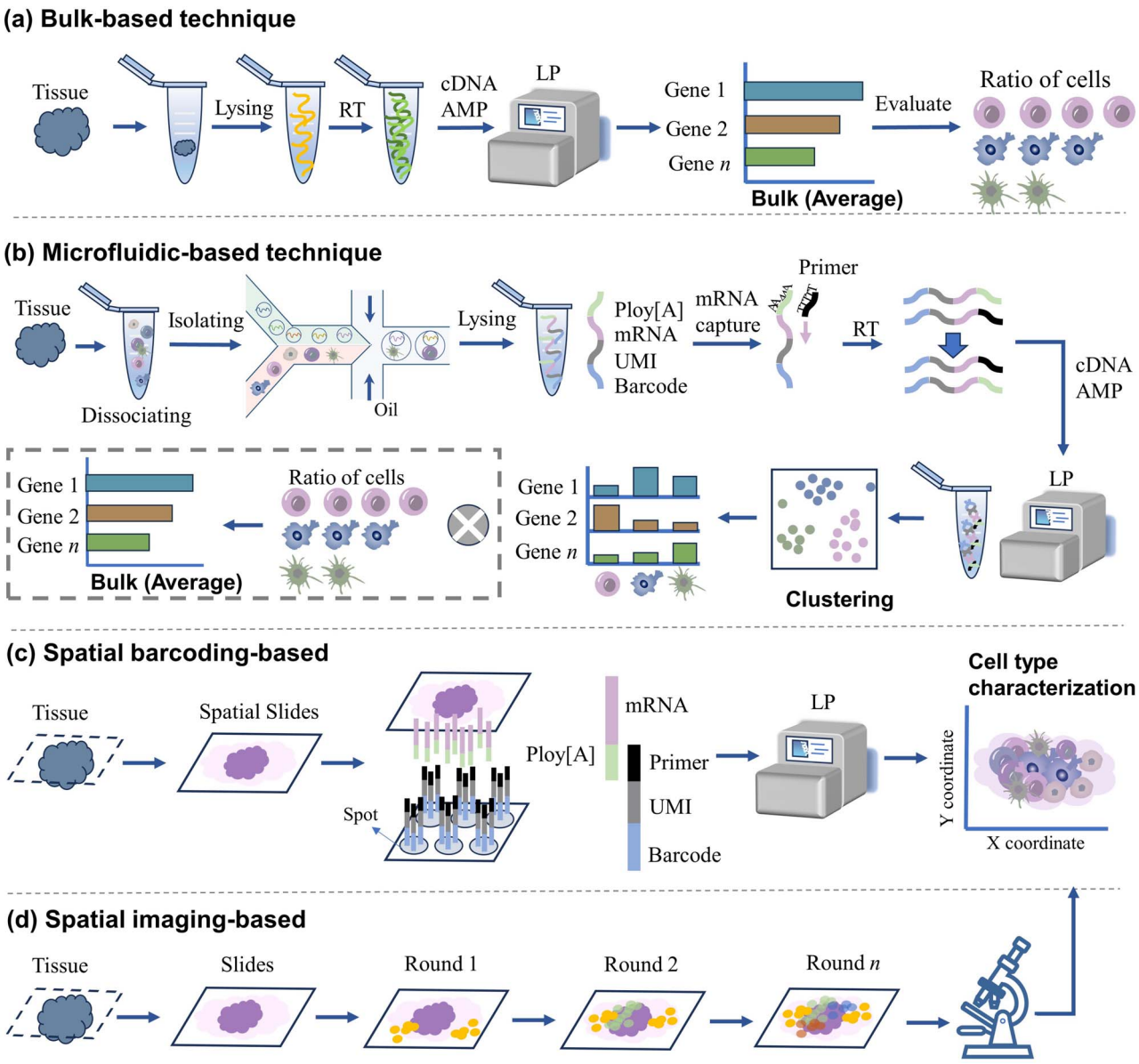

Figure 2. The sequencing pipeline for single-cell and ST data. (a) Bulk-based technique provides average gene expression profiles at the tissue level, with cell proportions estimated through deconvolution methods. (b) Microfluidic-based techniques isolate individual cells into droplets or wells, followed by barcoding and sequencing. (c) Spatial barcode-based techniques utilize cell barcodes to capture poly-adenylated RNA molecules *in situ* before reverse transcription. (d) Targeted *in situ* sequencing employs specifically designed probes to bind RNA or cDNA targets, leveraging *in situ* spatial information.

Microfluidic-based techniques exploit the inherent physical properties of cells for separation. These properties encompass cell size, shape, electrical polarizability, electrical impedance, density, deformability, magnetic susceptibility, and hydrodynamic characteristics [37]. In droplet-based microfluidics, individual cells are encapsulated within small droplets that are suspended in an immiscible fluid. Techniques such as InDrop [38], Drop-seq [39], and 10x Chromium [40] build upon this methodology. Microwell-based scRNA-seq methods—such as CytoSeq [41], Seq-Well [42], and Microwell-seq [43]—involve placing cells into discrete wells to ensure that each well contains either a single cell or none at all.

Reverse transcription. RNA cannot be directly sequenced within the cell. After cell lysis, the released RNA must be reverse transcribed to generate complementary DNA (cDNA). The poly(A) tailing method employs an oligo-dT primer that binds to the 3'-poly(A) tail of messenger RNA (mRNA), thereby facilitating its reverse transcription into cDNA. During this process, additional nucleotide sequences, such as cell-specific barcodes and uniform molecular identifiers (UMIs) for mRNA, are incorporated to uniquely label each cell and distinguish individual mRNA molecules.

cDNA amplification. Since mRNA is typically present in very low quantities within individual cells, it often proves inadequate for sequencing purposes. Therefore, cDNA amplification is necessary to produce sufficient amounts for subsequent library preparation. The most widely utilized method for this process is polymerase chain reaction-based amplification. Other methods such as *in vitro* transcription-based amplification and multiple displacement amplification are also employed in specific applications.

Sequencing library construction. The first step in library preparation involves converting nucleic acids into a sequencing

library, where DNA or RNA molecules are ligated to platform-specific adapters.

Spatial transcriptomes

ST techniques can be broadly categorized into two main types, based on how positional information is encoded before sequencing: (1) next-generation sequencing-based methods and (2) imaging-based methods.

Next-generation sequencing-based approaches encompass both the earlier microdissection-based techniques and the more widely adopted barcode-based approaches. Microdissection techniques isolate regions of interest through physical segmentation or optical selection, followed by collection for library preparation and sequencing. Microdissection-based techniques include tomo-seq [44], STRP-seq [45], Geo-seq [46], PIC-seq [47], TIVA [48], and NICHE-seq [49]. Of these, PIC-seq, TIVA, and NICHE-seq achieve cellular-level resolution [1]. However, physical segmentation is often performed manually, making it time-consuming. Additionally, optical selection requires the insertion of specialized markers into living cells or model organisms, which limits its application to formalin-fixed paraffin-embedded (FFPE) human samples. Accurately locating spatial locations remains a significant challenge, often resulting in relatively low spatial resolution.

The barcode-based sequencing technique, inspired by scRNA-seq, utilizes cell barcodes to capture poly-adenylated RNA molecules *in situ* before reverse transcription. This process is facilitated by a capture probe that incorporates a spatial barcode, a UMI, and poly-T oligonucleotides, followed by the synthesis of cDNA. Spatial barcodes operate analogously to cellular barcodes, ensuring an accurate mapping of transcriptomes obtained from tissue slices back to their original locations. The spatial resolution of this method depends on the distance between adjacent spots, achieving a maximum resolution of $\sim 0.5-0.7\mu m$, which facilitates subcellular analysis. However, enhancing spatial resolution often leads to compromises in detection sensitivity and gene coverage. Examples of barcode-based sequencing technique include 10x Visium [50], Slide-seq (V2) [51], HDST [52], Stereo-seq [53], Seq-scope [54, 55], and Decoder-seq [56]. Moreover, Open-ST [57] can generate ST in 3D.

Barcode-based sequencing techniques have limitations related to spot size, spatial resolution, and capture efficiency, which constrain their application to tissues with fine-scale cellular structures or low-abundance transcripts [58]. In contrast to barcode-based techniques, image-based methods directly leverage *in situ* spatial information without the need for spatial barcodes. Techniques such as *in situ* hybridization and targeted *in situ* sequencing (ISS) utilize gene-specific cDNA or RNA probes that bind to target sequences within fixed cells or tissues. Subsequently, spatial mapping of gene expression is accomplished by imaging, typically employing fluorescence or other markers. However, the number of detectable transcripts is constrained by optical limitations, allowing to detect only a few hundred targets. Ongoing technological advances aim to enhance the multiplexing capability. SeqFISH+ [59] increases the number of colors used in seqFISH [60] from four or five to 60 "pseudo colors" in its readout probe palette. In contrast, MERFISH [61] utilizes three-color imaging and only requires 23 rounds of imaging, whereas SeqFISH+ needs 80 rounds of hybridization. Both methods are capable of analyzing ~10 000 genes.

RNA techniques based on ISS can be divided into targeted and untargeted approaches. Targeted ISS involves the binding of RNA or cDNA targets with specifically designed probes, such as padlock probes, followed by rolling-circle amplification to replicate these targets for sequencing. In contrast, the untargeted ISS transcribes the transcript into cDNA using standard reverse transcription, which is followed by DNA amplification and sequencing. This approach does not require preselection of target genes, but may exhibit lower detection efficiency. Examples of untargeted ISS include STARmap [62] and ExSeq [63].

Database

The volume of sequencing data has grown exponentially with the rapid advances in single-cell and ST technologies, highlighting the need for curated databases, robust analysis pipelines, and effective visualization tools. This review collects 12 large-scale single-cell sequencing databases (Table 2) and seven ST databases (Table 3). A concise overview is provided in Table 4.

Single-cell omics highlights the significance of spatial context, which will increasingly be incorporated to develop multi-omics databases. The establishment of such a database not only emphasizes the integration of data sets from diverse sources, but also necessitates data preprocessing, analysis, visualization, user interaction, and other critical components.

Challenges and DL Methods in single-cell data analysis

Data sparsity

Single-cell transcriptomes typically encompass tens of thousands of genes and exhibit considerable variability in expression across individual cells. Many genes remain inactive within a particular cell type, and even within the same cell type, certain genes may be transiently unexpressed due to the dynamic nature of the transcriptome and fluctuations in the cell cycle state. Consequently, gene expression matrices tend to be highly sparse, which presents challenges in modeling feature spaces. These challenges include issues related to the curse of dimensionality, noise, and uncertainty. The overall structure of this section is shown in Fig. 3.

Curse of dimensionality

The curse of dimensionality arises from the exponential increase in the volume of space as the number of dimensions grows. This leads to sparsity and makes it more difficult to cover the space effectively with a limited number of observations. Consequently, the similarity between data points diminishes. To address this challenge, feature selection and dimensionality reduction techniques are frequently employed. Traditional methods for dimensionality reduction include parameterized approaches such as principal component analysis (PCA) [82], along with variants such as scPCA [83] and FastRNA [84]. Additionally, nonparametric methods such as t-distributed stochastic neighbor embedding (t-SNE) [85–87] and uniform manifold approximation and projection [88, 89] are also widely utilized.

Nonparametric methods aim to map high-dimensional data into a lower dimensional space while preserving local structure, typically described in terms of probabilities or metric learning. In contrast, parametric methods model the relationships between data points through explicit mathematical formulations with parameters estimated from training data. DL is a powerful parametric approach that employs neural networks for automated modeling, enabling end-to-end parameter learning directly from data. Compared with traditional methods, DL has demonstrated superior effectiveness in managing complex and high-dimensional feature spaces [27, 30, 90]. Scvis [91], a nonlinear

Table 2. Commonly used database of million-level single cells.

	Name	Type	Species	Datasets	Tissues	Cells (M)	Cell types
1	GEO [64]	Comprehensive	\	4348	\	\	\
2	HCA [65]	Comprehensive	1 (H)	200	80	19.9	200
3	SCP [66]	Disease	16	780	106	55.1	640
4	CELLxGENE [67]	Comprehensive	1 (H)	1634	\	98.6	942
5	PanglaoDB [68]	Comprehensive	2 (H&M)	\	258	5.6	\
6	ABC Atlas	Brain	23	166	\	4.0	34
7	CSEM [69]	Cancer	1 (H)	1466	74	7.3	80
8	EA [70]	Comprehensive	66	4451	\	5.9	\
9	HUSCH [71]	Comprehensive	1 (H)	185	45	3	270
10	DISCO [72]	Comprehensive	1 (H)	4593	107	18	\
11	EMBL-EBI [73]	Comprehensive	12	123	\	\	\
12	hECA [74]	Comprehensive	1 (H)	116	38	1.1	146

Note. The table summarizes the name, type, number of species (abbreviated as "Species"), number of datasets (abbreviated as "Datasets"), number of tissues (abbreviated as "Tissues"), number of cells (in millions, abbreviated as "Cells (M)"), and number of cell types (abbreviated as "Cell types") for each database. H: Human, M: Mus musculus. The symbol "\" indicates that the corresponding statistics are missing.

Table 3. Commonly used database of spatial transcriptomics data.

	Name	Type	Species	Datasets	Tissues	Samples (k)	Publications
1	SpatialDB [75]	Comprehensive	5	305	\	\	5
2	Aquila [76]	Disease	5	110	26	6.5	81
3	SOAR [77]	Disease	11	304	40	2.8	\
4	STOmicsDB [78]	Comprehensive	17	231	128	7.7	7339
5	SPASCER [79]	Comprehensive	4	1082	16	\	43
6	SODB [80]	Comprehensive	12	>2000	76	\	\
7	SORC [81]	Cancer	1 (H)	82	17	0.3	\

Note. The table summarizes the name, type, number of species (abbreviated as "Species"), number of datasets (abbreviated as "Datasets"), number of tissues (abbreviated as "Tissues"), number of samples (abbreviated as "Samples"), and number of publications (abbreviated as "Publications") for each database. H denotes Human. K is an abbreviation for thousand. The symbol "\" indicates that the corresponding statistics are missing.

dimensionality reduction method based on the variational autoencoder (VAE) framework, integrates generative modeling with variational inference (Fig. 4a). By using two distinct neural network structures, it facilitates bidirectional mapping from high-dimensional data to low-dimensional (i.e. cell embedding) space, thereby preserving the global structure of the data. Consider a high-dimensional scRNA-seq dataset $\mathcal{D} = \{\mathbf{x}_n\}_{n=1}^{N}$, which comprises N cells, where x_n denotes the gene expression vector for each cell. It is assumed that the observed data are generated from a low-dimensional prior distribution given by $p(x_n \mid z_n, \theta) = \text{Distribution}((\mu_\theta(z_n)), \sigma_\theta(z_n))$; this is typically modeled as a factorized standard normal distribution expressed as $p(\mathbf{z}_n \mid \theta) = \prod_{i=1}^{d} \mathcal{N}(z_{n,i} \mid 0, \mathbf{I})$, through a transformation parameterized by θ. This parameter is difficult to compute directly and is instead approximated using a neural network. For each cell, the generative distribution can be expressed as the following integral:

$$p(\mathbf{x}_n \mid \boldsymbol{\theta}) = \int p(\mathbf{z}_n \mid \boldsymbol{\theta})\, p(\mathbf{x}_n \mid \mathbf{z}_n, \boldsymbol{\theta})\, \mathrm{d}\mathbf{z}_n \tag{1}$$

However, computing the posterior distribution $\mathbf{p}(\mathbf{z_n} \mid \mathbf{x_n}, \theta)$ based on the observed data is intractable. To address this issue, a variational distribution $\mathbf{q}(\mathbf{z_n} \mid \mathbf{x_n}, \boldsymbol{\phi})$ is introduced as an approximation. It is assumed that the variational distribution follows a multivariate Gaussian distribution characterized by mean $\mu_\phi(\mathbf{x_n})$ and standard deviation $\sigma_\phi(\mathbf{x_n})$, both of which are functions of x_n, parameterized by a neural network. The model is then optimized to ensure that similar cells exhibit analogous posterior distributions. Consequently, the low-dimensional latent space effectively preserves the distance relations of the high-dimensional data, leading to efficient dimensionality reduction.

Noise issues

Biological noise in scRNA-seq data arises from the intrinsic randomness of biological systems and variations in cellular states. In contrast, experimental noise reflects nonbiological fluctuations due to technical limitations or random errors. Systematic biases, commonly referred to as batch effects, occur due to discrepancies in experimental conditions, instruments, reagents, and procedures across data batches. Furthermore, technical constraints or low capture efficiency often lead to missing values, resulting in a large number of zeros in the gene expression matrix. These zeros can obscure the true biological signal, a phenomenon known as dropout events. Batch effects and dropout events always co-occur in real-world datasets. Research has increasingly focused on batch effect correction and imputation methods to address these challenges.

Batch effect correction improves the comparability of datasets derived from different batches, ensuring that observed differences reflect genuine biological variation. This process involves mapping the identical cell types from various experiments to a common region within a latent space. Traditional methods, such as nearest neighbor matching (NNM) [93, 94], address this issue by aligning representations across batches. DL-based approaches focus on the hidden space where samples are mapped to semantic representations, facilitating better learning of the underlying patterns. These methods preserve essential biological signals while filtering out irrelevant features through data reconstruction. The CLEAR algorithm [95] removes batch effects by adding simulated batch-related noise during data augmentation to create positive and negative pairs, then uses contrastive learning to generate similar representations for positive pairs and distinct ones for negative pairs, integrating cells from different batches while

Table 4. A list of 12 large-scale single-cell sequencing databases and seven ST databases.

Name	Brief introduction
GEO [64]	A widely used repository for gene expression data, encompassing profiles from diverse species and experimental conditions. It includes 26 712 platforms and 4348 datasets, each assigned a unique identifier (GEO Accession ID), and provides standardized formats and annotations.
HCA [65]	Delivers a comprehensive atlas of human cells, detailing their molecular and spatial characteristics across various organs, tissues, developmental stages, and disease states.
SCP [66]	Facilitates the sharing and exploration of single-cell genomics data. It enables researchers to contribute datasets and create visualizations without additional development effort.
CELLxGENE [67]	Offers real-time tools for analyzing large-scale single-cell data. Its scalable and flexible framework allows users to adapt the code to specific analytical requirements.
PanglaoDB [68]	Provides preprocessed and precomputed analyses, simplifying data exploration. Its online interface supports queries on cell types, genetic pathways, and regulatory networks, removing the need for extensive preprocessing.
ABC Atlas	provides a platform for visualizingdata from multiple different cells in the mammalian brain and empower researchers to simultaneously explore and analyze multiple brain datasets.
CSEM [69]	Integrates scRNA-seq data from diverse human cancers, enabling researchers to explore immune profiles, gene expression dynamics, and metabolic reprogramming within tumor microenvironments.
EA [70]	An open-access platform that provides comprehensive information on gene and protein expression across species, tissues, cell types, and biological conditions, with tools for data exploration, visualization, and analysis.
HUSCH [71]	A comprehensive scRNA-seq database offering detailed cell-type annotations, gene expression visualizations, and functional analyses
DISCO [72]	A comprehensive single-cell omics database that integrates over 18 million cells with harmonized metadata, offering tools such as FastIntegration, CELLiD, and CellMapper for data integration, annotation, and projection onto global and tissue-specific atlases.
EMBL-EBI [73]	Manages a comprehensive suite of open data resources and tools for life sciences, including the Pathogens Portal, which provides extensive biomolecular data on over 200,000 pathogen species to support infection biology, pathogen surveillance, and public health research.
hECA [74]	Integrates over 1 million human cells from diverse datasets, providing advanced tools for data retrieval, multi-view biological representations, and customizable reference creation for applications in various biological studies.
SpatialDB [75]	The first manually curated ST database. It includes 24 datasets (305 sub-datasets) spanning five species, generated using eight ST technologies. It features 6,000 gene-cell-type associations and supports automatic cell type annotation.
Aquila [76]	Facilitates transcriptomics and proteomics analyses for both 2D and 3D experiments. and provides diverse visualizations, such as spatial cell distributions, expression patterns, and marker co-localization. Researchers can securely upload and analyze their ST data, enabling personalized exploration.
SOAR [77]	Hosts large-scale ST datasets with systematic management and analysis. It ensures consistency through uniform processing and annotation, enhancing reliability for comparative studies and benchmarking.
STOmicsDB [78]	Offers comprehensive analyses, including cell type annotation, spatial region and gene identification, and cell–cell interaction insights, facilitating deeper biological understanding.
SPASCER [79]	Specializes in advanced analyses such as ST deconvolution, spatial cell–cell interactions, gene pattern detection, and pathway enrichment, supporting more detailed spatial investigations.
SODB [80]	Accommodates a broad range of ST technologies with a user-friendly interface. It enables users to generate molecular markers for specific regions, enhancing flexibility in data exploration.
SORC [81]	The first ST database dedicated to cancer research, SORC includes 269 tissue slices from seven cancer types and integrates 46 single-cell data types. It provides a detailed spatial cell atlas, facilitating insights into tumor microenvironment interactions by uncovering specific genes and pathways.

Note. For single-cell sequencing, only data sets with million-level cells were included.

preserving cell-type differences (Fig. 4b). Similarly, BERMUDA [96] aligns batch distributions using the maximum mean discrepancy loss, facilitating the integration of data in the latent space.

Imputation methods are designed to reconstruct missing gene expression values by distinguishing technical noise from real biological zeros, relying on observed data patterns. Various approaches have been developed to address this issue, including traditional matrix factorization techniques, similarity modeling, statistical approaches, and DL strategies. Matrix factorization methods mitigate noise by preserving the dominant low-rank signal while discarding extraneous components. Traditional matrix factorization techniques are based on singular value decomposition and nonnegative matrix factorization (NMF), such as ALRA [97]. Similarity-based approaches leverage relationships between cells to infer missing values. Actually, they leverage gene expression profiles from other cells. An example is MAGIC [98], which employs a k-nearest neighbor algorithm to smooth and impute data based on the neighborhood of similar cells. ScImpute [99] also relies on local structure for imputation. Statistical modeling methods employ predefined or probabilistic models to fit observed data. For example, SAVER [100] assumes gene expression and follows a negative binomial distribution modeled through a Poisson-gamma mixture. It estimates the parameters using an empirical Bayesian approach with Poisson LASSO regression, and outputs the posterior means as imputed values. DL methods reframe traditional matrix operations as neural network layers, transforming parameter estimation into an optimization problem. Common frameworks include autoencoder-based models and generative architectures that adaptively learn complex patterns to enhance imputation accuracy.

AEs are widely used for imputation, effectively integrating dimensionality reduction and denoising within a unified framework. As unsupervised learning models, AEs transform high-dimensional input data into a lower dimensional latent space,

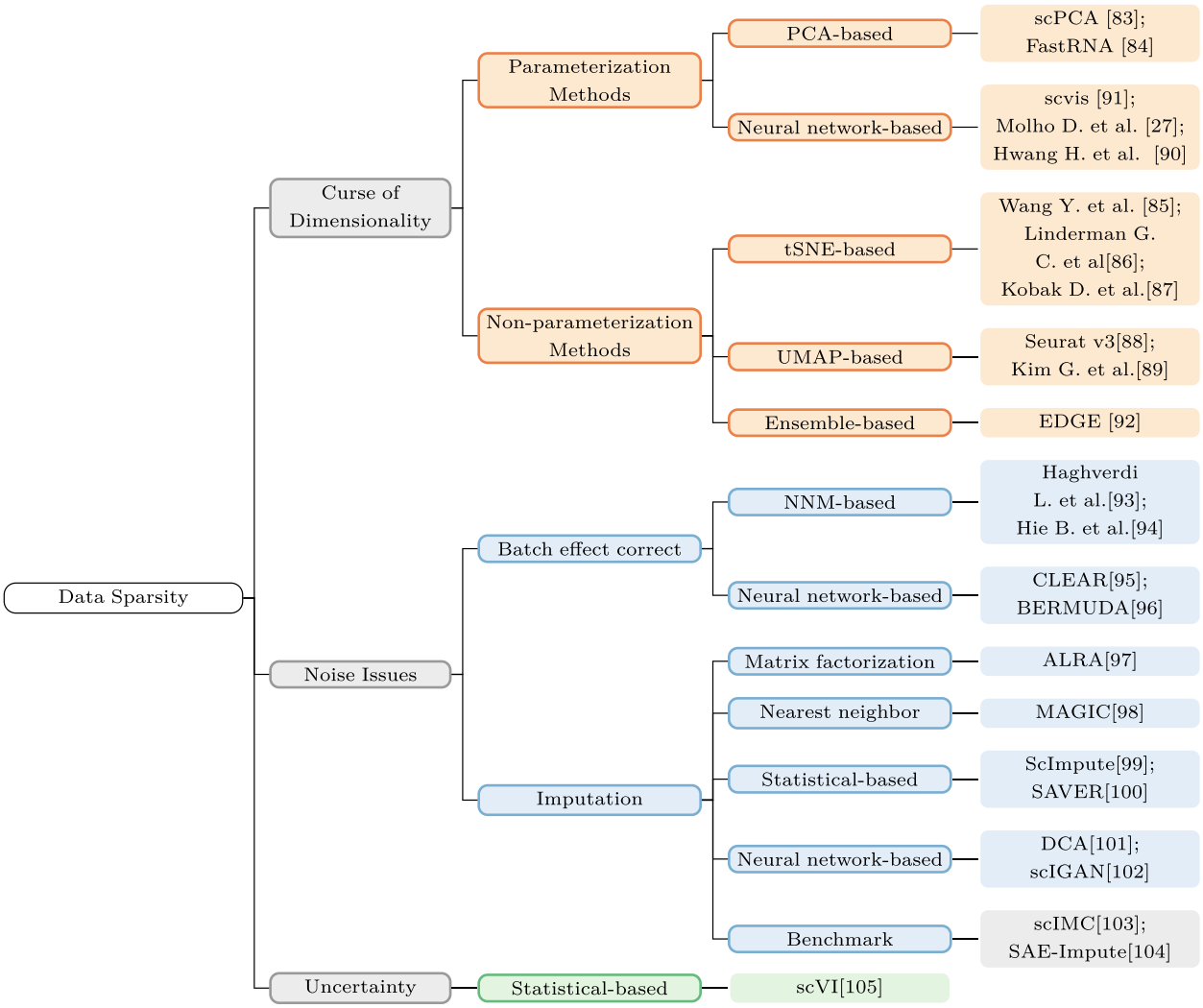

Figure 3. The structure of section "Data Sparsity" and related methods.
Note. The tree chart outlines the challenges associated with processing sparse single-cell data, focusing on issues including the curse of dimensionality, noise, and uncertainty.

preserving essential features while eliminating redundant information. This latent representation provides contextual insights that facilitate the imputation process. The decoder reconstructs the original input from this compact representation with the objective of generating an output that closely resembles the initial data. For example, DCA [101] employs an autoencoder with a zero-inflated negative binomial (ZINB) noise model to infer key parameters such as the mean, dispersion, and dropout probabilities associated with gene expression data (Fig. 4c). The decoder produces a denoised reconstruction that is well aligned with the modeled data distribution. Variants of AEs, such as VAEs [106], conditional autoencoders [107], and sparse autoencoders [108], offer additional flexibility tailored to specific applications.

Generative models, such as GANs, address the limitations of similarity-based methods that often lead to over-smoothed imputations. GAN-based models are designed to learn the underlying data distribution and generate new samples that closely resemble the denoised data. ScIGAN [102] generates synthetic single-cell profiles instead of directly estimating missing values from observed data. This strategy minimizes overfitting to dominant cell types while improving imputation for rare cell populations. The distinctive design of scIGAN involves transforming real gene expression data into a two-dimensional image representation. The generator synthesizes gene expression profiles from latent variables, whereas the discriminator distinguishes between real and synthetic images. Both networks are trained adversarially and their performance is refined by iterative competition.

We summarized and reanalyzed the imputation performance of 12 methods, including scImpute [99], SAVER [100], ALRA [97], MAGIC [98], scTSSR [109], DCA [101], DrImpute [110], DeepImpute [111], AutoImpute [112], scIGANs [102], and scGAIN [113], across five benchmark datasets [102, 103]. For each method, we assessed two key performance metrics: imputation consistency and clustering performance. Imputation consistency was evaluated using metrics such as F1 score, AUC, and accuracy (ACC), while clustering performance was assessed using NMI and ARI. We averaged these metrics across the datasets to get an overall performance score for each method. The results revealed that DCA and scIGANs each achieved the highest imputation consistency across the two benchmarks. Notably, DCA displaying robust clustering performance in three out of five datasets (Fig. 5).

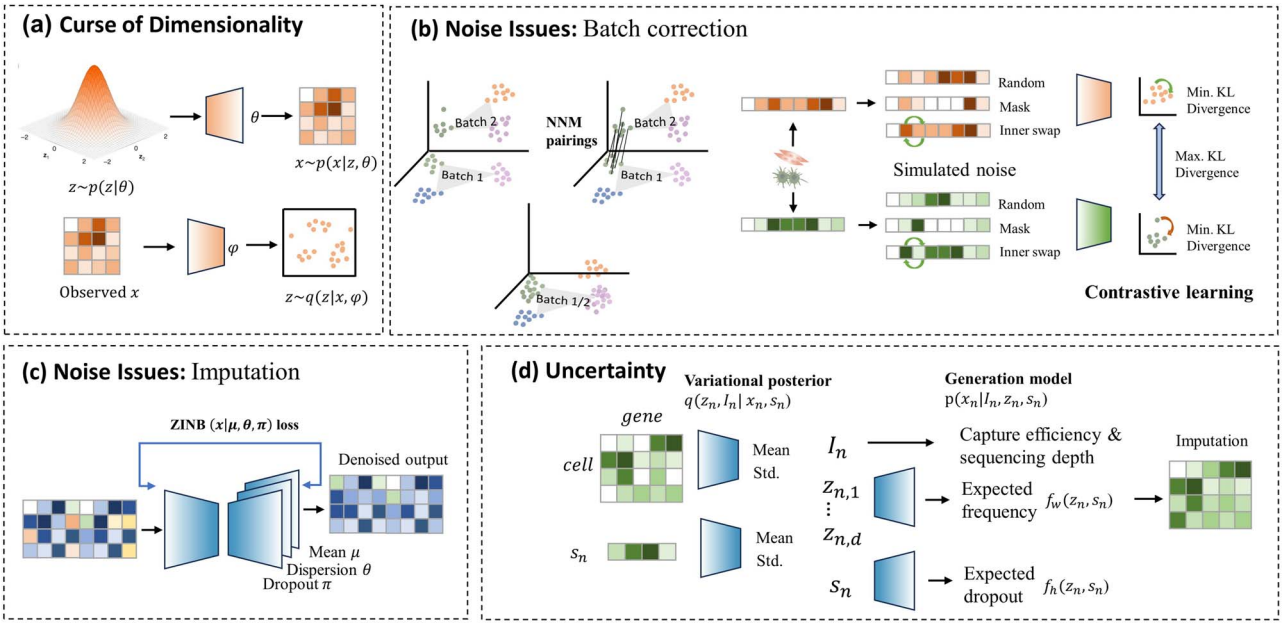

Figure 4. The challenges and typical approaches for data sparsity.
Note. Neural networks often modeling data representations in complex latent spaces, particularly in scenarios with increased factors of variability, such as uncertainties in experimental processes. (a) For curse of dimensionality, we plotted the framework of scvis [91]. (b) For batch correction, the neural networks (CLEAR [95]) shares the same objective as the NNM [93]. Both approaches use distance to measure the similarity between samples, facilitating the clustering of samples of the same type across different batches. (c) For imputation, the VAE-based method (DCA [101]) is used for noise separation through data reconstruction. (d) For modeling uncertainty, scVI incorporates stochastic factors inherent in the sequencing process, providing a framework to better capture and account for variability in the data.

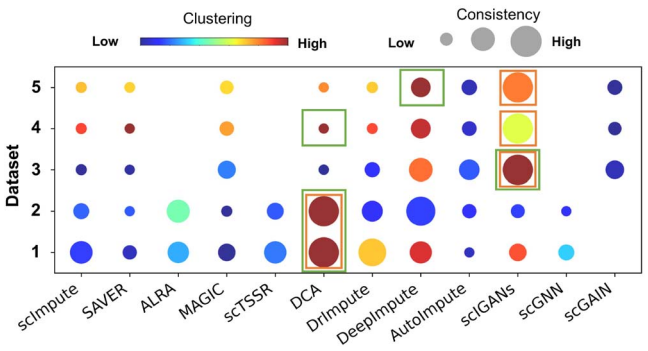

Figure 5. Revisualize the benchmark results for data imputation from five benchmark datasets [102, 103].
Note. It illustrates that the performance of different methods varies significantly across different benchmark datasets (with missing values represented as blanks). DCA and scIGANs each achieved the highest imputation consistency in both benchmarks, with DCA performing best in benchmark 1 and scIGANs in benchmark 2. In benchmark 1 (dataset 1 and 2), "clustering" represents the average value of clustering evaluation metrics, including NMI and ARI, while "consistency" includes PCC. In benchmark 2 (dataset 3-5), "clustering" represents the mean of the NMI and ARI, and "consistancy" refers to the mean metrics of F1, AUC, and ACC. The orange rectangle indicates the largest point size (imputation consistency), while the green rectangle represents the highest color value (clustering performance) across all methods. Red represents the highest color value. Dataset 1 and Dataset 2 represent six scRNA-seq datasets simulated by Splatter [114] with dropout rates of 0.42 and 0.78, respectively [103]. Datasets 3-5 refer to CIDR [115] simulation datasets with dropout rates of 0.71, 0.83, and 0.87 [102].

Uncertainty

Uncertainty issues typically arise from factors that contribute to ambiguity during analysis, decision-making, or prediction. This is caused by insufficient information, measurement errors, inaccurate model assumptions, or other sources of variability. In addition to the aforementioned approaches for addressing batch effects and dropout events, uncertainty quantification can enhance model selection and performance evaluation. This procedure facilitates the mitigation of uncertainties arising from experimental data and model assumptions [116]. An example is scVI [105], which explicitly incorporates batch annotations and addresses batch effects through conditional independence assumptions (Fig. 4d). This approach effectively isolates batch-related factors from the data, thereby reducing the uncertainties associated with batch differences and improving gene expression analysis. scVI models the observed expression x_{ng} of each gene g in each cell n as a random sample from a ZINB distribution denoted as $p(x_{ng}|z_n, s_n, l_n)$. Here, z_n represents a low-dimensional Gaussian vector that captures biological differences between cells. l_n is a one-dimensional Gaussian variable that accounts for variation due to differences in capture efficiency and sequencing depth. It serves as a cell-specific scaling factor. s_n denotes the batch annotation of the cell (if available). Employing variational inference and reparameterization techniques, scVI optimizes the posterior distribution via a variational lower bound. By incorporating sources of uncertainty, including cell-specific and batch-dependent features, this approach effectively preserves the information inherent in the original data. In contrast, posterior correction methods may rely on fixed assumptions, which can lead to the loss of critical information or introduce bias.

Data diversity

The "central dogma" delineates the flow of genetic information from DNA to RNA and subsequently to proteins, establishing a foundational framework for understanding how gene expression influences cellular functions. Omics data, obtained through high-throughput techniques, provide a systematic characterization of the various molecular components within an organism, including the genome, transcriptome, proteome, and

metabolome. Recent advances in multichannel sequencing now allow simultaneous measurement of multiple types of omics data. Current transcriptome-focused multimodal techniques include combinations such as gDNA-mRNA [137, 138], mRNA-methylation [139, 140], mRNA-ATAC [141–143], mRNA-proteome [144, 145], and mRNA-methylation-ATAC [146, 147]. The observed heterogeneity encompasses intra-sample heterogeneity, inter-sample heterogeneity, and variability across species and individuals. Intra-sample heterogeneity results from differences in sequencing depth, coverage, and data type. Inter-sample heterogeneity is caused by variations in experimental design, sample handling, and sequencing protocols. The diversity and complexity inherent in single-cell data present significant challenges, particularly when it comes to aligning and integrating paired and unpaired datasets. "Paired" data refer to multimodal datasets derived from the same sample, whereas "unpaired" data consist of multimodal datasets from different samples or platforms. Multi-omics analysis integrates these diverse data types to facilitate a comprehensive understanding of organismal heterogeneity and regulatory mechanisms. The overall structure of this section is shown in Fig. 6.

Multimodal data alignment

This section focuses on the alignment of multimodal data, including multi-omics data as well as paired single-cell and ST data, with the goal of uncovering intrinsic patterns in the alignment of homologous data.

Scale discrepancies

A critical challenge in multi-omics data integration is addressing the scale discrepancies between different modalities. These variations in the types, biological significance, and spatial distribution of features across different omics layers, such as transcriptomics, proteomics, and genomics, contribute to the complexity of associating multi-omics data. This issue complicates the alignment of shared patterns and preservation of unique modality-specific characteristics. The "shared patterns" refer to common biological significance, such as cell types or regulatory elements, which are present across different omics datasets. Several approaches have been proposed to mitigate these challenges. For example, LIGER [117] and iNMF [118] utilize NMF and its variants to decompose multi-omics data into lower dimensional latent structures called "latent metagene factors." These methods extract both shared and modality-specific factors. It reconstructs each dataset by combining these factors, ensuring that both shared biological signals and dataset-specific variations are preserved. By aligning features at the shared and modality-specific levels, these methods address scale discrepancies, facilitating more accurate integration of diverse datasets. (Fig. 7a). iNMF builds on LIGER by enabling online learning to progressively incorporate new data, making it well suited for handling large, evolving multi-omics datasets by adapting to changes over time and maintaining robustness to dataset variability. Similarly, models like scAI [119], MultiVI [120] and scMVAE [122], which adopt a VAE framework, handle scale discrepancies by generating modality-specific latent representations. MultiVI averages these representations to form a joint latent space, while scMVAE incorporates a shared latent space with separate modality-specific decoders for different omics data types. These models impose constraints on the latent space to minimize distance between representations, effectively reducing the impact of scale differences. Additional VAE-based models include Cobolt [121] and scMM [123]. Furthermore, scMVP [124] introduces a Gaussian mixture model prior within a VAE framework, providing a more accurate representation of complex data distributions. Each Gaussian component in the mixture corresponds to a distinct latent "cluster" that can represent distinct underlying biological processes or cell types, whereas a standard VAE may not fully separate them.

Resolution mismatch

Integrating ST with scRNA-seq data is essential for achieving high-resolution insights into gene expression within the spatial context of tissues. However, a significant challenge arises from the resolution mismatch: many ST sequencing platform typically captures gene expression across tens to thousands of cells per spatial spot, whereas scRNA-seq provides single-cell resolution. To address this challenge, recent approaches focus on aligning these datasets through shared latent spaces and advanced computational strategies. Methods such as Seurat [126], LIGER [117], and Harmony [127] utilize mutual nearest neighbors (MNN) and shared latent representations to integrate scRNA-seq and ST data, facilitating cell-type label transfer and enhancing weak ST signals. Alternatively, RCTD [128] integrates reference scRNA-seq data to model cell-type-specific gene expression profiles, estimating their relative abundance within spatial spots using a Poisson distribution and maximum likelihood estimation. SoScope [129] employs a multimodal DL framework that integrates transcriptomic, histone, DNA, and protein data with high-resolution images, utilizing a variational Bayesian inference network to infer precise spatial maps (Fig. 7b). These strategies collectively provide solutions for overcoming resolution mismatches, enabling more accurate spatial gene expression mapping.

Modeling spatial information

Unlike single-cell omics data, spatial omics incorporates spatial context, making the modeling of spatial information a critical challenge in data integration, slice registration, and the alignment of molecular-scale data (e.g. pathology). Many open-source frameworks have been developed, including SpatialData [130], SSGATE [148], STAligner [131], and SpatialGlue [149]. SpatialGlue models spatial information by constructing spatial and feature neighbor graphs for each modality and using GNN encoders to learn modality-specific graph representations. It then integrates these representations through within- and cross-modality attention layers, preserving spatial relationships and the relative importance of each modality in the final integrated output. STAligner aligns 2D tissue slices using a triplet-based method to identify MNN with similar gene expression patterns, enabling accurate coordinate registration and 3D tissue reconstruction (Fig. 7c). ST-Net [132], on the other hand, models the correlation between ST and spatially continuous molecular pathology, employing an end-to-end neural network to predict spatially resolved transcriptomics.

Recent studies have highlighted advancements in ST techniques [150], although challenges such as the trade-off between spatial resolution and measurement throughput remain. Nevertheless, modeling spatial information in ST by bridging imaging and sequencing holds great potential for integrating diverse modalities from histopathology and single-cell data. This integration will enhance our understanding of tissue organization, microenvironmental interactions, and histopathology.

Integration of multi-source data

The integration of multi-source data enhances the predictive power of models, offers more robust conclusions, and uncovers

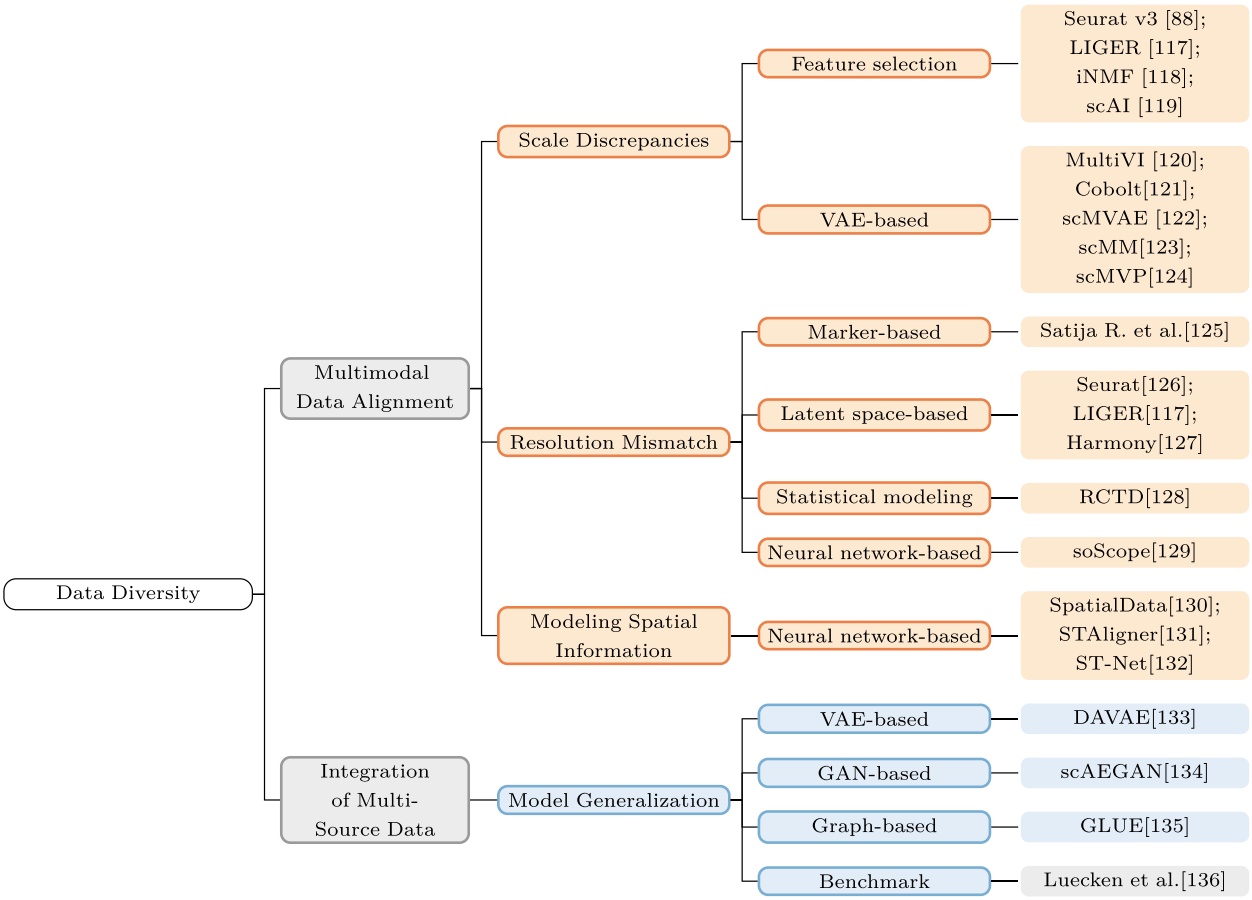

Figure 6. The structure of section "Data Diversity" and related methods.
Note. The tree chart outlines the challenges associated with processing multi-view single-cell data, focusing on issues including multimodal data alignment, and the integration of multi-source data.

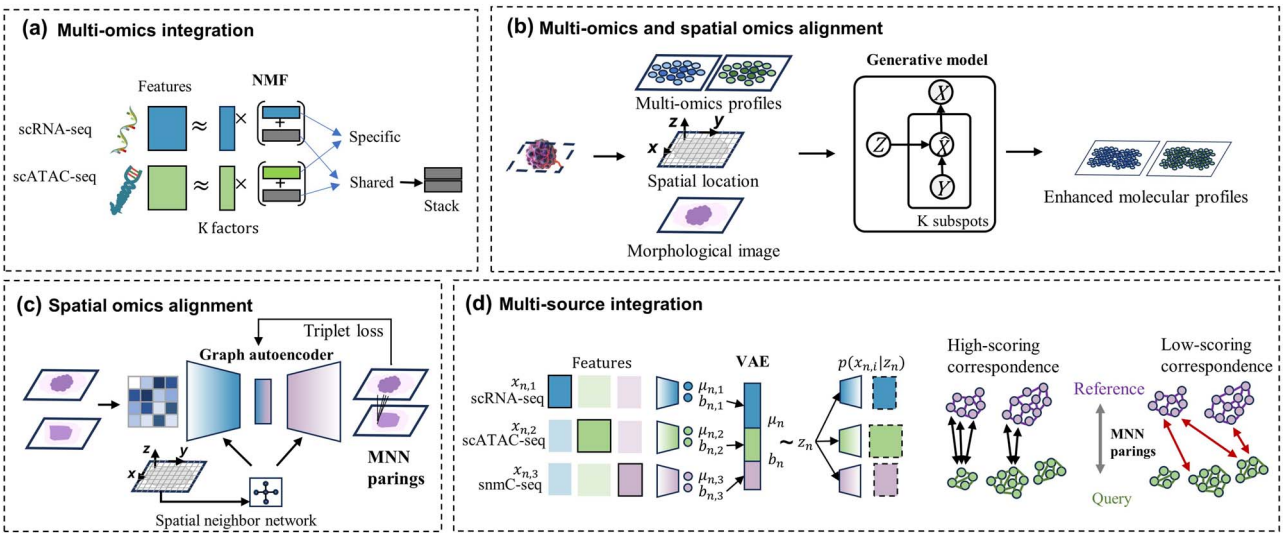

Figure 7. The challenges and typical approaches for data diversity.
Note. The integration of multi-modal and multi-omics data via DL is a trend in the study of data diversity. (a) For multi-omics integration, we plotted the framework of LIGER [117]. Multi-omics data alignment aims to align similar features while preserving the unique characteristics of each modality. (b) For multi-omics and ST alignment, soScope [129] jointly infers high-resolution spatial maps using a variational Bayesian inference network. (c) For spatial alignment, STAligner [131] enables coordinate registration of stacked consecutive slices and 3D tissue reconstruction based on MNN and graph autoencoder. (d) For multi-source integration, Seurat v3 [126] integrates scRNA-seq experiments with scATAC-seq based on anchors. GLUE [135] integrates omics-specific autoencoders to model regulatory interactions between omics layers.

hidden patterns that individual datasets might miss. One key motivation for integrating multi-source data is the increasing availability of diverse datasets from various platforms, which makes it possible to build robust models by combining different modalities. This availability creates opportunities for better generalization across different conditions and populations.

Model generalization

Integrating unpaired datasets, such as those derived from different samples or sequencing platforms, often requires aligning independent feature spaces. A central challenge in this context is ensuring model generalization—the ability of models to perform well across heterogeneous datasets with varying structures and distributions.

Seurat v3 [126] integrates diverse single-cell datasets across technologies and modalities by identifying common anchors, which are based on high-variability features. It returns a corrected data matrix for all datasets, enabling joint analysis in a single workflow, while projecting either discrete labels or continuous data for information transfer between reference and query datasets (Fig. 7d). By focusing on shared structure, Seurat v3 enhances the model's ability to generalize to new data sources while minimizing modality-specific noise. Similarly, generative DL models have been leveraged to improve model generalization by capturing complex semantic relationships in multi-source data. For instance, DAVAE [133] integrates large-scale, unpaired data by incorporating a domain-adversarial classifier, which encourages the model to learn features that are invariant across different domains (i.e. datasets or modalities). This reduces modality-specific biases and enables the model to generalize better across heterogeneous data, making it robust to differences in dataset distributions. Another example is scAEGAN [134], which enables data integration and model generalization by learning domain-invariant representations through the use of adversarial training (Wasserstein GANs) and cycle consistency. The cycle consistency loss ensures that the learned mappings between domains preserve the inherent structure of the data, while the Wasserstein GAN framework helps generate samples from distributions that are close to the real data distribution, promoting better alignment across domains. GLUE integrates omic-specific autoencoders with graph-based coupling and adversarial alignment, supporting regulatory inference across unpaired multi-omics datasets. When integrating scRNA-seq and scATAC-seq data, genes and accessible chromatin regions (ATAC peaks) are treated as graph vertices, with edges linking accessible regions to their potential downstream genes. Adversarial multimodal alignment is then performed iteratively, guided by feature embeddings from the graph (Fig. 7d).

Recent benchmark for multi-source data integration has collected 19 methods for data integration on seven benchmark datasets, as detailed by Luecken *et al.* [136], indicating that scANVI [151], Scanorama [94], scVI [105], and scGen [152] perform well on complex integration tasks. After reanalyzing the overall metrics, we observed the relative robustness of scANVI model in preserving biological consistency and batch correction across all datasets. However, it is shown that none of the methods demonstrate consistent optimal performance across all datasets and metrics (Fig. 8).

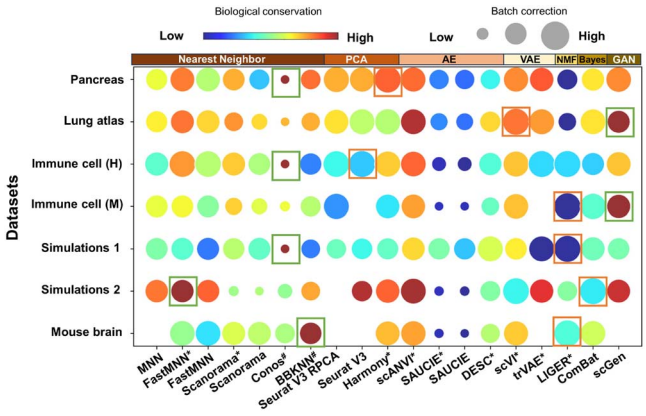

Figure 8. Revisualize the benchmark results for multi-source data integration from seven benchmark datasets [136].
Note. This analysis includes 19 reported methods, including backbones based on nearest neighbors (NN), PCA, AE, VAE, NMF, Bayes, and GAN. All datasets are scRNA-seq data, where Simulation 1 and Simulation 2 are simulated by Splatter [114]. It is shown that none of the methods demonstrate consistent optimal performance across all datasets and metrics (with missing values represented as blanks). We also observe that ScANVI is relatively effective in preserving biological consistency and batch correction across all datasets. The orange rectangle represents the largest point size (batch correction), and the green rectangle indicates the points with the highest color value (biological conservation), closest to red on the color bar. "Biological conservation" refers to preserving key biological features across single-cell data, which can be assessed through both label-based and label-free methods. This includes classical metrics such as neighborhood assessment and cluster matching, along with new metrics for rare cell types, cell-cycle variance, highly variable genes (HVGs), and cellular trajectories before and after integration. "∗" indicates HVGs, "#" indicates graph representation.

Data scarcity

The overall structure of this section is shown in Fig. 9.

Missing data annotation

Single cell data have reached sequencing scales of hundreds of millions. Due to the significant labor and time required in laboratory settings, existing datasets often present challenges in obtaining large-scale biological annotations. In addition, single-cell data contain many complex biological factors, making it difficult to obtain a reasonable and reliable ground truth. For instance, benchmark datasets for analyzing the evolution of single-cell populations require follow-up samples with known evolutionary trajectories and developmental timelines. However, obtaining such samples under experimental conditions is challenging [163]. Moreover, the reliability of model evaluation often depends on high-quality data annotations. For example, since regulatory interactions in databases are aggregated from broad datasets and lack specificity to particular biological systems, it is unreliable to evaluate the performance of gene regulatory network (GRN) inference algorithms. A key technical solution to this problem is the construction of simulation datasets.

It has been extensively employed to evaluate and compare computational methods, concentrate on specific biological features, and establish more precise benchmarks [155]. Cao *et al.* [156] conducted a comprehensive review of various simulation approaches. They proposed a framework for systematic benchmarking, highlighting the ability of these approaches to capture biological signals and higher order interactions. However, most existing methods generate simulated data tailored to specific evaluation objectives, such as clustering or differential gene expression analysis. There are few tools specifically designed to create datasets that are applicable across diverse scenarios.

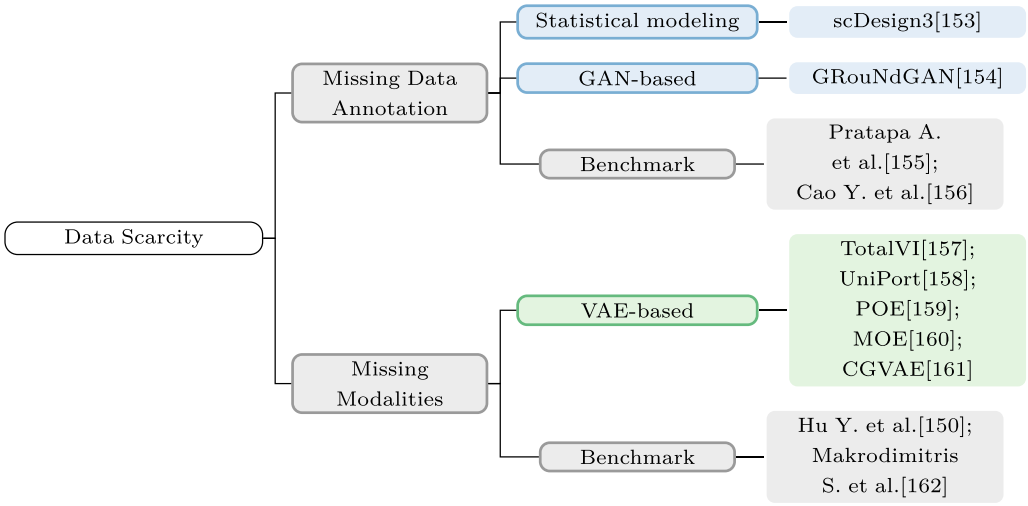

Figure 9. The structure of section "Data Scarcity" and related methods.
Note. The tree chart outlines the challenges related to the scarcity of high-quality single-cell data, emphasizing issues such as missing data annotation and missing modalities.

scDesign3 [153] employs statistical modeling methods to generate single-cell multi-omics data and ST data with known cell proportions. It standardizes generative modeling approaches across various data modalities, rather than focusing on only one modality. Given a cell state covariate x_i (factors such as cell type, cell pseudotime, and cell spatial locations) and experimental design covariates z_i (such as batch effects and experimental conditions), the measurement values Y_{ij} are modeled according to a specific distribution $F_j\left(\cdot \mid \mathbf{X}_i, \mathbf{z}_i; \mu_{ij}, \sigma_{ij}, p_{ij}\right)$. This is formulated as a generalized additive model for location scale and shape (GAMLSS), characterized by its position, proportion, and shape parameters.

The model is parametrically represented, incorporating specific link functions for each feature j's (distribution functions $\theta_j\left(\mu_{ij}\right)$, such as Gaussian (Normal), Bernoulli, Poisson, ZINB) These link functions correspond to the mean parameter μ_{ij}, the scale parameter σ_{ij}(e.g. standard deviation or dispersion), and zero-inflation proportion parameter p_{ij}. For instance, the specific link functions $\theta_j\left(\mu_{ij}\right)$ for features j maps the mean parameter μ_{ij} to the model's linear predictor. This mapping depends on the chosen distribution function and consists of four key components:

$$\theta_j\left(\mu_{ij}\right) = \alpha_{j0} + \alpha_{jb_i} + \alpha_{jc_i} + f_{jc_i}\left(\mathbf{x}_i\right) \tag{2}$$

The specific intercept α_{j0} for feature j represents the mean of feature j in the absence of other influencing factors. The batch effect α_{jb_i} captures the influence of batch b_i on feature j. The conditional effect α_{jc_i} denotes the impact of condition c_i on feature j. The cell state covariates x_i, such as the effects associated with different cell types on feature j, are also considered. In scDesign3, the joint distribution of cellular features is constructed using a marginal cumulative distribution function and a copula model with parameters estimated by a maximum likelihood method. These parameters can be adjusted to generate synthetic data reflecting varying sequencing depths and cell states. Furthermore, scDesign3 is capable of producing ST data based on cell type proportions derived from single-cell sequencing, and simulating realistic ATAC-seq datasets at both the count and read levels. Additionally, it generates multi-omics datasets by integrating separate omics datasets, such as RNA expression or DNA methylation.

In summary, statistical modeling provides a highly interpretable framework for data generation and analysis, and its integration with DL is emerging as a significant trend.

GlouNdGAN [154] is a causal model-based data generation method that allows the generation of realistic simulated data aligned with the underlying principles of GRN. The architecture of GlouNdGAN consists of five sub-networks: Causal Controller, Target Generator, Critic, Labeler, and Anti-labeler. The Causal Controller generates expression values for transcription factors (TFs), while the Target Generators produce expression values for target genes based on the causal GRN framework. The Critic estimates the Wasserstein distance between the generated data and the real data to ensure that the target gene expression is causally related to TF expression. The Labeler predicts TF expression based on generated and real target gene expression, while the Anti-labeler estimates TF expression solely from generated target gene expression. This model pretrains the TF expression generation module and subsequently generates expression values for other genes via the Target Generator. GlouNdGAN allows researchers to simulate interference by manipulating TF expression values during the inference phase. This enables an accurate comparison of gene expression before and after interference while keeping other parameters constant. Additionally, by performing mutation experiments on TF expression for specific cell types, researchers observed alterations in the characteristics of the generated cells. This process helps validate the function of TFs and their relationship to phenotypic labels. This capability positions GlouNdGAN as an ideal tool for *in situ* interference experiments.

As discussed above, simulation data generation serves as a valuable tool for elucidating biological mechanisms in contexts where high-quality data are lacking. It allows the creation of diverse datasets with controllable parameters and facilitates the evaluation of model performance.

Missing modalities

Although genetic information is transferred from DNA to RNA and then to proteins, each modality captures distinct biological information, making it impossible for one modality to substitute for another. It has been demonstrated that multimodal analysis enhances the overall understanding of cellular heterogeneity.

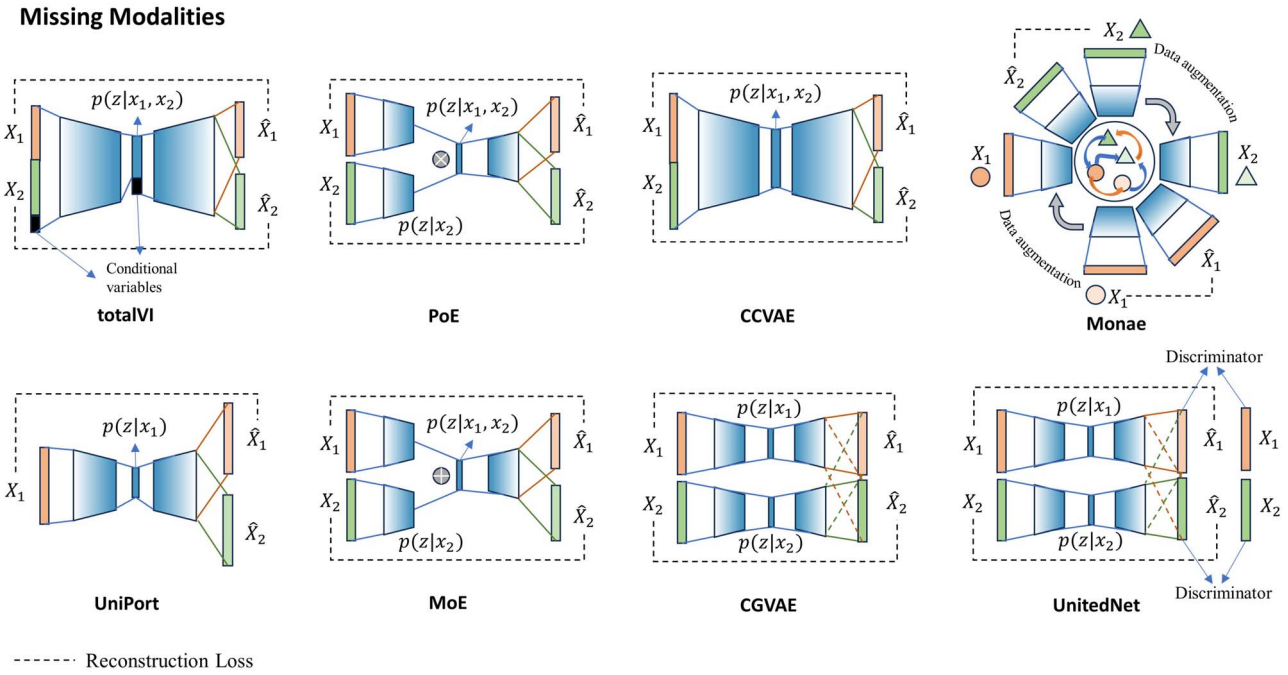

Figure 10. The challenges and typical approaches for data scarcity.
Note. DL-based methods mainly rely on VAE architectures, using either single-modality or dual-modality joint embeddings for feature modeling to learn a shared latent space. We plot the basic architectures of the following VAE-based methods: totalVI [157], UniPort [158], POE [159], MOE [160], CCVAE [164], and CGVAE [161]. Recently, additional strategies have been incorporated into generative models, such as introducing regularized discriminators (UnitedNet [165]) and employing data augmentation strategies (Monae [166]).

However, multi-channel sequencing typically incurs higher costs compared with single-channel sequencing. DL-based solutions commonly rely on VAE architectures that use either single-modality or multi-modality joint embeddings for shared latent space modeling [162]. The difference lies in how the latent variables are modeled (Fig. 10). TotalVI [157] is trained on the joint embeddings of the two modalities with separate reconstruction. UniPort [158] trains a single-modality embedding to reconstruct two different modalities, compelling the encoder to learn features that are predictive of both. In the Product of Experts (POE) model [159], the joint latent variable is derived as a product of each modality's. Unlike POE, Mixture of Experts (MOE) [160] employs the sum of the joint latent variables for data reconstruction. Constrained Graph Variational Autoencoders (CGVAE) [161] learns feature embeddings for each modality individually while applying constraints to ensure that each modality can reconstruct both itself and the other modalities. Based on the benchmark results, Makrodimitris, S. *et al*. [162] concluded that different joint embeddings can be used for different downstream tasks.

Scenarios of modality completion are often related to data sparsity. Monae [166] employs data imputation to perform data denoising and modality completion simultaneously, constrained by a cross-modal prediction loss. Specifically, the embeddings of the same cell type from the same modality, are input into the student latent space projection after data augmentation (i.e. adding a random dropout mask). These embeddings then form positive pairs with the teacher projection. In the feature extraction phase, a graph encoder-decoder reconstruction process extracts embedding features from multiple modalities, and contrastive learning is applied to minimize the spatial distance between embeddings of the same cell type. Therefore, when discriminative information from one modality is lacking, the latent space embeddings of other modalities can be leveraged for reconstruction. UnitedNet [165] combines multimodal ensemble with cross-modal prediction in a multi-task learning framework, trained with cross-modal prediction loss alongside generator and discriminator losses.

We have collected 17 methods, including BABEL [167], CMAE [168], LIGER [117], Seurat [126], cTP-net [169], scArches [170], scMoGNN [171], scVAEIT [172], sciPENN [173], totalVI [157], Generalized Linear Model (GLM), MCIA [174], MOFA [175], CGVAE [161], ccVAE [164], PoE [159], MoE [160], to revisualize their modality prediction performance on four benchmark datasets [162, 176]. The results shown are a reanalysis of previously published data (Fig. 11). Among all the methods, totalVI shows highest cell–cell Pearson correlation coefficient (PCC) on predicted modalities and PoE shows better performance than other VAE-based models.

Data correlation

Understanding the relationship between biological systems and external factors is crucial to gain deeper insights into cellular dynamics and the interactions between cells and their environment. Modeling data correlation involves capturing these complex interactions and dependencies, which are affected by both spatial and temporal variations, as well as biological prior knowledge. The overall structure of this section is shown in Fig. 12.

Modeling spatiotemporal dependencies

Modeling temporal and spatial dependencies is crucial for the analysis of single-cell and ST data, as numerous biological processes exhibit dynamic spatiotemporal correlations. Examples include cell differentiation during development [194], the spatial organization of cells within tissues [195], disease progression pathways [196], and variations in immune responses over time and space. Capturing these features can reveal dynamical shifts in cell states, cell–cell interactions, and complex tissue or disease structures. Spatiotemporal omics data encompass

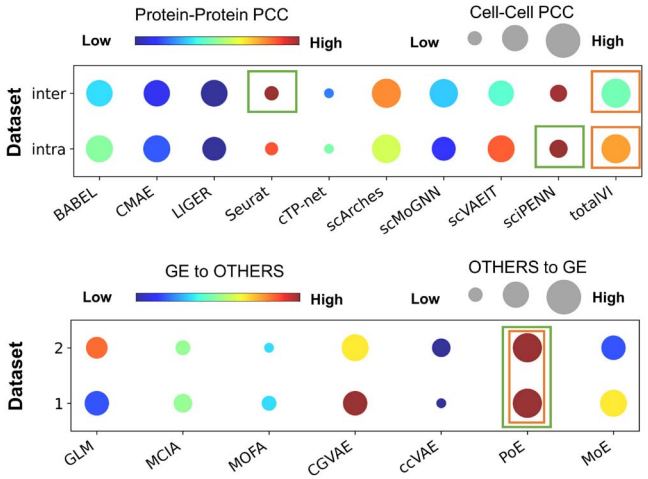

Figure 11. Revisualize the benchmark results for modality prediction from four benchmark datasets [162, 176].
Note. The orange rectangle represents the largest point size, and the green rectangle indicates the points with the highest color value, closest to red on the color bar. In Benchmark 1 (Dataset "inter" and Dataset "intra", the size of the bubbles represents the abundance of a protein (or chromatin accessibility) across two cells, while the color indicates the Pearson correlation coefficients (PCCs) between pairs of proteins. In Benchmark 2 (Dataset 1 and Dataset 2), 'GE to OTHERS' refers to the average accuracy of translation from GE to other modalities, and vice versa. Intra-dataset refers to the use of a single dataset, while inter-dataset involves using two datasets from the same organ or tissue. The results represent the average metric across 23 intra-dataset comparisons and 10 inter-dataset comparisons, respectively. All datasets are publicly available. Dataset 1 and Dataset 2 are all from TCGA Database [162].

longitudinal molecular profiles from patients, molecular profiles across developmental stages, and continuous spatiotemporal omics maps. However, making comparisons across varying scales, biological samples, or conditions remains a challenging task. For example, establishing statistical correlations between samples requires the alignment of temporal and spatial coordinates across individuals or biological scales. Current approaches mainly rely on regression-based and graph-based models [197].

Among regression-based models, Gaussian process-based probabilistic models are widely used [177–179]. These models are effective in capturing trends of continuous variability for both time series and spatially distributed data. MEFISTO [180] leverages the Gaussian process to model latent factors by incorporating temporal and spatial information, as well as grouped data, into the covariates within the Gaussian kernel. The covariance function consists of two components: one that describes relationships across different groups (e.g. sample sets or experimental conditions), and another that captures smooth variations such as spatial locations or time points. This dual structure allows Gaussian processes (GP) to account for both inter-group heterogeneity and intra-group covariate variation. By ensuring that samples located closer together in the covariate space share more similar latent factors, the Gaussian process effectively models the continuity of relations between data points. However, the main limitation of GP modeling is its cubic time complexity, making it infeasible for large datasets. Methods such as inducing points and low-precision arithmetic can reduce this complexity. For example, SpaVAE [198] derives the ELBO of VAEs on a mini-batch to perform GP regression, transforming the typical $O(N^3)$ complexity into $O(bp^2 + p^3)$, where b is the mini-batch size and p is the number of inducing points.

Another approach for jointly modeling omics latent features z involves the use of graph models. Markov random fields (MRF) are undirected graphical models that assume the distribution of each node depends only on the labels of its neighboring nodes. Compared with nonparametric regression models like GP, MRF offer greater computational efficiency, as they do not require inference of the complete covariance structure across all samples. Qian Zhu *et al.* [181] proposed a Markovian property for spatial patterns, which constrains the correlations between neighboring nodes. By assuming that labels of neighboring cells, including gene expression states or cell types, exhibit a degree of similarity, the joint probability distribution over the spatial domain can be factorized into a product of local neighborhood probability distributions. The probability distribution for the label c_i associated with the current node s_i is jointly modeled using both its neighboring nodes' labels C_{Ni} and its own gene expression x_i:

$$P\left(c_i \mid s_i, x_i, c_{N_i}\right) = \frac{1}{Z} P\left(x_i \mid c_i, s_i\right) P\left(c_i \mid s_i, c_{N_i}\right) \tag{3}$$

Giotto [182] utilizes MRF with conditional probability distributions, such as Gaussian or Poisson, to enhance spatial clustering. This approach effectively captures smooth and continuous expression changes, thereby facilitating the identification of spatially structured cell populations.

DL-based graph frameworks are increasingly employed to explore cell–cell interactions, including recurrent neural network (RNN), CNN, and GNN. In GNNs, cells are represented as nodes, with edges denoting potential interactions, such as those between ligand-receptor pairs. This approach effectively integrates spatial data and gene expression profiles to reveal interaction patterns. For instance, DeepLinc [183] posits that neighboring cells are more likely to interact than cells further apart (Fig. 13a). It constructs a cell adjacency graph based on the physical distance between cells. The model then learns embedding features that reflect the likelihood of interactions by aggregating information from each cell and its neighbors. Using variational graph autoencoders (VGAEs) and adversarial networks, DeepLinc employs unsupervised learning techniques to uncover intrinsic associations between the cell adjacency graph and gene expression profiles, thereby reconstructing interaction networks. NCEM [184] utilizes GNN to reconstruct gene expression vectors from cell type labels and niche composition, which are represented through graph-level predictors and adjacency matrices derived from spatial proximity. While NCEM incorporates a linear model for mathematical interpretability, experimental results demonstrate that its nonlinear variant significantly outperforms the linear one. Different from GNNs, CNNs [185] aggregate features from neighboring regions in images through convolution operations. RNNs [186], on the other hand, propagate information from adjacent spatial points using recurrent connections.

Overall, DL frameworks exhibit considerable flexibility in analyzing spatiotemporal omics data.

Modeling prior knowledge

Single-cell data is characterized by sparse features, multi-source heterogeneity, and lack of high-quality labels, making it unreliable to draw experimental conclusions solely from the observed data. However, the incorporation of prior biological knowledge such as gene regulatory networks, cell type characteristics, and developmental trajectories, can significantly enhance the accuracy and interpretability of the analysis. Nonetheless, effectively

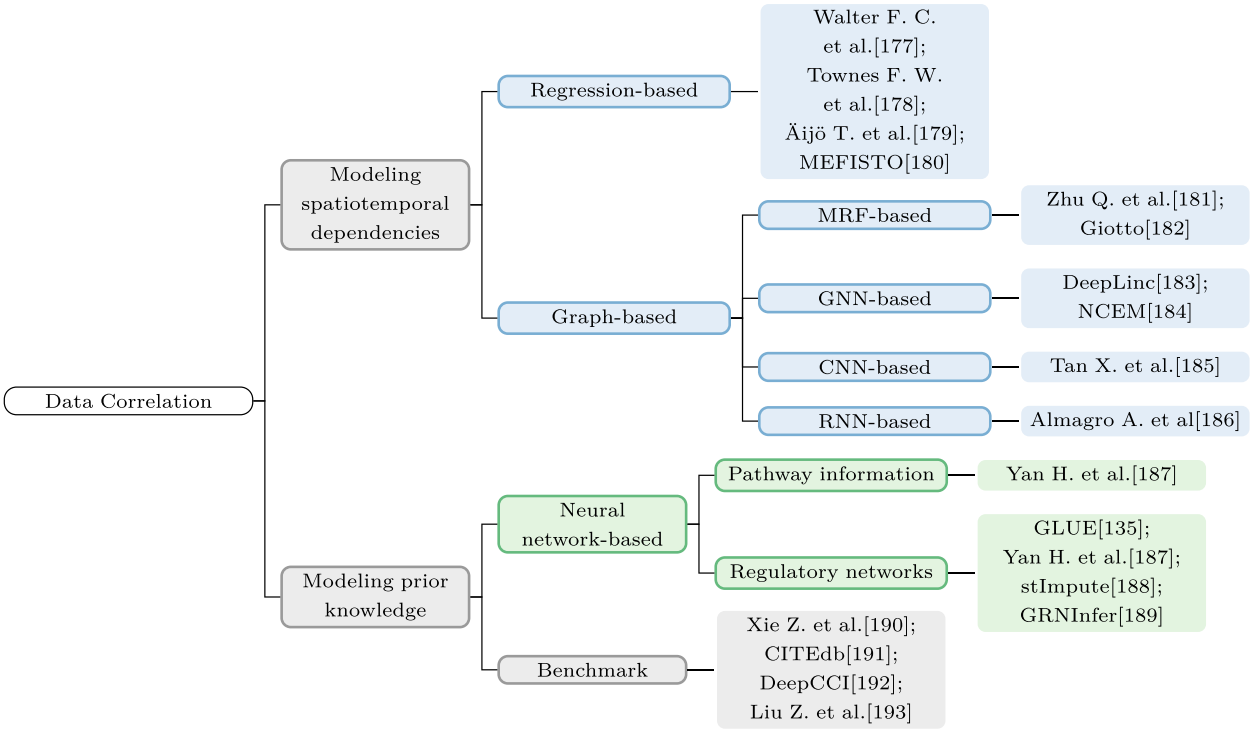

Figure 12. The structure of section "Data correlation" and related methods.
Note. The tree chart outlines the challenges related to data correlation, emphasizing challenges such as modeling spatiotemporal dependencies and prior knowledge.

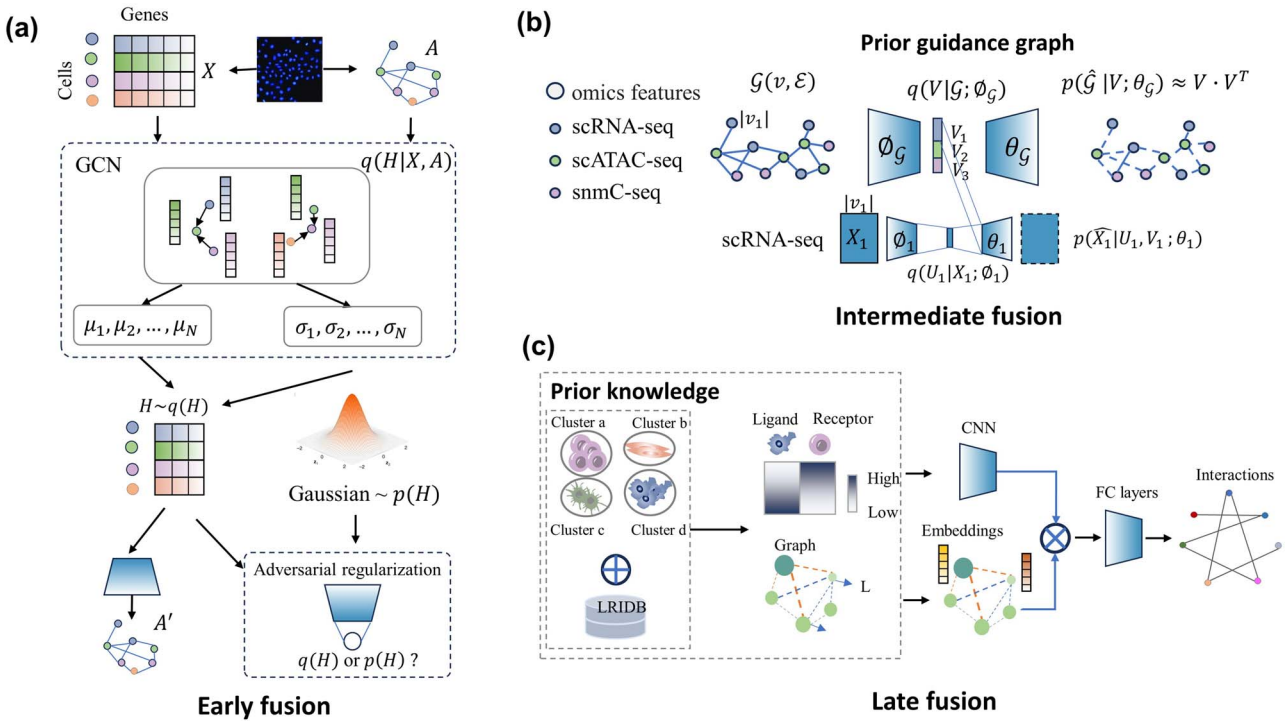

Figure 13. The challenges and typical approaches for data correlation.
Note. DL methods for integrating prior knowledge and modeling data correlation typically use three fusion strategies: early, intermediate, and late fusion. (a) For early fusion, DeepLinc [183] constructs a cell adjacency graph based on the physical distance between cells and learns graph embedding features. (b) For intermediate fusion, GLUE [135] fuses nodes representation features (genes or accessible chromatin regions) from each modality. (c) For late fusion, DeepCCI [192] uses the LRIDB database to identify receptors and predicts interactions by combining ResNet and GCN outputs.

integrating prior knowledge while avoiding potential biases and overfitting remains a challenging task.

Prior knowledge mainly refers to pathway information and regulatory networks [187] obtained from databases. Pathway information concerns to molecular interactions and biochemical reactions that drive specific biological processes, such as signal transduction, metabolism, and cellular activity. This information aids in elucidating how cells respond to external stimuli or internal changes. In the context of single-cell and multi-omics analyses, pathway information is used to infer gene-gene interactions, characterize cell types, and determine cell states. It is typically sourced from well-established databases such as KEGG [199], Reactome [200], and WikiPathways [201]. Regulatory networks involve the interactions among genes, TFs, proteins, and other biomolecules that regulate gene expression and cellular function. These networks are commonly represented as graph where nodes denote biomolecules (e.g. genes or proteins) while edges indicate regulatory relationships (e.g. activation or inhibition). Databases such as STRING [202] and GeneMANIA [203] provide valuable insights into protein-protein interactions and gene-gene interactions, respectively. Regulatory networks facilitate the understanding of the mechanisms governing gene expression regulation, the identification of key regulatory factors, and the revelation of cell-specific patterns in gene expression.

GLUE [135] integrates multi-omics data through a guidance graph, where nodes represent features from various modalities, such as genes in scRNA-seq data and accessible chromatin regions in ATAC-seq data(Fig. 13b). Graphs establish connections between ATAC peaks and RNA genes based on overlapping gene bodies or promoter regions. A variational posterior is employed to reconstruct the guidance map and its latent space is used as a prior for multi-omics data reconstruction. The decoder computes an inner product of feature and cell embeddings to ensure consistent embedding directions across different modalities. Hongxi Yan *et al.* [187] aggregate gene features within the same biological pathway to obtain pathway-level features for predictive modeling. They utilize the KEGG database together with an ensemble gradient method to identify key pathways, which significantly enhances model interpretability.

Some methods use prior knowledge from existing databases to initialize edge features in gene-gene interaction networks. For instance, DeepCCI [192] uses the constructed LRIDB database to define receptors and establishes interaction networks between cell clusters based on known ligand-receptor (L–R) pairs, predicting interactions by a combination of ResNet and graph convolutional network (GCN) outputs (Fig. 13c). stImpute [188] leverages the ESM-2 protein language model to embed proteins and constructs a network of gene relationships using cosine similarity. GRNInfer [189] incorporates gene regulatory relationships from RegNetwork [204] as prior information to construct a gene graph network.

Furthermore, we have collected 10 methods, including CellChat [205], CellPhoneDB [206], iTALK [207], LIANA [208], NATMI [209], scMLnet [210], SingleCellSignalR [211], Connectome [212], CytoTalk [213], and CellCall [214]. These methods leverage existing L-R pair knowledge to infer cell–cell communication, and we reanalyze their cell–cell interaction prediction performance on five benchmark datasets [190–193]. Among all the methods, CellPhoneDB ranks among the top across all benchmark datasets, demonstrating the robustness of extracting cellular context (Fig. 14).

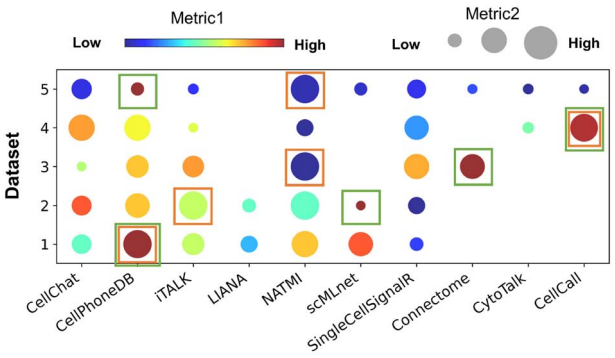

Figure 14. Revisualize the benchmark results for cell–cell interactions from five benchmark datasets [190–193].
Note. While none of the methods demonstrate consistent optimal performance across all datasets, we find that CellPhoneDB ranks among the top in all benchmarks, demonstrating its relative robustness. "Metric 1 and Metric 2 are both accuracy metrics for cell–cell interactions. In Benchmark 1 (Datasets 1 and 2), 'Metric 1' refers to precision, and 'Metric 2' refers to F1 score. In Benchmark 2 (Dataset 3), 'Metric 1' refers to the sum of communication scores, and 'Metric 2' refers to the count of active LR pairs. In Benchmark 3 (Dataset 4), 'Metric 1' refers to AUC, and 'Metric 2' refers to precision. In Benchmark 4 (Dataset 5), 'Metric 1' refers to the distance enrichment score, and 'Metric 2' refers to F1 score. The orange rectangle represents the largest point size ("Metric2"), while the green rectangle indicates the points with the highest color value ("Metric1"), closest to red on the color bar. The blanks represent missing performance results. All the datasets are publicly available scRNA-seq data.

Conclusion and future perspectives

We provide a systematic review of challenges in single-cell data analysis and advances in DL methods. Many studies focus only on partial comparisons using specific datasets, often overlooking the performance of the model in other application scenarios. To address this, we have compiled new statistics by aggregating results from the latest benchmarks (presented in Figs 5, 8, 11, 14), offering a clear reference for selecting the appropriate model to compare on specific tasks for new algorithms developed later.

Innovative AI method for single-cell and ST data analysis

Recently, foundational models have emerged as a focus of single-cell omics. By leveraging self-supervised pretraining on large unlabeled scRNA-seq datasets, these models capture complex features and patterns, producing unified representations that can be fine-tuned for specific downstream tasks [215–218]. For instance, SCimilarity [219] enables rapid querying of cell states for cell type annotation. scMulan [220] converts single-cell transcriptomic data along with rich metadata (e.g. cell type, spatial context, and temporal aspects) into "cell sentences" (c-sentences), achieving superior performance in tasks like zero-shot cell annotation and batch correction. scGPT [218], which is based on the GPT architecture, employs self-supervised pretraining with condition tokens to model gene interactions within cells, incorporating cell type labels for tasks such as cell type prediction, and applies a "binning" technique to ensure semantic alignment across diverse datasets.

Future advancements in large language models, such as OpenAI's O1(https://openai.com/o1/), and agent-based methods, are expected to further enhance single-cell and ST analysis. O1 leverages large-scale reinforcement learning algorithms to achieve chain-of-thought reasoning, thereby improving inference accuracy. Agent-based approaches, such as ReAct [221], integrate

real-time observations to guide decision-making, facilitating more efficient and proactive error correction. These models offer considerable potential for providing interpretable inferences across a variety of downstream tasks in single-cell analysis.

However, in contrast to parametric modeling approaches, end-to-end networks often operate as "black boxes," providing limited interpretability of the inference processes. Moreover, these networks are typically trained on specific datasets and lack dynamic updates, which constrains their generalizability and robustness on unseen or uncertain data. For scientific scenarios, it is crucial to strike a balance between interpretability and accuracy. As indicated by previous studies, it may be feasible to enhance interpretability by establishing connections between prior models and observed outcomes through interpretable parametric rules.

Benchmark datasets and evaluation metrics

Relevant benchmarks have been established for various stages of single-cell sequencing data analysis workflows, including imputation [103], cell identification [222], clustering [223], gene regulatory networks [224], cell–cell interactions [225], and multi-omics data integration [226]. However, several studies have mentioned that the datasets and evaluation metrics employed in these benchmarks do not accurately reflect the strengths and weaknesses of contemporary algorithms [155]. Moreover, as large-scale sequencing data continues to evolve, algorithms that previously demonstrated strong performance may no longer be applicable in different application contexts [227]. These datasets may exhibit increased heterogeneity due to samples derived from diverse sequencing platforms or continuous-time and continuous-space settings.

Current evaluation metrics mainly emphasize performance metrics such as accuracy, AUC, and RMSE, with a limited focus on biological relevance. To ensure the biological validity of model predictions, systematic validation through *in vitro* experiments is essential. For example, trends in gene expression, molecular properties, or cellular behavior can be compared with experimental results to confirm biological significance. Therefore, it is important to establish benchmarks that provide a more comprehensive and objective assessment of the generalizability and applicability of algorithms. Data simulations that generate benchmark datasets based on well-defined rules can provide a diverse array of labeled data for evaluation. This approach has already been applied to tasks such as cell identity recognition and modeling of gene regulatory networks [153, 154], highlighting its potential as a robust tool for evaluating model performance.

Application of DL in practical scenarios

Here, we summarize the applications of single-cell and ST in biology, medicine, and clinical practice, providing an overview of the background for DL applications in these fields.

In biology, single-cell and ST focus on embryonic [228], tissue [229], and organ development [230]. These techniques facilitate the identification and classification of various cell types and lineages, while providing insights into the evolution of cell populations throughout organogenesis.

In precision medicine, the analysis of single-cell transcriptomic data is critical for investigating disease heterogeneity [231], identifying distinct subclones within diseases [232], discovering critical disease biomarkers [233], characterizing interactions between normal and diseased cells [234], elucidating relevant signaling pathways [235], and predicting resistance [236]. Single-cell transcriptomics facilitates the construction of comprehensive cellular maps that significantly enhance the discovery of novel biomarkers and therapeutic targets [237]. In additon, scRNA-seq has demonstrated significant potential for improving patient outcomes and accelerating the development of personalized therapies [238, 239]. Moreover, the integration of multi-omics data, such as genomic, transcriptomic, proteomic, and metabolomic information, plays a pivotal role in enhancing the depth and accuracy of disease modeling. By incorporating diverse data hierarchies, multi-omics approaches offer a comprehensive view of biological processes, providing deeper insights at the cellular, tissue, and organ levels. This, in turn, facilitates more accurate predictions of disease progression, treatment response [240], and the identification of novel therapeutic targets [241].

In clinical applications, scRNA-seq plays a crucial role in characterizing patient-specific features [242]. It assists in the identification of biomarkers for patient stratification [243], elucidates the underlying mechanisms of drug action and resistance [244], supports the development of personalized treatment strategies, and enables monitoring of drug response and disease progression [245].

Key Points

- This review discusses four major challenges and related deep learning approaches in single-cell and spatial transcriptomics data analysis.
- This review curates 21 datasets from nine benchmarks covering 58 computational methods and provides insights into selecting the most appropriate approach for a specific scenario.
- This review outlines three future research directions regarding data, methods, and applications for single-cell and spatial omics data analysis.

Author Contributions

Shuang Ge and Zhixiang Ren collected and reviewed literature. Shuang Ge, Shuqing Sun, Huan Xu, Qiang Cheng and Zhixiang Ren drafted the manuscript. All authors read and approved the final manuscript.

Conflict of interest: None declared.

Data availability

Data supporting this study are included within the article.

References

1. Sun F, Li H, Sun D. *et al*. Single-cell omics: experimental workflow, data analyses and applications. *Sci China Life Sci* 2024;**68**:5–102.
2. Fan J, Slowikowski K, Zhang F. Single-cell transcriptomics in cancer: computational challenges and opportunities. *Exp Mol Med* 2020;**52**:1452–65. https://doi.org/10.1038/s12276-020-0422-0
3. Anonymous. Method of the year 2013. *Nat Methods* 2014;**11**:1. https://doi.org/10.1038/nmeth.2801
4. Li H, Hsieh K, Wong PK. *et al*. Single-cell pathogen diagnostics for combating antibiotic resistance. *Nat. Rev. Methods Primers* 2023;**3**:6.

5. Wang L, Mo S, Li X. *et al*. Single-cell rna-seq reveals the immune escape and drug resistance mechanisms of mantle cell lymphoma. *Cancer Biol Med* 2020;**17**:726. https://doi.org/10.20892/j.issn.2095-3941.2020.0073
6. Keller L, Pantel K. Unravelling tumour heterogeneity by single-cell profiling of circulating tumour cells. *Nat Rev Cancer* 2019;**19**:553–67. https://doi.org/10.1038/s41568-019-0180-2
7. Zeng Q, Mousa M, Nadukkandy AS. *et al*. Understanding tumour endothelial cell heterogeneity and function from single-cell omics. *Nat Rev Cancer* 2023;**23**:544–64. https://doi.org/10.1038/s41568-023-00591-5
8. Zhang S, Fang W, Zhou S. *et al*. Single cell transcriptomic analyses implicate an immunosuppressive tumor microenvironment in pancreatic cancer liver metastasis. *Nat Commun* 2023;**14**:5123.
9. Wang T, Peng J, Fan J. *et al*. Single-cell multi-omics profiling of human preimplantation embryos identifies cytoskeletal defects during embryonic arrest. *Nat Cell Biol* 2024;**26**:263–77. https://doi.org/10.1038/s41556-023-01328-0
10. Marx V. Method of the year: spatially resolved transcriptomics. *Nat Methods* 2021;**18**:9–14. https://doi.org/10.1038/s41592-020-01033-y
11. Myung Y, de Sá AGC, Ascher DB. Deep-pk: deep learning for small molecule pharmacokinetic and toxicity prediction. *Nucleic Acids Res* 2024;**52**:469–75. https://doi.org/10.1093/nar/gkae254
12. Bennett WFD, He S, Bilodeau CL. *et al*. Predicting small molecule transfer free energies by combining molecular dynamics simulations and deep learning. *J Chem Inf Model* 2020;**60**:5375–81. https://doi.org/10.1021/acs.jcim.0c00318
13. Chowdhury R, Bouatta N, Biswas S. *et al*. Single-sequence protein structure prediction using a language model and deep learning. *Nat Biotechnol* 2022;**40**:1617–23. https://doi.org/10.1038/s41587-022-01432-w
14. Pearce R, Zhang Y. Deep learning techniques have significantly impacted protein structure prediction and protein design. *Curr Opin Struct Biol* 2021;**68**:194–207. https://doi.org/10.1016/j.sbi.2021.01.007
15. Catacutan DB, Alexander J, Arnold A. *et al*. Machine learning in preclinical drug discovery. *Nat Chem Biol* 2024;**20**:960–73. https://doi.org/10.1038/s41589-024-01679-1
16. Erfanian N, Heydari AA, Feriz AM. *et al*. Deep learning applications in single-cell genomics and transcriptomics data analysis. *Biomed Pharmacother* 2023;**165**:115077. https://doi.org/10.1016/j.biopha.2023.115077
17. Bao S, Li K, Yan C. *et al*. Deep learning-based advances and applications for single-cell rna-sequencing data analysis. *Brief Bioinform* 2022;**23**:bbab473.
18. Flores M, Liu Z, Tinghe Zhang M. *et al*. Deep learning tackles single-cell analysis—a survey of deep learning for scrna-seq analysis. *Brief Bioinform* 2022;**23**:bbab531.
19. Brendel M, Chang S, Bai Z. *et al*. Application of deep learning on single-cell rna sequencing data analysis: a review. *Genomics Proteomics Bioinf*. 2022;**20**:814–35. https://doi.org/10.1016/j.gpb.2022.11.011
20. Taud H, Mas J-F. Multilayer perceptron (mlp). *Geomatic approaches for modeling land change scenarios* 2018;451–5.
21. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. *Science* 2006;**313**:504–7.
22. Goodfellow I, Pouget-Abadie J, Mirza M. *et al*. Generative adversarial networks. *Communications of the ACM* 2020;**63**:139–44. https://doi.org/10.1145/3422622
23. LeCun Y, Boser B, Denker JS. *et al*. Backpropagation applied to handwritten zip code recognition. *Neural Comput* 1989;**1**:541–51. https://doi.org/10.1162/neco.1989.1.4.541
24. Scarselli F, Marco Gori A, Tsoi C. *et al*. The graph neural network model. *IEEE Trans Neural Netw* 2008;**20**:61–80.
25. Argelaguet R, Cuomo ASE, Stegle O. *et al*. Computational principles and challenges in single-cell data integration. *Nat Biotechnol* 2021;**39**:1202–15. https://doi.org/10.1038/s41587-021-00895-7
26. Ma A, McDermaid A, Jennifer X. *et al*. Integrative methods and practical challenges for single-cell multi-omics. *Trends Biotechnol* 2020;**38**:1007–22. https://doi.org/10.1016/j.tibtech.2020.02.013
27. Molho D, Ding J, Tang W. *et al*. Deep learning in single-cell analysis. *ACM Transactions on Intelligent Systems and Technology* 2024;**15**:1–62. https://doi.org/10.1145/3641284
28. Zahedi R, Ghamsari R, Argha A. *et al*. Deep learning in spatially resolved transcriptfomics: a comprehensive technical view. *Brief Bioinform* 2024;**25**:bbae082.
29. Ali, Heydari A, Sindi SS. Deep learning in spatial transcriptomics: learning from the next next-generation sequencing. *Biophysics Reviews* 2023;**4**:011306.
30. Ma Q, Dong X. Deep learning shapes single-cell data analysis. *Nat Rev Mol Cell Biol* 2022;**23**:303–4. https://doi.org/10.1038/s41580-022-00466-x
31. Tang F, Barbacioru C, Wang Y. *et al*. Mrna-seq whole-transcriptome analysis of a single cell. *Nat Methods* 2009;**6**:377–82. https://doi.org/10.1038/nmeth.1315
32. Herzenberg LA, Sweet RG, Herzenberg LA. Fluorescence-activated cell sorting. *Sci Am* 1976;**234**:108–18. https://doi.org/10.1038/scientificamerican0376-108
33. Hagemann-Jensen M, Ziegenhain C, Chen P. *et al*. Single-cell rna counting at allele and isoform resolution using smart-seq3. *Nat Biotechnol* 2020;**38**:708–14. https://doi.org/10.1038/s41587-020-0497-0
34. Salmen F, De Jonghe J, Kaminski TS. *et al*. High-throughput total rna sequencing in single cells using vasa-seq. *Nat Biotechnol* 2022;**40**:1780–93. https://doi.org/10.1038/s41587-022-01361-8
35. Hahaut V, Pavlinic D, Carbone W. *et al*. Fast and highly sensitive full-length single-cell rna sequencing using flash-seq. *Nat Biotechnol* 2022;**40**:1447–51. https://doi.org/10.1038/s41587-022-01312-3
36. Miltenyi S, Müller W, Weichel W. *et al*. High gradient magnetic cell separation with macs. *Cytometry: The Journal of the International Society for Analytical Cytology* 1990;**11**:231–8. https://doi.org/10.1002/cyto.990110203
37. Gossett DR, Weaver WM, Mach AJ. *et al*. Label-free cell separation and sorting in microfluidic systems. *Anal Bioanal Chem* 2010;**397**:3249–67. https://doi.org/10.1007/s00216-010-3721-9
38. Klein AM, Mazutis L, Akartuna I. *et al*. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* 2015;**161**:1187–201. https://doi.org/10.1016/j.cell.2015.04.044
39. Macosko EZ, Basu A, Satija R. *et al*. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 2015;**161**:1202–14. https://doi.org/10.1016/j.cell.2015.05.002
40. Zheng GXY, Terry JM, Belgrader P. *et al*. Massively parallel digital transcriptional profiling of single cells. *Nat Commun* 2017;**8**:14049.
41. Christina Fan H, Fu GK, Fodor SPA. Combinatorial labeling of single cells for gene expression cytometry. *Science* 2015;**347**:1258367. https://doi.org/10.1126/science.1258367

42. Gierahn TM, Wadsworth MH, Hughes TK. *et al*. Seq-well: portable, low-cost rna sequencing of single cells at high throughput. *Nat Methods* 2017;**14**:395–8. https://doi.org/10.1038/nmeth.4179
43. Han X, Wang R, Zhou Y. *et al*. Mapping the mouse cell atlas by microwell-seq. *Cell* 2018;**172**:1091–107. https://doi.org/10.1016/j.cell.2018.02.001
44. Junker JP, Noel ES, Guryev V. *et al*. Genome-wide rna tomography in the zebrafish embryo. *Cell* 2014;**159**:662–75. https://doi.org/10.1016/j.cell.2014.09.038
45. Schede HH, Schneider CG, Stergiadou J. *et al*. Spatial tissue profiling by imaging-free molecular tomography. *Nat Biotechnol* 2021;**39**:968–77. https://doi.org/10.1038/s41587-021-00879-7
46. Chen J, Suo S, Tam PPL. *et al*. Spatial transcriptomic analysis of cryosectioned tissue samples with geo-seq. *Nat Protoc* 2017;**12**:566–80. https://doi.org/10.1038/nprot.2017.003
47. Giladi A, Cohen M, Medaglia C. *et al*. Dissecting cellular crosstalk by sequencing physically interacting cells. *Nat Biotechnol* 2020;**38**:629–37. https://doi.org/10.1038/s41587-020-0442-2
48. Lovatt D, Ruble BK, Lee J. *et al*. Transcriptome in vivo analysis (tiva) of spatially defined single cells in live tissue. *Nat Methods* 2014;**11**:190–6. https://doi.org/10.1038/nmeth.2804
49. Medaglia C, Giladi A, Stoler-Barak L. *et al*. Spatial reconstruction of immune niches by combining photoactivatable reporters and scrna-seq. *Science* 2017;**358**:1622–6. https://doi.org/10.1126/science.aao4277
50. Ståhl PL, Salmén F, Vickovic S. *et al*. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* 2016;**353**:78–82. https://doi.org/10.1126/science.aaf2403
51. Stickels RR, Murray E, Kumar P. *et al*. Highly sensitive spatial transcriptomics at near-cellular resolution with slide-seqv2. *Nat Biotechnol* 2021;**39**:313–9. https://doi.org/10.1038/s41587-020-0739-1
52. Vickovic S, Eraslan G, Salmén F. *et al*. High-definition spatial transcriptomics for in situ tissue profiling. *Nat Methods* 2019;**16**:987–90. https://doi.org/10.1038/s41592-019-0548-y
53. Chen A, Liao S, Cheng M. *et al*. Spatiotemporal transcriptomic atlas of mouse organogenesis using dna nanoball-patterned arrays. *Cell* 2022;**185**:1777–92.
54. Cho C-S, Xi J, Si Y. *et al*. Microscopic examination of spatial transcriptome using seq-scope. *Cell* 2021;**184**:3559–72. https://doi.org/10.1016/j.cell.2021.05.010
55. Kim Y, Cheng W, Cho C-S. *et al*. Seq-scope: repurposing illumina sequencing flow cells for high-resolution spatial transcriptomics. *Nat Protoc* 2024;**20**:643–89.
56. Cao J, Zheng Z, Sun D. *et al*. Decoder-seq enhances mrna capture efficiency in spatial rna sequencing. *Nat Biotechnol* 2024;**42**:1735–46.
57. Schott M, León-Periñán D, Splendiani E. *et al*. Open-st: high-resolution spatial transcriptomics in 3d. *Cell* 2024;**187**:3953–72. https://doi.org/10.1016/j.cell.2024.05.055
58. Chen J, Larsson L, Swarbrick A. *et al*. Spatial landscapes of cancers: insights and opportunities. *Nat Rev Clin Oncol* 2024;**21**:660–74. https://doi.org/10.1038/s41571-024-00926-7
59. Eng C-HL, Lawson M, Zhu Q. *et al*. Transcriptome-scale super-resolved imaging in tissues by rna seqfish+. *Nature* 2019;**568**:235–9. https://doi.org/10.1038/s41586-019-1049-y
60. Lubeck E, Coskun AF, Zhiyentayev T. *et al*. Single-cell in situ rna profiling by sequential hybridization. *Nat Methods* 2014;**11**:360–1. https://doi.org/10.1038/nmeth.2892
61. Xia C, Fan J, Emanuel G. *et al*. Spatial transcriptome profiling by merfish reveals subcellular rna compartmentalization and cell cycle-dependent gene expression. *Proc Natl Acad Sci* 2019;**116**:19490–9. https://doi.org/10.1073/pnas.1912459116
62. Wang X, Allen WE, Wright MA. *et al*. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* 2018;**361**:eaat5691.
63. Alon S, Goodwin DR, Sinha A. *et al*. Expansion sequencing: spatially precise in situ transcriptomics in intact biological systems. *Science* 2021;**371**:eaax2656.
64. Barrett T, Wilhite SE, Ledoux P. *et al*. Ncbi geo: archive for functional genomics data sets—update. *Nucleic Acids Res* 2012;**41**:D991–5. https://doi.org/10.1093/nar/gks1193
65. Regev A, Teichmann SA, Lander ES. *et al*. The human cell atlas. *Elife* 2017;**6**:e27041.
66. Tarhan L, Bistline J, Chang J. *et al*. Single cell portal: An interactive home for single-cell genomics data. bioRxiv. 2023;548886. https://doi.org/10.1101/2023.07.13.548886
67. Megill C, Martin B, Weaver C. *et al*. Et al., Cellxgene: a performant, scalable exploration platform for high dimensional sparse matrices. bioRxiv. 2021;438318. https://doi.org/10.1101/2021.04.05.438318
68. Franzén O, Gan L-M, Björkegren JLM. Panglaodb: a web server for exploration of mouse and human single-cell rna sequencing data. *Database* 2019;**2019**:baz046.
69. Zeng J, Zhang Y, Shang Y. *et al*. Cancerscem: a database of single-cell expression map across various human cancers. *Nucleic Acids Res* 2022;**50**:D1147–55. https://doi.org/10.1093/nar/gkab905
70. Moreno P, Fexova S, George N. *et al*. Expression atlas update: gene and protein expression in multiple species. *Nucleic Acids Res* 2022;**50**:D129–40. https://doi.org/10.1093/nar/gkab1030
71. Shi X, Zhiguang Y, Ren P. *et al*. Husch: An integrated single-cell transcriptome atlas for human tissue gene expression visualization and analyses. *Nucleic Acids Res* 2023;**51**:D1029–37. https://doi.org/10.1093/nar/gkac1001
72. Li M, Zhang X, Ang KS. *et al*. Disco: a database of deeply integrated human single-cell omics data. *Nucleic Acids Res* 2022;**50**:D596–602. https://doi.org/10.1093/nar/gkab1020
73. Li W, Cowley A, Uludag M. *et al*. The embl-ebi bioinformatics web and programmatic tools framework. *Nucleic Acids Res* 2015;**43**:W580–4. https://doi.org/10.1093/nar/gkv279
74. Chen S, Luo Y, Gao H. *et al*. Heca: the cell-centric assembly of a cell atlas. *Iscience* 2022;**25**:104318. https://doi.org/10.1016/j.isci.2022.104318
75. Fan Z, Chen R, Chen X. Spatialdb: a database for spatially resolved transcriptomes. *Nucleic Acids Res* 2020;**48**:D233–7. https://doi.org/10.1093/nar/gkz934
76. Zheng Y, Chen Y, Ding X. *et al*. Aquila: a spatial omics database and analysis platform. *Nucleic Acids Res* 2023;**51**:D827–34. https://doi.org/10.1093/nar/gkac874
77. Li Y, Dennis S, Hutch MR. *et al*. Soar elucidates disease mechanisms and empowers drug discovery through spatial transcriptomics. bioRxiv. 2022;488596. https://doi.org/10.1101/2022.04.17.488596
78. Zhicheng X, Wang W, Yang T. *et al*. Stomicsdb: a comprehensive database for spatial transcriptomics data sharing, analysis and visualization. *Nucleic Acids Res* 2024;**52**:D1053–61. https://doi.org/10.1093/nar/gkad933
79. Fan Z, Luo Y, Huifen L. *et al*. Spascer: spatial transcriptomics annotation at single-cell resolution. *Nucleic Acids Res* 2023;**51**:D1138–49. https://doi.org/10.1093/nar/gkac889
80. Yuan Z, Pan W, Zhao X. *et al*. Sodb facilitates comprehensive exploration of spatial omics data. *Nat Methods* 2023;**20**:387–99. https://doi.org/10.1038/s41592-023-01773-7

81. Zhou W, Minghai S, Jiang T. *et al*. Sorc: An integrated spatial omics resource in cancer. *Nucleic Acids Res* 2024;**52**:D1429–37. https://doi.org/10.1093/nar/gkad820
82. Maćkiewicz A, Ratajczak W. Principal components analysis (pca). *Comput Geosci* 1993;**19**:303–42. https://doi.org/10.1016/0098-3004(93)90090-R
83. Boileau P, Hejazi NS, Dudoit S. Exploring high-dimensional biological data with sparse contrastive principal component analysis. *Bioinformatics* 2020;**36**:3422–30. https://doi.org/10.1093/bioinformatics/btaa176
84. Lee H, Han B. Fastrna: An efficient solution for pca of single-cell rna-sequencing data based on a batch-accounting count model. *The American Journal of Human Genetics* 2022;**109**:1974–85. https://doi.org/10.1016/j.ajhg.2022.09.008
85. Wang Y, Chen L, Jo J. *et al*. Joint t-sne for comparable projections of multiple high-dimensional datasets. *IEEE Trans Vis Comput Graph* 2021;**28**:623–32.
86. Linderman GC, Rachh M, Hoskins JG. *et al*. Fast interpolation-based t-sne for improved visualization of single-cell rna-seq data. *Nat Methods* 2019;**16**:243–5. https://doi.org/10.1038/s41592-018-0308-4
87. Kobak D, Berens P. The art of using t-sne for single-cell transcriptomics. *Nat Commun* 2019;**10**:5416.
88. Becht E, McInnes L, Healy J. *et al*. Dimensionality reduction for visualizing single-cell data using umap. *Nat Biotechnol* 2019;**37**:38–44. https://doi.org/10.1038/nbt.4314
89. Kim G, Chun H. Similarity-assisted variational autoencoder for nonlinear dimension reduction with application to single-cell rna sequencing data. *BMC bioinformatics* 2023;**24**:432.
90. Hwang H, Jeon H, Yeo N. *et al*. Big data and deep learning for rna biology. *Exp Mol Med* 2024;**56**:1293–321.
91. Ding J, Condon A, Shah SP. Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. *Nat Commun* 2018;**9**:2002.
92. Sun X, Liu Y, An L. Ensemble dimensionality reduction and feature gene extraction for single-cell rna-seq data. *Nat Commun* 2020;**11**:5853.
93. Haghverdi L, Lun ATL, Morgan MD. *et al*. Batch effects in single-cell rna-sequencing data are corrected by matching mutual nearest neighbors. *Nat Biotechnol* 2018;**36**:421–7. https://doi.org/10.1038/nbt.4091
94. Hie B, Bryson B, Berger B. Efficient integration of heterogeneous single-cell transcriptomes using scanorama. *Nat Biotechnol* 2019;**37**:685–91. https://doi.org/10.1038/s41587-019-0113-3
95. Han W, Cheng Y, Chen J. *et al*. Self-supervised contrastive learning for integrative single cell rna-seq data analysis. *Brief Bioinform* 2022;**23**:bbac377.
96. Wang T, Johnson TS, Shao W. *et al*. Bermuda: a novel deep transfer learning method for single-cell rna sequencing batch correction reveals hidden high-resolution cellular subtypes. *Genome Biol* 2019;**20**:1–15.
97. Linderman GC, Zhao J, Roulis M. *et al*. Zero-preserving imputation of single-cell rna-seq data. *Nat Commun* 2022;**13**:192.
98. Van Dijk D, Sharma R, Nainys J. *et al*. Recovering gene interactions from single-cell data using data diffusion. *Cell* 2018;**174**:716–29. https://doi.org/10.1016/j.cell.2018.05.061
99. Li WV, Li JJ. An accurate and robust imputation method scimpute for single-cell rna-seq data. *Nat Commun* 2018;**9**:997.
100. Huang M, Wang J, Torre E. *et al*. Saver: gene expression recovery for single-cell rna sequencing. *Nat Methods* 2018;**15**:539–42. https://doi.org/10.1038/s41592-018-0033-z
101. Eraslan G, Simon LM, Mircea M. *et al*. Single-cell rna-seq denoising using a deep count autoencoder. *Nat Commun* 2019;**10**:390.
102. Xu Y, Zhang Z, You L. *et al*. Scigans: single-cell rna-seq imputation using generative adversarial networks. *Nucleic Acids Res* 2020;**48**:e85–5. https://doi.org/10.1093/nar/gkaa506
103. Dai C, Jiang Y, Yin C. *et al*. Scimc: a platform for benchmarking comparison and visualization analysis of scrna-seq data imputation methods. *Nucleic Acids Res* 2022;**50**:4877–99. https://doi.org/10.1093/nar/gkac317
104. Bai L, Ji B, Wang S. Sae-impute: Imputation for single-cell data via subspace regression and auto-encoders. *BMC bioinformatics* 2024;**25**:317.
105. Lopez R, Regier J, Cole MB. *et al*. Deep generative modeling for single-cell transcriptomics. *Nat Methods* 2018;**15**:1053–8. https://doi.org/10.1038/s41592-018-0229-2
106. Kingma DP, Welling M. Auto-encoding variational bayes. *Stat* 2014;**1050**:1. https://doi.org/10.1371/journal.pone.0079190
107. Sohn K, Lee H, Yan X. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems* 2015;**2**:3483–91.
108. Ng A. *et al*. Sparse autoencoder. *CS294A Lecture notes* 2011;**72**:1–19.
109. Jin K, Ou-Yang L, Zhao X-M. *et al*. Sctssr: gene expression recovery for single-cell rna sequencing using two-side sparse self-representation. *Bioinformatics* 2020;**36**:3131–8. https://doi.org/10.1093/bioinformatics/btaa108
110. Gong W, Kwak I-Y, Pota P. *et al*. Drimpute: imputing dropout events in single cell rna sequencing data. *BMC bioinformatics* 2018;**19**:1–10.
111. Arisdakessian C, Poirion O, Yunits B. *et al*. Deepimpute: An accurate, fast, and scalable deep neural network method to impute single-cell rna-seq data. *Genome Biol* 2019;**20**:1–14.
112. Talwar D, Mongia A, Sengupta D. *et al*. Autoimpute: autoencoder based imputation of single-cell rna-seq data. *Sci Rep* 2018;**8**:16329.
113. Gunady MK, Kancherla J, Bravo HC. *et al*. Scgain: single cell rna-seq data imputation using generative adversarial networks. bioRxiv. 2019;837302. https://doi.org/10.1101/837302
114. Zappia L, Phipson B, Oshlack A. Splatter: simulation of single-cell rna sequencing data. *Genome Biol* 2017;**18**:174.
115. Lin P, Troup M, Ho JWK. Cidr: ultrafast and accurate clustering through imputation for single-cell rna-seq data. *Genome Biol* 2017;**18**:1–11.
116. Gawlikowski J, Tassi CRN, Ali M. *et al*. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review* 2023;**56**:1513–89.
117. Welch JD, Kozareva V, Ferreira A. *et al*. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell* 2019;**177**:1873–87. https://doi.org/10.1016/j.cell.2019.05.006
118. Gao C, Liu J, Kriebel AR. *et al*. Iterative single-cell multi-omic integration using online learning. *Nat Biotechnol* 2021;**39**:1000–7. https://doi.org/10.1038/s41587-021-00867-x
119. Jin S, Zhang L, Nie Q. Scai: An unsupervised approach for the integrative analysis of parallel single-cell transcriptomic and epigenomic profiles. *Genome Biol* 2020;**21**:1–19.
120. Ashuach T, Gabitto MI, Koodli RV. *et al*. Multivi: deep generative model for the integration of multimodal data. *Nat Methods* 2023;**20**:1222–31. https://doi.org/10.1038/s41592-023-01909-9
121. Gong B, Zhou Y, Purdom E. Cobolt: integrative analysis of multimodal single-cell sequencing data. *Genome Biol* 2021;**22**:1–21.
122. Zuo C, Chen L. Deep-joint-learning analysis model of single cell transcriptome and open chromatin accessibility data. *Brief Bioinform* 2021;**22**:bbaa287.

123. Minoura K, Abe K, Nam H. *et al*. A mixture-of-experts deep generative model for integrated analysis of single-cell multiomics data. *Cell reports methods* 2021;**1**:100071. https://doi.org/10.1016/j.crmeth.2021.100071
124. Li G, Shaliu F, Wang S. *et al*. A deep generative model for multi-view profiling of single-cell rna-seq and atac-seq data. *Genome Biol* 2022;**23**:20.
125. Satija R, Farrell JA, Gennert D. *et al*. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* 2015;**33**: 495–502. https://doi.org/10.1038/nbt.3192
126. Stuart T, Butler A, Hoffman P. *et al*. William M Mauck, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *Cell* 2019;**177**: 1888–902. https://doi.org/10.1016/j.cell.2019.05.031
127. Korsunsky I, Millard N, Fan J. *et al*. Fast, sensitive and accurate integration of single-cell data with harmony. *Nat Methods* 2019;**16**:1289–96. https://doi.org/10.1038/s41592-019-0619-0
128. Cable DM, Murray E, Zou LS. *et al*. Robust decomposition of cell type mixtures in spatial transcriptomics. *Nat Biotechnol* 2022;**40**: 517–26. https://doi.org/10.1038/s41587-021-00830-w
129. Li B, Bao F, Hou Y. *et al*. Tissue characterization at an enhanced resolution across spatial omics platforms with deep generative model. *Nat Commun* 2024;**15**:6541.
130. Marconato L, Palla G, Yamauchi KA. *et al*. Spatialdata: An open and universal data framework for spatial omics. *Nat Methods* 2024;**22**:58–62.
131. Zhou X, Dong K, Zhang S. Integrating spatial transcriptomics data across different conditions, technologies and developmental stages. *Nature Computational Science* 2023;**3**:894–906. https://doi.org/10.1038/s43588-023-00528-w
132. He B, Bergenstråhle L, Stenbeck L. *et al*. Integrating spatial gene expression and breast tumour morphology via deep learning. *Nature biomedical engineering* 2020;**4**:827–34. https://doi.org/10.1038/s41551-020-0578-x
133. Jialu H, Zhong Y, Shang X. A versatile and scalable single-cell data integration algorithm based on domain-adversarial and variational approximation. *Brief Bioinform* 2022;**23**:bbab400.
134. Khan SA, Lehmann R, Morentin X M-D. *et al*. Scaegan: unification of single-cell genomics data by adversarial learning of latent space correspondences. *PloS One* 2023;**18**:e0281315. https://doi.org/10.1371/journal.pone.0281315
135. Cao Z-J, Gao G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nat Biotechnol* 2022;**40**:1458–66. https://doi.org/10.1038/s41587-022-01284-4
136. Luecken MD, Büttner M, Chaichoompu K. *et al*. Benchmarking atlas-level data integration in single-cell genomics. *Nat Methods* 2022;**19**:41–50. https://doi.org/10.1038/s41592-021-01336-8
137. Rodriguez-Meira A, Buck G, Clark S-A. *et al*. Unravelling intratumoral heterogeneity through high-sensitivity single-cell mutational analysis and parallel rna sequencing. *Mol Cell* 2019;**73**: 1292–305. https://doi.org/10.1016/j.molcel.2019.01.009
138. Dey SS, Kester L, Spanjaard B. *et al*. Integrated genome and transcriptome sequencing of the same cell. *Nat Biotechnol* 2015;**33**: 285–9. https://doi.org/10.1038/nbt.3129
139. Angermueller C, Clark SJ, Lee HJ. *et al*. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat Methods* 2016;**13**:229–32. https://doi.org/10.1038/nmeth.3728
140. Hou Y, Guo H, Cao C. *et al*. Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. *Cell Res* 2016;**26**:304–19. https://doi.org/10.1038/cr.2016.23
141. Cao J, Cusanovich DA, Ramani V. *et al*. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* 2018;**361**:1380–5. https://doi.org/10.1126/science.aau0730
142. Chen S, Lake BB, Zhang K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat Biotechnol* 2019;**37**:1452–7. https://doi.org/10.1038/s41587-019-0290-0
143. Ma S, Zhang B, LaFave LM. *et al*. Chromatin potential identified by shared single-cell profiling of rna and chromatin. *Cell* 2020;**183**:1103–16. https://doi.org/10.1016/j.cell.2020.09.056
144. Gerlach JP, van Buggenum JAG, Tanis SEJ. *et al*. Combined quantification of intracellular (phospho-) proteins and transcriptomics from fixed single cells. *Sci Rep* 2019;**9**:1469.
145. Stoeckius M, Hafemeister C, Stephenson W. *et al*. Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods* 2017;**14**:865–8. https://doi.org/10.1038/nmeth.4380
146. Wang Y, Yuan P, Yan Z. *et al*. Single-cell multiomics sequencing reveals the functional regulatory landscape of early embryos. *Nat Commun* 2021;**12**:1247.
147. Clark SJ, Argelaguet R, Kapourani C-A. *et al*. Scnmt-seq enables joint profiling of chromatin accessibility dna methylation and transcription in single cells. *Nat Commun* 2018;**9**:781.
148. Lv T, Zhang Y, Liu J. *et al*. Multi-omics integration for both single-cell and spatially resolved data based on dual-path graph attention auto-encoder. *Brief Bioinform* 2024;**25**:bbae450.
149. Long Y, Ang KS, Sethi R. *et al*. Deciphering spatial domains from spatial multi-omics with spatialglue. *Nat Methods* 2024;**21**:1658–67.
150. Yunfei H, Xie M, Li Y. *et al*. Benchmarking clustering, alignment, and integration methods for spatial transcriptomics. *Genome Biol* 2024;**25**:212. https://doi.org/10.1186/s13059-024-03361-0
151. Chenling X, Lopez R, Mehlman E. *et al*. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Mol Syst Biol* 2021;**17**:e9620. https://doi.org/10.15252/msb.20209620
152. Mohammad Lotfollahi F, Wolf A, Theis FJ. Scgen predicts single-cell perturbation responses. *Nat Methods* 2019;**16**:715–21. https://doi.org/10.1038/s41592-019-0494-8
153. Song D, Wang Q, Yan G. *et al*. scdesign3 generates realistic in silico data for multimodal single-cell and spatial omics. *Nat Biotechnol* 2024;**42**:247–52. https://doi.org/10.1038/s41587-023-01772-1
154. Zinati Y, Takiddeen A, Emad A. Groundgan: Grn-guided simulation of single-cell rna-seq data using causal generative adversarial networks. *Nat Commun* 2024;**15**:4055.
155. Pratapa A, Jalihal AP, Law JN. *et al*. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat Methods* 2020;**17**:147–54. https://doi.org/10.1038/s41592-019-0690-6
156. Cao Y, Yang P, Yang JYH. A benchmark study of simulation methods for single-cell rna sequencing data. *Nat Commun* 2021;**12**:6911.
157. Gayoso A, Steier Z, Lopez R. *et al*. Joint probabilistic modeling of single-cell multi-omic data with totalvi. *Nat Methods* 2021;**18**: 272–82. https://doi.org/10.1038/s41592-020-01050-x
158. Cao K, Gong Q, Hong Y. *et al*. A unified computational framework for single-cell data integration with optimal transport. *Nat Commun* 2022;**13**:7419.
159. Mike W, Goodman N. Multimodal generative models for scalable weakly-supervised learning. *Advances in neural information processing systems* 2018;**31**:5580–90.

160. Shi Y, Paige B, Torr P. *et al*. Variational mixture-of-experts autoencoders for multi-modal deep generative models. *Advances in neural information processing systems* 2019;**32**: 15718–29.
161. Liu Q, Allamanis M, Brockschmidt M. *et al*. Constrained graph variational autoencoders for molecule design. *Advances in neural information processing systems* 2018;**31**:7806–15.
162. Makrodimitris S, Pronk B, Abdelaal T. *et al*. An in-depth comparison of linear and non-linear joint embedding methods for bulk and single-cell multi-omics. *Brief Bioinform* 2024;**25**:bbad416.
163. Lähnemann D, Köster J, Szczurek E. *et al*. Eleven grand challenges in single-cell data science. *Genome Biol* 2020;**21**:1–35.
164. Svahn C, Sysoev O. Ccvae: A variational autoencoder for handling censored covariates. In:*2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 709–14. New York, USA, Wani MA: IEEE COMPUTER SOC, 2022.
165. Tang X, Zhang J, He Y. *et al*. Explainable multi-task learning for multi-modality biological data analysis. *Nat Commun* 2023;**14**:2546.
166. Tang Z, Chen G, Chen S. *et al*. Modal-nexus auto-encoder for multi-modality cellular data integration and imputation. *Nature. Communications* 2024;**15**:9021.
167. Wu KE, Yost KE, Chang HY. *et al*. Babel enables cross-modality translation between multiomic profiles at single-cell resolution. *Proc Natl Acad Sci* 2021;**118**:e2023070118.
168. Yang KD, Belyaeva A, Venkatachalapathy S. *et al*. Multi-domain translation between single-cell imaging and sequencing data using autoencoders. *Nat Commun* 2021;**12**:31.
169. Zhou Z, Ye C, Wang J. *et al*. Surface protein imputation from single cell transcriptomes by deep neural networks. *Nat Commun* 2020;**11**:651.
170. Lotfollahi M, Naghipourfar M, Luecken MD. *et al*. Mapping single-cell data to reference atlases by transfer learning. *Nat Biotechnol* 2022;**40**:121–30. https://doi.org/10.1038/s41587-021-01001-7
171. Lan M, Zhang S, Gao L. Efficient generation of paired single-cell multiomics profiles by deep learning. *Advanced Science* 2023;**10**:2301169.
172. Jin-Hong D, Cai Z, Roeder K. Robust probabilistic modeling for single-cell multimodal mosaic integration and imputation via scvaeit. *Proc Natl Acad Sci* 2022;**119**:e2214414119. https://doi.org/10.1073/pnas.2214414119
173. Lakkis J, Schroeder A, Kenong S. *et al*. A multi-use deep learning method for cite-seq and single-cell rna-seq data integration with cell surface protein prediction and imputation. *Nature machine intelligence* 2022;**4**:940–52. https://doi.org/10.1038/s42256-022-00545-w
174. Meng C, Kuster B, Culhane AC. *et al*. A multivariate approach to the integration of multi-omics datasets. *BMC bioinformatics* 2014;**15**:1–13.
175. Argelaguet R, Arnol D, Bredikhin D. *et al*. Mofa+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol* 2020;**21**:1–17.
176. Yinlei H, Wan S, Luo Y. *et al*. Benchmarking algorithms for single-cell multi-omics prediction and integration. *Nat Methods* 2024;**21**:2182–94.
177. Walter FC, Stegle O, Velten B. Fishfactor: a probabilistic factor model for spatial transcriptomics data with subcellular resolution. *Bioinformatics* 2023;**39**:btad183.
178. William Townes F, Engelhardt BE. Nonnegative spatial factorization applied to spatial genomics. *Nat Methods* 2023;**20**: 229–38. https://doi.org/10.1038/s41592-022-01687-w
179. Äijö T, Müller CL, Bonneau R. Temporal probabilistic modeling of bacterial compositions derived from 16s rrna sequencing. *Bioinformatics* 2018;**34**:372–80. https://doi.org/10.1093/bioinformatics/btx549
180. Velten B, Braunger JM, Argelaguet R. *et al*. Identifying temporal and spatial patterns of variation from multimodal data using mefisto. *Nat Methods* 2022;**19**:179–86. https://doi.org/10.1038/s41592-021-01343-9
181. Zhu Q, Shah S, Dries R. *et al*. Identification of spatially associated subpopulations by combining scrnaseq and sequential fluorescence in situ hybridization data. *Nat Biotechnol* 2018;**36**: 1183–90. https://doi.org/10.1038/nbt.4260
182. Dries R, Zhu Q, Dong R. *et al*. Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biol* 2021;**22**:1–31.
183. Li R, Yang X. De novo reconstruction of cell interaction landscapes from single-cell spatial transcriptome data with deeplinc. *Genome Biol* 2022;**23**:124.
184. Fischer DS, Schaar AC, Theis FJ. Modeling intercellular communication in tissues using spatial graphs of cells. *Nat Biotechnol* 2023;**41**:332–6. https://doi.org/10.1038/s41587-022-01467-z
185. Tan X, Andrew S, Tran M. *et al*. Spacell: integrating tissue morphology and spatial gene expression to predict disease cells. *Bioinformatics* 2020;**36**:2293–4. https://doi.org/10.1093/bioinformatics/btz914
186. Armenteros JJA, Sønderby CK, Sønderby SK. *et al*. Deeploc: prediction of protein subcellular localization using deep learning. *Bioinformatics* 2017;**33**:3387–95. https://doi.org/10.1093/bioinformatics/btx431
187. Yan H, Weng D, Dongguo Li YG. *et al*. Prior knowledge-guided multilevel graph neural network for tumor risk prediction and interpretation via multi-omics data integration. *Brief Bioinform* 2024;**25**:bbae184.
188. Zeng Y, Song Y, Zhang C. *et al*. Imputing spatial transcriptomics through gene network constructed from protein language model. *Communications Biology* 2024;**7**:1271.
189. Li J, Yang F, Wang F. *et al*. Integrating prior knowledge with graph encoder for gene regulatory inference from single-cell rna-seq data. In:*2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 102–7. New York, USA: IEEE, 2022.
190. Xie Z, Li X, Mora A. A comparison of cell–cell interaction prediction tools based on scrna-seq data. *Biomolecules* 2023;**13**:1211. https://doi.org/10.3390/biom13081211
191. Shan N, Yao L, Guo H. *et al*. Citedb: a manually curated database of cell–cell interactions in human. *Bioinformatics* 2022;**38**: 5144–8. https://doi.org/10.1093/bioinformatics/btac654
192. Yang W, Wang P, Luo M. *et al*. Deepcci: a deep learning framework for identifying cell–cell interactions from single-cell rna sequencing data. *Bioinformatics* 2023;**39**:btad596.
193. Liu Z, Sun D, Wang C. Evaluation of cell-cell interaction methods by integrating single-cell rna sequencing data with spatial information. *Genome Biol* 2022;**23**:218.
194. Bar-Joseph Z, Gitter A, Simon I. Studying and modelling dynamic biological processes using time-series gene expression data. *Nat Rev Genet* 2012;**13**:552–64. https://doi.org/10.1038/nrg3244
195. Larsson L, Frisén J, Lundeberg J. Spatially resolved transcriptomics adds a new dimension to genomics. *Nat Methods* 2021;**18**:15–8. https://doi.org/10.1038/s41592-020-01038-7
196. Seferbekova Z, Lomakin A, Yates LR. *et al*. Spatial biology of cancer evolution. *Nat Rev Genet* 2023;**24**:295–313. https://doi.org/10.1038/s41576-022-00553-x

197. Velten B, Stegle O. Principles and challenges of modeling temporal and spatial omics data. *Nat Methods* 2023;**20**:1462–74. https://doi.org/10.1038/s41592-023-01992-y
198. Tian T, Zhang J, Lin X. *et al*. Dependency-aware deep generative models for multitasking analysis of spatial omics data. *Nat Methods* 2024;**21**:1501–13.
199. Kanehisa M. The kegg database. In: Bock G, Goode JA (eds.), *'In silico'simulation of Biological Processes: Novartis Foundation Symposium 247*, Vol. **247**, pp. 91–103. Hoboken, NJ: Wiley Online Library, 2002.
200. Fabregat A, Jupe S, Matthews L. *et al*. The reactome pathway knowledgebase. *Nucleic Acids Res* 2018;**46**:D649–55. https://doi.org/10.1093/nar/gkx1132
201. Agrawal A, Balci H, Hanspers K. *et al*. Wikipathways 2024: next generation pathway database. *Nucleic Acids Res* 2024;**52**:D679–89. https://doi.org/10.1093/nar/gkad960
202. Szklarczyk D, Gable AL, Nastou KC. *et al*. The string database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res* 2021;**49**:D605–12. https://doi.org/10.1093/nar/gkaa1074
203. Warde-Farley D, Donaldson SL, Comes O. *et al*. The genemania prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res* 2010;**38**:W214–20. https://doi.org/10.1093/nar/gkq537
204. Liu Z-P, Wu C, Miao H. *et al*. Regnetwork: An integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. *Database* 2015;**2015**:bav095. https://doi.org/10.1093/database/bav095
205. Jin S, Guerrero-Juarez CF, Zhang L. *et al*. Inference and analysis of cell-cell communication using cellchat. *Nat Commun* 2021;**12**:1088.
206. Efremova M, Vento-Tormo M, Teichmann SA. *et al*. Cellphonedb: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nat Protoc* 2020;**15**:1484–506. https://doi.org/10.1038/s41596-020-0292-x
207. Wang Y, Wang R, Zhang S. *et al*. Italk: An r package to characterize and illustrate intercellular communication bioRxiv. 2019;507871.
208. Dimitrov D, Türei D, Garrido-Rodriguez M. *et al*. Comparison of methods and resources for cell-cell communication inference from single-cell rna-seq data. *Nat Commun* 2022;**13**:3224.
209. Hou R, Denisenko E, Ong HT. *et al*. Predicting cell-to-cell communication networks using natmi. *Nat Commun* 2020;**11**:5011.
210. Cheng J, Zhang J, Zhongdao W. *et al*. Inferring microenvironmental regulation of gene expression from single-cell rna sequencing data using scmlnet with an application to covid-19. *Brief Bioinform* 2021;**22**:988–1005. https://doi.org/10.1093/bib/bbaa327
211. Cabello-Aguilar S, Alame M, Kon-Sun-Tack F. *et al*. Single-cellsignalr: inference of intercellular networks from single-cell transcriptomics. *Nucleic Acids Res* 2020;**48**:e55–5. https://doi.org/10.1093/nar/gkaa183
212. Raredon MSB, Yang J, Garritano J. *et al*. Computation and visualization of cell–cell signaling topologies in single-cell systems data using connectome. *Sci Rep* 2022;**12**:4187.
213. Yuxuan H, Peng T, Gao L. *et al*. Cytotalk: De novo construction of signal transduction networks using single-cell transcriptomic data. *Sci Adv* 2021;**7**:eabf1356.
214. Zhang Y, Liu T, Xuesong H. *et al*. Cellcall: integrating paired ligand–receptor and transcription factor activities for cell–cell communication. *Nucleic Acids Res* 2021;**49**:8520–34. https://doi.org/10.1093/nar/gkab638
215. Yang F, Wang W, Wang F. *et al*. Scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. *Nature Machine Intelligence* 2022;**4**:852–66. https://doi.org/10.1038/s42256-022-00534-z
216. Theodoris CV, Xiao L, Chopra A. *et al*. Transfer learning enables predictions in network biology. *Nature* 2023;**618**:616–24. https://doi.org/10.1038/s41586-023-06139-9
217. Hao M, Gong J, Zeng X. *et al*. Large-scale foundation model on single-cell transcriptomics. *Nat Methods* 2024;**21**:1481–91.
218. Cui H, Wang C, Maan H. *et al*. Scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nat Methods* 2024;**21**:1470–80.
219. Heimberg G, Kuo T, DePianto D. *et al*. Scalable querying of human cell atlases via a foundational model reveals commonalities across fibrosis-associated macrophages. bioRxiv. 2023;549537. https://doi.org/10.1101/2023.07.18.549537
220. Bian H, Chen Y, Dong X. *et al*. Scmulan: A multitask generative pre-trained language model for single-cell analysis. In: Jian M (ed.), *International Conference on Research in Computational Molecular Biology*, pp. 479–82. Berlin: Springer, 2024.
221. Yao S, Zhao J, Yu D. *et al*. React: synergizing reasoning and acting in language models (2022). arXiv preprint arXiv:2210.03629. 2023. https://doi.org/10.48550/arXiv.2210.03629
222. Abdelaal T, Michielsen L, Cats D. *et al*. A comparison of automatic cell identification methods for single-cell rna sequencing data. *Genome Biol* 2019;**20**:1–19.
223. Wang J, Zou Q, Lin C. A comparison of deep learning-based pre-processing and clustering approaches for single-cell rna sequencing data. *Brief Bioinform* 2022;**23**:bbab345.
224. Mompel P B-ı, Wessels L, Müller-Dott S. *et al*. Gene regulatory network inference in the era of single-cell multi-omics. *Nat Rev Genet* 2023;**24**:739–54. https://doi.org/10.1038/s41576-023-00618-5
225. Wang S, Zheng H, Choi JS. *et al*. A systematic evaluation of the computational tools for ligand-receptor-based cell–cell interaction inference. *Brief Funct Genomics* 2022;**21**:339–56. https://doi.org/10.1093/bfgp/elac019
226. Athaya T, Ripan RC, Li X. *et al*. Multimodal deep learning approaches for single-cell multi-omics data integration. *Brief Bioinform* 2023;**24**:bbad313.
227. Khan SA, Maillo A, Lagani V. *et al*. Reusability report: learning the transcriptional grammar in single-cell rna-sequencing data using transformers. *Nature. Machine Intelligence* 2023;**5**:1437–46.
228. Proks M, Salehin N, Brickman JM. Deep learning-based models for preimplantation mouse and human embryos based on single-cell rna sequencing. *Nat Methods* 2024;**22**:207–16.
229. Dohmen J, Baranovskii A, Ronen J. *et al*. Identifying tumor cells at the single-cell level using machine learning. *Genome Biol* 2022;**23**:123.
230. Biancalani T, Scalia G, Buffoni L. *et al*. Deep learning and alignment of spatially resolved single-cell transcriptomes with tangram. *Nat Methods* 2021;**18**:1352–62. https://doi.org/10.1038/s41592-021-01264-7
231. Halawani R, Buchert M, Chen Y-PP. Deep learning exploration of single-cell and spatially resolved cancer transcriptomics to unravel tumour heterogeneity. *Comput Biol Med* 2023;**164**:107274. https://doi.org/10.1016/j.compbiomed.2023.107274
232. Qin F, Cai G, Amos CI. *et al*. A statistical learning method for simultaneous copy number estimation and subclone clustering with single-cell sequencing data. *Genome Res* 2024;**34**:85–93.
233. Alamin M, Sultana MH, Babarinde IA. *et al*. Single-cell rna-seq data analysis reveals functionally relevant biomarkers of early

brain development and their regulatory footprints in human embryonic stem cells (hescs). *Brief Bioinform* 2024;**25**:bbae230.
234. Yang W, Wang P, Shouping X. *et al*. Deciphering cell–cell communication at single-cell resolution for spatial transcriptomics with subgraph-based graph attention network. *Nat Commun* 2024;**15**:7101.
235. Ji X, Pei Q, Zhang J. *et al*. Single-cell sequencing combined with machine learning reveals the mechanism of interaction between epilepsy and stress cardiomyopathy. *Front Immunol* 2023;**14**:1078731.
236. Wegmann R, Bonilla X, Casanova R. *et al*. Single-cell landscape of innate and acquired drug resistance in acute myeloid leukemia. *Nat Commun* 2024;**15**:9402.
237. Van de Sande B, Lee JS, Mutasa-Gottgens E. *et al*. Applications of single-cell rna sequencing in drug discovery and development. *Nat Rev Drug Discov* 2023;**22**:496–520. https://doi.org/10.1038/s41573-023-00688-4
238. Tang Z, Liu X, Li Z. *et al*. Sparx: elucidate single-cell spatial heterogeneity of drug responses for personalized treatment. *Brief Bioinform* 2023;**24**:bbad338.
239. Ianevski A, Nader K, Driva K. *et al*. Single-cell transcriptomes identify patient-tailored therapies for selective co-inhibition of cancer clones. *Nat Commun* 2024;**15**:8579.
240. Chakraborty S, Sharma G, Karmakar S. *et al*. Multi-omics approaches in cancer biology: new era in cancer therapy. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease* 2024;**1870**:167120. https://doi.org/10.1016/j.bbadis.2024.167120
241. Si S, Hongyan Liu LX, Zhan S. Identification of novel therapeutic targets for chronic kidney disease and kidney function by integrating multi-omics proteome with transcriptome. *Genome Med* 2024;**16**:84.
242. Sorin M, Rezanejad M, Karimi E. *et al*. Single-cell spatial landscapes of the lung tumour immune microenvironment. *Nature* 2023;**614**:548–54. https://doi.org/10.1038/s41586-022-05672-3
243. Niu X, Man X, Zhu J. *et al*. Identification of the immune-associated characteristics and predictive biomarkers of keratoconus based on single-cell rna-sequencing and bulk rna-sequencing. *Front Immunol* 2023;**14**:1220646.
244. Qi R, Zou Q. Trends and potential of machine learning and deep learning in drug study at single-cell level. *Research* 2023;**6**:0050.
245. Liu Y, Li H, Zeng T. *et al*. Integrated bulk and single-cell transcriptomes reveal pyroptotic signature in prognosis and therapeutic options of hepatocellular carcinoma by combining deep learning. *Brief Bioinform* 2024;**25**:bbad487.

© The Author(s) 2025. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited.
For commercial re-use, please contact journals.permissions@oup.com
Briefings in Bioinformatics, 2025, **26(2)**, bbaf136
https://doi.org/10.1093/bib/bbaf136
Review