genes

MDPI

Review

Progress in the Application of Machine Learning in the Field of Single-Cell and Spatial Transcriptomics

Yan Zhu [1,2,3,†], Ziling Hao [1,2,3,†], Li Zhu [1,2,3], Linyuan Shen [1,2,3], Yihui Liu [4,*] and Mailin Gan [1,2,3,*]

[1] Farm Animal Germplasm Resources and Biotech Breeding Key Laboratory of Sichuan Province, Sichuan Agricultural University, Chengdu 611130, China; 2024302105@stu.sicau.edu.cn (Y.Z.); 2022002001@stu.sicau.edu.cn (Z.H.); zhuli@sicau.edu.cn (L.Z.)
[2] State Key Laboratory of Swine and Poultry Breeding Industry, Sichuan Agricultural University, Chengdu 611130, China
[3] Key Laboratory of Livestock and Poultry Multi-Omics, Ministry of Agriculture and Rural Affairs, College of Animal and Technology, Sichuan Agricultural University, Chengdu 611130, China
[4] Sichuan Province General Station of Animal Husbandry, Chengdu 610066, China
* Correspondence: yihuisky@163.com (Y.L.); ganmailin@sicau.edu.cn (M.G.)
† These authors contributed equally to this work.

Abstract

The rapid evolution of transcriptome sequencing technologies has driven significant breakthroughs across the life sciences. The advent of single-cell RNA-sequencing (scRNA-seq) has enabled gene expression profiling at single-cell resolution, whereas spatial transcriptomics further contextualizes these transcriptional profiles within preserved tissue morphology. Concurrently, advancements in artificial intelligence have introduced unprecedented opportunities in bioinformatics. As a core component of artificial intelligence, machine learning (ML) substantially outperforms traditional computational methods in deciphering complex, high-dimensional biological data. This review systematically summarizes the significant advantages of integrating ML algorithms into transcriptomic workflows. By leveraging these advanced computational tools, researchers can efficiently extract comprehensive biological insights, elucidate intricate Gene Regulatory Networks, and generate intuitive visualizations. Ultimately, ML-driven transcriptomics provides a robust technical foundation for disease diagnosis, drug discovery, and precision medicine. These advancements underscore the pivotal role of ML in transforming transcriptomic data analysis into an intelligent, highly precise, and multidimensional discipline, thereby accelerating future biological discoveries.

Keywords: machine learning; deep learning; transcriptomics; RNA-sequencing; single-cell transcriptomics; spatial transcriptomics; spatiotemporal transcriptome

Academic Editor: Quan Zou

Received: 20 April 2026
Revised: 20 May 2026
Accepted: 25 May 2026
Published: 26 May 2026

1. Introduction

Transcription is the process by which RNA is synthesized from a DNA template via complementary base pairing. Broadly defined, the transcriptome encompasses the complete set of RNA transcripts, including messenger RNA (mRNA), ribosomal RNA (rRNA), transfer RNA (tRNA), and non-coding RNA (ncRNA), produced under specific physiological conditions. Transcriptomic analysis facilitates the molecular-level investigation of gene expression dynamics across specific cells, tissues, or organs during various developmental stages or under diverse environmental stimuli.

Initially, the primary bottleneck in transcriptomic studies was the efficient generation of sequencing data. The advent of next-generation sequencing (NGS) [1] has largely

overcome this limitation. Consequently, high-throughput transcriptomic profiling is now widely utilized to elucidate gene expression patterns, investigate disease pathogenesis, and characterize crop stress resistance. However, this technological shift has introduced new analytical challenges. The exponential growth of sequencing data yields highly complex, multidimensional expression profiles that vary significantly across cell types, developmental phases, and environmental contexts.

Relying solely on classical univariate statistical tools is increasingly inadequate for deciphering such complex data [2]. Furthermore, advanced modalities like single-cell and spatial transcriptomics are highly computationally intensive and present significant interpretational hurdles [3,4]. As the pace of data generation far outstrips the capacity of manual analysis, vast amounts of transcriptomic data in public repositories remain underutilized. To address these high-dimensional challenges, modern analytical pipelines do not abandon classical methods; rather, they seamlessly integrate foundational statistical techniques (e.g., principal component analysis, PCA) as preliminary dimensionality reduction steps within broader, highly automated machine learning (ML) frameworks. Although early applications of ML to gene expression data emerged in the early 21st century, these initial approaches lacked accessibility and were not incorporated into standardized analytical workflows [5].

In recent years, artificial intelligence (AI) has experienced unprecedented growth, driven largely by significant advancements in machine learning (ML) and the rapid emergence of deep learning technologies [6,7]. Consequently, ML has been extensively integrated into biological and medical research, extending even into routine clinical practice [8–12]. The application of ML algorithms to transcriptomic analysis has become a standard paradigm; for instance, methods such as support vector machines (SVMs) and clustering are now routinely employed for differential expression and enrichment analyses [13,14] (Figure 1). As robust computational tools, ML models efficiently extract critical biological features from massive datasets, thereby accelerating research in animal genetics and breeding, disease diagnosis, epigenetic regulation, and beyond.

Before discussing specific applications, it is essential to define "machine learning" clearly within the context of transcriptomics and to delineate its relationship with classical statistics. Historically, classical statistical methods, such as linear or logistic regression, linear discriminant analysis (LDA), and regularized models (e.g., Lasso and Elastic Net), were firmly rooted in statistical inference, focusing on parameter estimation and hypothesis testing under strict distributional assumptions. However, in the modern omics era, the boundary between statistics and machine learning has evolved into a continuum. When repurposed to optimize predictive accuracy and perform automated feature selection on high-dimensional, highly collinear transcriptomic data, these classical methods function as foundational "statistical learning" algorithms under the broader ML umbrella [15]. Modern deep learning extends this continuum by deploying complex, multi-layered neural networks that bypass rigid statistical assumptions to capture highly non-linear biological representations.

Given the exponential growth of literature in this field, an exhaustive catalog of ML applications across single-cell and spatial transcriptomics is unfeasible. Therefore, this review is intentionally curated to establish a cohesive evolutionary narrative [3]. Rather than treating advanced methodologies as isolated tools, it is crucial to understand their relationships as a continuous computational evolution driven by increasingly complex biological bottlenecks [4].

We trace this progression by examining how the field iteratively evolved to resolve the core challenges of data representation, normalization, and integration. Initially, classical statistical learning methods (such as PCA and linear regression) were sufficient for bulk

RNA-seq, providing essential dimensionality reduction for continuous, densely populated matrices. However, as the field transitioned to single-cell resolution, the extreme sparsity (dropout events) and massive batch effects inherent in scRNA-seq rendered rigid linear assumptions inadequate [16]. This biological bottleneck necessitated a leap toward deep generative models (such as VAEs). Architectures like scVI [17] fundamentally revolutionized normalization and integration by projecting cells into non-linear, probabilistic latent spaces, inherently impute missing values and seamlessly correcting batch effects without strict parametric constraints.

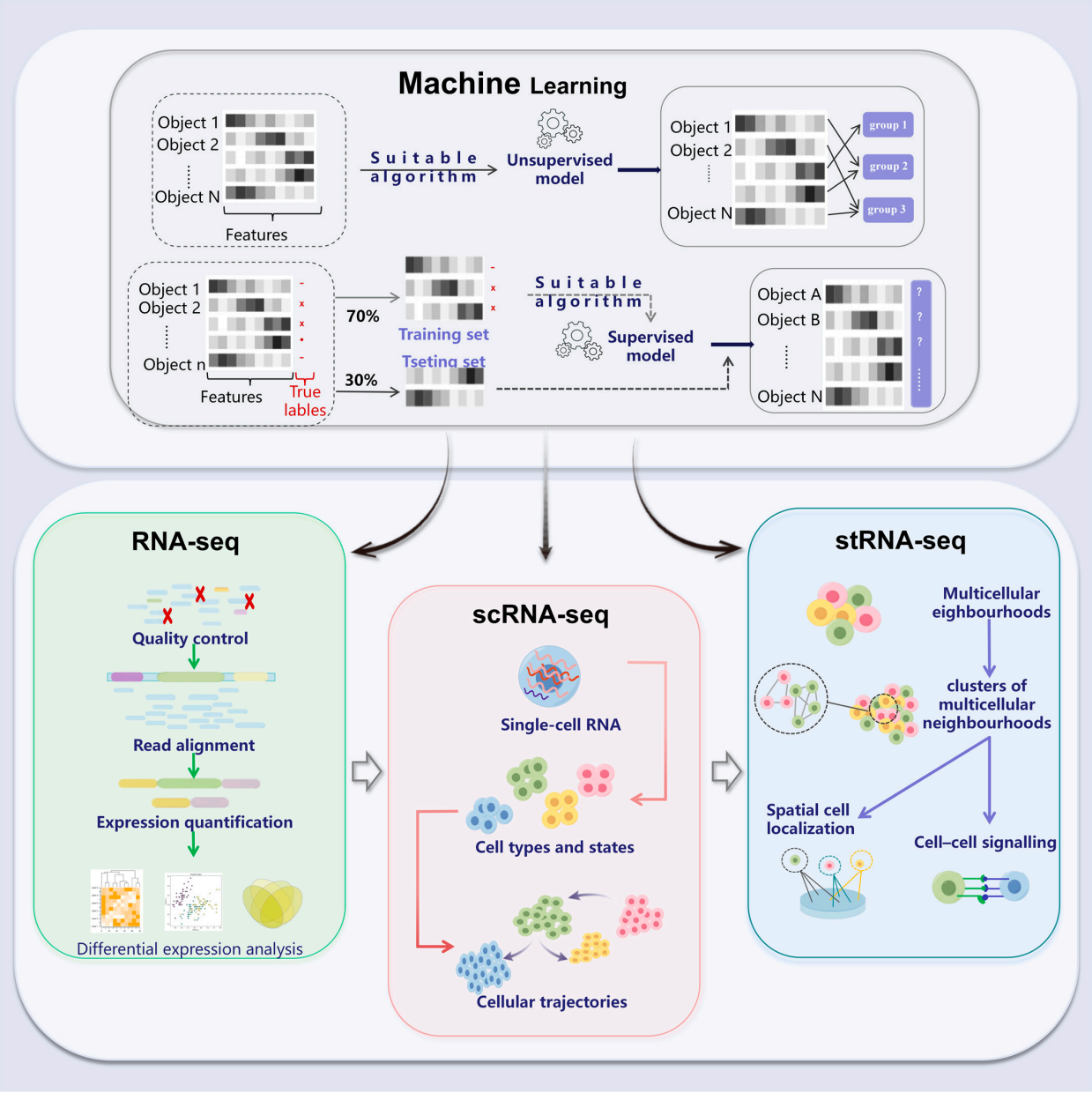

Figure 1. Machine learning has been applied in the analysis flow in various fields of the transcriptome. Machine learning algorithms are mainly divided into supervised and unsupervised categories. Machine learning techniques with these two algorithms as the core have been widely used in various fields of transcriptomics.

Yet, deep generative models still process cells as isolated transcriptomic bags, ignoring the critical biological context of the tissue microenvironment. The subsequent advent of spatial transcriptomics required a methodology capable of interpreting physical topology [18].

Consequently, Graph Neural Networks (GNNs) were adopted to explicitly bridge this gap, modeling spatial coordinates and cell–cell communication networks by treating physical proximity as mathematical edges [19,20]. Finally, while GNNs excel at localized spatial tasks, they typically require retraining for new datasets and struggle with cross-tissue generalizations. To resolve this ultimate challenge of universal representation, the field is currently adopting Transformer-based foundation models [21]. Pre-trained on massive, multi-tissue corpora, these architectures utilize self-attention to capture global gene–gene dependencies, enabling zero-shot predictions and universal cell embeddings.

By connecting these distinct mathematical innovations through the lens of algorithmic evolution, this review systematically illustrates how each architectural paradigm directly compensated for the biological and computational limitations of its predecessor.

2. Modern Machine Learning Architectures in Transcriptomics

As established in the preceding historical trajectory, the application of machine learning (ML) to transcriptomics has transitioned from isolated algorithmic sorting into a deeply integrated computational continuum. Historically, ML applications in biology were broadly dichotomized into supervised and unsupervised learning tasks [5]. At its foundational core, deep learning (DL) transcended these rigid boundaries by deploying artificial neural networks with multiple hidden layers to automatically extract hierarchical feature representations directly from raw expression matrices. In contrast to classical ML methods, which typically rely on shallow architectures and manual feature engineering, DL inherently captures complex, non-linear biological relationships. This capacity for automated representation learning makes it exceptionally suited for analyzing high-dimensional transcriptomic profiles.

While these fundamental statistical and algorithmic categorizations remain conceptually relevant, the exponential growth of transcriptomic data—shifting from dense bulk matrices to sparse single-cell profiles and intricate spatial topologies—has driven the field toward highly specialized, task-specific DL architectures [16]. Consequently, modern analytical pipelines have evolved far beyond basic, dense neural networks. To provide a rigorous, problem-oriented analysis of current methodologies, this section classifies modern computational frameworks not as a fragmented list of software packages but according to their underlying architectural paradigms, each designed to overcome a distinct biological bottleneck: deep generative models for sparsity and dropout correction, graph-based representation learning for microenvironmental topology mapping, attention mechanisms for global gene regulatory modeling, and the modern adaptation of classical algorithms for highly interpretable downstream validation.

2.1. Deep Generative Models and Latent Space Learning

Single-cell RNA-sequencing (scRNA-seq) datasets are inherently noisy and highly sparse, presenting significant computational challenges such as severe batch effects and "dropout" events—instances where expressed transcripts fail to be detected [16]. Traditional linear dimensionality reduction techniques frequently fail to delineate the complex, non-linear manifolds characteristic of such biological data. Consequently, deep generative models, particularly Autoencoders (AEs) and Variational Autoencoders (VAEs), have emerged as foundational computational architectures.

These models operate by compressing high-dimensional gene expression profiles into a lower-dimensional, dense latent space before reconstructing the original data. This information bottleneck architecture forces the network to learn essential, denoised biological variations while effectively discarding technical noise. In modern analytical pipelines, VAEs function as standard instruments for missing value imputation, batch effect correction,

and multi-omic data integration. For instance, the widely adopted scvi-tools framework leverages probabilistic VAEs to facilitate scalable, end-to-end single-cell analyses [22]. Furthermore, architectural innovations such as scArches integrate deep generative models with transfer learning to project novel single-cell query datasets onto massive reference atlases without the computationally prohibitive need to retrain models from scratch. This approach efficiently mitigates batch effects and enhances the identification of rare cell states [23].

2.2. Graph-Based Representation Learning

Cells do not operate in isolation; their developmental trajectories, signaling interactions, and spatial localizations are fundamentally interconnected. Graph Neural Networks (GNNs) and graph convolutional networks (GCNs) have revolutionized the field by enabling computational models to explicitly capture and represent these complex, higher-order structural relationships.

Within a graph-based framework, individual cells or spatial capture spots are conceptualized as nodes, while their physical proximities or signaling interactions are defined as edges. This architecture is indispensable for spatial transcriptomics, where spatial context is critical for decoding the tissue microenvironment. Pioneering frameworks such as *SpaGCN* utilize GCNs to integrate gene expression profiles, spatial coordinates, and histological imaging, thereby accurately identifying spatial domains and spatially variable genes [19]. Furthermore, advanced models like *GraphST* employ graph-based self-supervised contrastive learning to fully exploit spatial neighborhoods. These approaches significantly outperform traditional non-spatial clustering methods in deciphering tissue architecture and performing cell-type deconvolution [24].

2.3. Attention Mechanisms and Foundation Models

A significant paradigm shift in contemporary bioinformatics is the adaptation of attention mechanisms and Transformer architectures, originally pioneered in natural language processing, to transcriptomic analysis. Attention mechanisms are uniquely suited for modeling long-range dependencies, enabling networks to dynamically weigh the importance of specific genes and capture complex regulatory interactions across the entire transcriptome, irrespective of their linear genomic positions.

This architectural evolution has catalyzed the development of "foundation models" in single-cell biology. Through self-supervised pre-training on diverse datasets comprising tens of millions of cells, these models learn highly generalizable representations of cellular states and underlying Gene Regulatory Networks. A prominent example is Geneformer, a context-aware Transformer model that successfully facilitates zero-shot and few-shot predictions in network biology and disease target discovery [25]. Similarly, scGPT represents a substantial advancement toward a universal foundation model for single-cell multi-omics, demonstrating state-of-the-art performance in cell-type annotation, genetic perturbation prediction, and batch integration via generative pre-training [26].

2.4. Classical Machine Learning and Its Modern Adaptation

While deep learning architectures now dominate complex representation learning, classical machine learning algorithms, such as Random Forests (RFs), Support Vector Machines (SVMs), and k-Nearest Neighbors (KNNs), remain vital, albeit repositioned, components of the pipeline [5].

Rather than directly processing raw, high-dimensional sequencing counts, these traditional algorithms are now predominantly deployed in downstream stages following deep learning-based dimensionality reduction. For instance, KNN graphs are routinely constructed from dense latent representations to serve as the topological foundation for

Louvain or Leiden clustering algorithms within standard toolkits like Seurat, ultimately defining robust cell subpopulations [27]. Furthermore, supervised ensemble models are frequently applied to these refined feature sets for targeted downstream tasks, including marker gene selection, clinical trajectory forecasting, and disease severity classification. By operating on lower-dimensional, denoised representations, classical ML algorithms continue to deliver the high interpretability, statistical rigor, and computational efficiency required for clinical and biological validation [16].

3. Application of Machine Learning in Transcriptomics Studies

The transition from classical machine learning to advanced deep learning in transcriptomics is not merely a shift in computational preference but a necessary response to the growing complexity, sparsity, and dimensionality of single-cell and spatial data [28]. Methodologically, this evolution follows a clear trajectory: early classical methods (such as PCA and simple clustering) provided initial dimensionality reduction; probabilistic generative models (like VAEs) emerged to rigorously address batch effects and dropout noise; subsequently, Graph Neural Networks (GNNs) were adopted to capture complex cell-cell and spatial topologies; and, most recently, Transformer-based foundation models have been introduced to leverage large-scale, self-supervised representation learning [4]. Understanding these methodological shifts is crucial for selecting the appropriate tool for specific biological inquiries.

Presently, the field of transcriptomics is broadly categorized into bulk RNA-sequencing, single-cell RNA-sequencing (scRNA-seq), and spatial/spatiotemporal transcriptomics. Figure 2 illustrates the parallel biological development of these sequencing technologies alongside their corresponding machine learning milestones. While high-throughput bulk RNA-seq became widely established around 2008–2009 [29], driving initial improvements in quantification efficiency, the advent of scRNA-seq in 2009 fundamentally shifted the analytical paradigm toward single-cell resolution. Subsequently, the formal proposition of spatial transcriptomics in 2016 introduced critical physical coordinates, enabling researchers to decode spatiotemporal gene expression dynamics within intact tissue architectures.

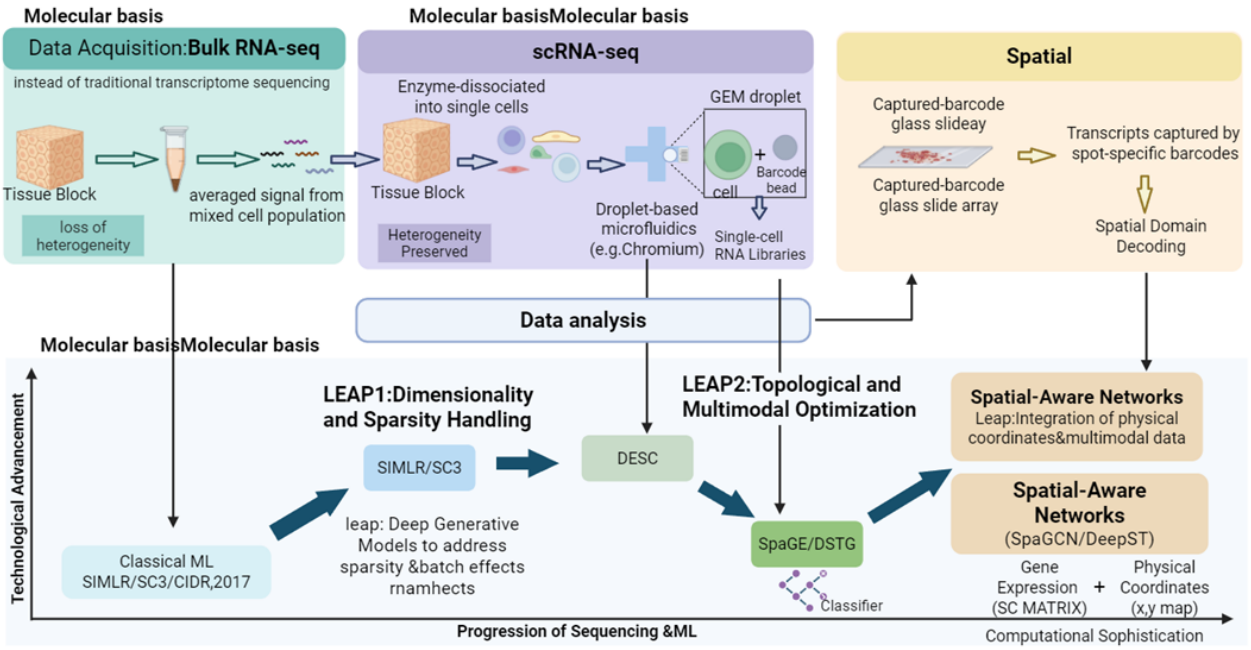

Figure 2. The methodological evolution and molecular basis of machine learning in transcriptomics. The diagram illustrates the biological progression of sequencing technologies alongside

the corresponding computational leaps. Molecular Basics: The field evolved from Bulk RNA-sequencing (requiring tissue homogenization) to single-cell RNA-seq (requiring single-cell dissociation and droplet microfluidics), and finally to spatial transcriptomics (utilizing tissue sections on barcoded arrays to preserve physical morphology). Algorithmic Leaps: Corresponding to these technological shifts, computational methods have undergone significant leaps. Early single-cell analysis relied on classical statistical clustering (e.g., SC3, CIDR). As data sparsity grew, a methodological leap occurred toward deep generative architectures (e.g., DESC, gimVI) for robust representation learning. The emergence of spatial transcriptomics necessitated another critical leap: algorithms (e.g., SpaGE, DSTG) evolved to jointly optimize gene expression embeddings alongside physical tissue topologies. Legend: Green solid/dashed boxes denote bulk RNA-seq milestones and methods; purple boxes represent single-cell RNA-seq algorithms; and yellow/orange boxes indicate spatial transcriptomics applications. Dashed boxes denote the formal proposition milestones of the respective technologies.

3.1. Machine Learning for Transcriptome Sequencing

While initial RNA-sequencing (RNA-seq) protocols emerged around 2008, their revolutionary impact on transcriptomics was comprehensively established by 2009 [29]. This technology relies on high-throughput sequencing to quantify RNA sequences and expression levels within specific cells, tissues, organs, or physiological states [30]. As the earliest and most widely adopted omics technology, bulk RNA-seq is essential for elucidating biological phenomena and the transcriptional regulatory mechanisms underlying complex phenotypes. However, the exponential growth of high-throughput sequencing data poses significant analytical challenges [31]. In this context, machine learning (ML) has emerged as a powerful analytical tool, injecting new vitality into bioinformatics. ML is now applied across both upstream and downstream stages of RNA-seq analysis to process high-dimensional datasets, automatically extract robust features, alleviate computational burdens, and enhance analytical automation.

From a methodological perspective, it is critical to highlight which specific ML models are designed for bulk data. Because bulk RNA-seq averages signals across millions of cells, the resulting data matrices are typically dense, continuous, and free from the extreme sparsity characteristic of single-cell data. Consequently, classical supervised ML algorithms, such as Support Vector Machines (SVMs), Random Forests (RFs), and gradient boosting (e.g., XGBoost), are specifically well-suited for bulk transcriptomics. These decision tree and margin-based methods excel at handling the "small sample, high feature" ($p >> n$) datasets typical of clinical cohorts, providing highly interpretable biomarker selection and disease classification. In contrast, applying massive deep neural networks directly to bulk expression data often leads to severe overfitting due to insufficient independent sample sizes.

A standard RNA-seq pipeline encompasses both experimental and computational phases, with the latter broadly divided into raw data quality control (QC), reference genome alignment, expression quantification, differential expression analysis, and functional annotation [9] (Figure 3). Traditionally, the read alignment step—matching short sequences to a known reference genome—does not directly utilize ML. Instead, it relies on specialized heuristic tools, such as Bowtie, STAR, HISAT2, and BWA. However, ML algorithms are predominantly applied in the subsequent preprocessing and downstream stages. During QC, algorithms like principal component analysis (PCA) and Autoencoders (AEs) can automatically identify outliers and technical noise, thereby ensuring data integrity for subsequent analyses. Furthermore, traditional normalization metrics (e.g., RPKM, FPKM, and TPM) adjust for sequencing depth but can introduce systematic biases. Because classical statistical methods often assume specific data distributions (e.g., normal distributions) that empirical transcriptomic data may violate, ML frameworks offer a robust alternative.

They adapt to complex, non-linear data structures without strict parametric assumptions, ultimately reducing human bias and yielding more accurate gene expression estimates.

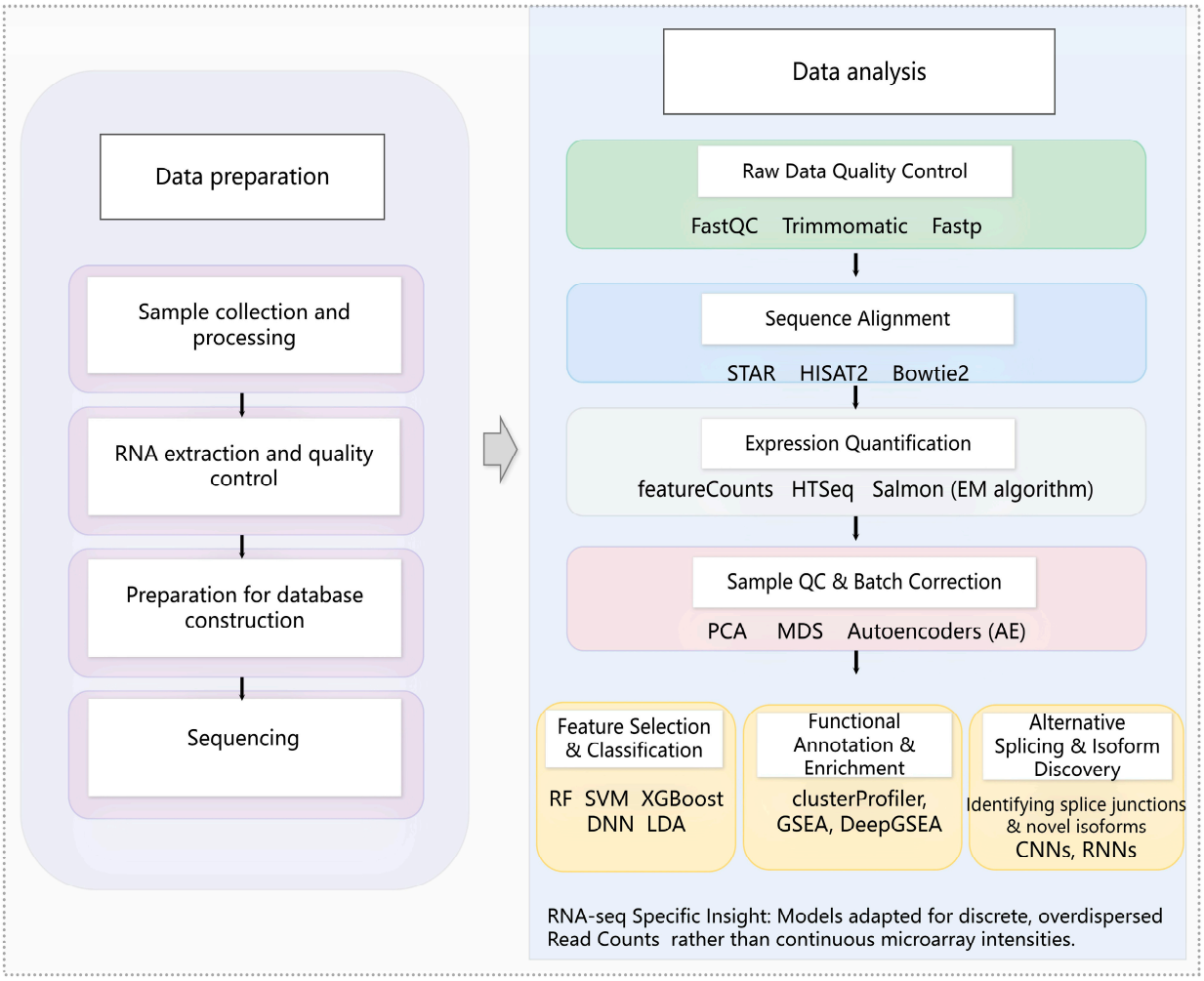

Figure 3. Flow of the transcriptome sequencing process. The left shows the transcriptome data extraction process and the transcriptome data analysis process on the right. Machine learning is mainly used in the analysis of transcriptome data.

In downstream applications, while PCA [11] is routinely utilized for exploratory dimensionality reduction and variance visualization, supervised algorithms like linear discriminant analysis (LDA) and RF [10] are deployed for differential expression analysis. These methods efficiently select highly discriminative genes from massive candidate pools, providing crucial insights into biological processes and disease mechanisms. Finally, to annotate the biological functions of these transcripts, supervised ML algorithms like SVM [14,32] can accurately predict and classify genes with unknown functions based on established functional networks, significantly improving overall annotation efficiency.

Beyond enhancing sequence alignment and quantification, machine learning algorithms excel at processing transcriptomic profiles to predict clinical outcomes and screen for key driver genes [33]. Crucially, this classification capability enables ML frameworks to navigate qualitative or categorical decisions that cannot be represented by continuous variables alone. Such applications typically involve discrete biological classifications, including assigning cellular identities (e.g., T-cell versus B-cell), diagnosing qualitative disease states (e.g., healthy versus tumor), or predicting binary clinical responses (e.g., responder versus non-responder to targeted therapies). By classifying sample subtypes

and predicting disease progression, these tools significantly advance the development of precision medicine and personalized therapy.

The application of supervised ML to disease classification has a rich history. For instance, as early as 2007, supervised models were applied to legacy microarray data to identify gene expression patterns distinguishing Parkinson's disease (PD) patients from healthy controls [34]. More recently, ML algorithms such as k-Nearest Neighbors (KNNs) [35] and SVM [36] have been deployed on modern transcriptomic datasets to discover robust PD biomarkers. Furthermore, the SVM-core NetWAS algorithm [37] leverages transcriptomic functional networks to enhance the interpretation of genome-wide association study (GWAS) results, identifying phenotype-associated genes more effectively. In the realm of oncology, comprehensive evaluations of ML methodologies applied to next-generation sequencing (NGS) data have demonstrated exceptional efficacy in predicting cancer tissue of origin (TOO) [38,39]. These NGS-based predictive models achieved high prediction accuracies (73–99%), substantially outperforming traditional microarray-based techniques (54–100% [40]) and demonstrating immense potential for diagnosing carcinomas of unknown primary origin.

To mitigate the inherent transcriptomic challenges of data overfitting and extreme sensitivity to biological noise, rigorous algorithmic benchmarking is often essential in biomarker discovery [41]. For example, in autoimmune research, researchers utilized PCA and t-SNE for dimensionality reduction, followed by SVM and Random Forest (RF) models, to predict rheumatoid arthritis (RA) activity and treatment response. These models successfully identified highly predictive genes, such as *MGAT5*, *TNFSF10*, *ACAT1*, *CEBPB*, and *MMEL1*, whose expression levels strongly correlate with RA pathophysiology. Similarly, a comprehensive caret-based ML framework evaluated 13 distinct algorithms (including decision trees, Lasso/Elastic Net, GBM, and NNET) to discover radiation biomarkers in mouse renal tissues [42]. Following extensive optimization, the KNN algorithm emerged as the superior model, effectively classifying absorbed radiation doses and identifying seven novel biomarker candidates (e.g., *Brf2*, *Ccng1*), thereby circumventing the high computational overhead of previous genetic algorithm hybrid methods [43]. In infectious disease research, an evaluation of logistic regression, Naive Bayes, SVM, and XGBoost models for predicting COVID-19 severity determined XGBoost to be the most accurate [44]. Crucially, by integrating SHapley Additive exPlanations (SHAP), researchers could biologically interpret the model's decision-making process, identifying key genes (e.g., *COX14*, *LAMB2*, *DOLK*) and clinical features (e.g., absolute neutrophil count, viremia categories) as severity drivers. These interpretable ML outputs provide clinicians with highly actionable insights into early disease trajectories. Detailed benchmarking parameters and validation metrics for these studies are summarized in Table 1.

Table 1. Application of partial machine learning in transcriptome sequencing.

Assignment	Model Name	Machine Learning Algorithm	Validation (Methods, Data Set, Results)	Published Time	Reference
To evaluate the prediction accuracy of the samples in the training set	Not reported	KNN	The qPCR validation of 13 risk markers in two independent test sets showed that PD patients could be accurately identified based on these 13 markers, with a sensitivity of 90% and a specificity of 94%.	2012	[35]
Classifier trained to distinguish between PD patients and healthy controls	Not reported	SVM	On the independent test set, the model distinguished well between Parkinson's disease patients and healthy controls (AUC of 0.74).	2017	[36]

Table 1. *Cont.*

Assignment	Model Name	Machine Learning Algorithm	Validation (Methods, Data Set, Results)	Published Time	Reference
Reordering of genes based on the connectivity patterns of tissue-specific functional networks	NetWAS	SVM	Application to hypertension GWAS data revealed that NetWAS could more effectively identify genes associated with hypertension, and that the genes identified by NetWAS were significantly enriched with functional annotations and drug targets related to hypertension.	2015	[37]
Identified genes that to distinguish different dose groups and tissue types as radiation biomarker candidates	caret-KNN	KNN	The KNN model achieved high accuracy (0.86) and K values (0.49,0.49) in both the training set samples and the test set, and high R^2 values (0.97,0.99) in dose prediction.	2023	[42]
Predict the COVID-19 severity	Not reported	XGBoost	The RNA-seq data, clinical characteristics, and comorbidity information of 299 hospitalized COVID-19 patients were analyzed and compared. Compared with LR, Naive Bayes, and SVM, the XGBoost model performed the best in predicting COVID-19 severity with 95% accuracy and an AUC of 0.99.	2024	[44]

3.2. Machine Learning for Single-Cell Transcriptome Sequencing

Before delving into specialized deep learning architectures, it is essential to establish the standard single-cell RNA-seq (scRNA-seq) analytical workflow. Currently, routine scRNA-seq analysis is predominantly driven by two comprehensive, industry-standard frameworks: Seurat [27] in the R (4.6.0) ecosystem and Scanpy [45] in Python (3.14). A conventional pipeline typically progresses through six highly standardized steps: (1) quality control (QC) to filter out dead cells and multiplets; (2) normalization and feature selection to identify highly variable genes; (3) linear dimensionality reduction (typically PCA), followed by non-linear visualization (e.g., UMAP or t-SNE); (4) data integration and batch correction to remove technical confounding factors; (5) graph-based clustering (e.g., Louvain or Leiden algorithms); and, finally, (6) cell-type annotation using established marker genes.

Unlike bulk data, single-cell transcriptomics introduces severe mathematical challenges: extreme sparsity, "dropout" events (zero-inflation), and complex, non-linear developmental manifolds. Classical linear tools and decision trees are inherently insufficient to handle this degree of biological and technical noise. Consequently, deep generative models (such as VAEs) and Transformer-based architectures have been specifically adapted for single-cell data. These neural network-based methods are uniquely capable of learning robust, low-dimensional latent spaces that inherently impute missing dropout values, seamlessly correct massive batch effects, and enable zero-shot cell-type annotations across diverse tissues, establishing them as the indispensable core of modern scRNA-seq analysis.

Understanding this baseline pipeline is critical for clarifying precisely where advanced machine learning models intersect with standard workflows. For instance, classical statistical methods like SC3 and CIDR aim to optimize the clustering and dimensionality reduction steps. In contrast, generative models like scVI and DESC primarily replace standard batch correction and integration protocols with robust probabilistic or deep embedded frameworks. Meanwhile, modern Transformer-based foundation models (such as scBERT or CelltypeGPT) are largely designed to automate and enhance the final cell-type annotation step via zero-shot or supervised learning. By conceptualizing these advanced architectures

as modular upgrades to standard Seurat or Scanpy pipelines, researchers can design more robust, task-specific analytical strategies.

As the fundamental units of life, organisms comprise a diverse array of cells that undergo developmental differentiation to form complex tissues and organs with distinct lineages and specialized functions. Understanding biological systems at single-cell resolution is essential for elucidating the altered molecular mechanisms underlying various diseases [46]. While early methods for single-cell molecular analysis emerged in the late 20th century, the advent of true single-cell RNA-sequencing (scRNA-seq) in 2009 marked a revolutionary milestone [47]. Driven by continuous improvements in experimental protocols and high-throughput sequencing platforms, the scale of scRNA-seq experiments has grown exponentially—from profiling a single cell in 2009 to exceeding one million cells per study by 2017 [47]. Today, the acquisition of scRNA-seq data is ubiquitously employed across diverse fields of biology and precision medicine [38,48,49].

Conventional bulk RNA-sequencing (bulk RNA-seq) typically profiles entire organs or tissue homogenates, yielding an averaged snapshot of global gene expression. This macroscopic approach inherently masks the profound spatial and temporal transcriptomic heterogeneity that exists among distinct cell subpopulations [50]. The emergence of scRNA-seq effectively resolves this limitation by distinguishing individual cell types and capturing independent, cell-specific expression profiles. As scRNA-seq technologies have rapidly matured, the primary research bottleneck has shifted from experimental data acquisition to the computational analysis of massive datasets. A standard scRNA-seq analytical pipeline typically encompasses raw data dimensionality reduction, denoising [51], and subsequent classification tasks—most notably, graph-based clustering [52] and cell-type annotation.

Cluster analysis is particularly pivotal, as it reveals cellular heterogeneity and developmental diversity by delineating specific cell types and states. However, compared to bulk RNA-seq, scRNA-seq datasets present unique and severe mathematical challenges. A defining obstacle is extreme data sparsity, characterized by the prevalent "dropout" phenomenon (zero-inflation), where expressed transcripts fail to be detected due to minute starting RNA quantities [53,54]. Furthermore, technical artifacts inevitably introduce batch effects when integrating datasets generated across different protocols, platforms, or biological replicates [55]. Compounding these issues is a high degree of biological and technical noise, including stochastic transcriptional bursting, mRNA degradation, and inadequate sequencing depth [56]. Because many features (genes) in scRNA-seq matrices are zero or near-zero, this pervasive noise obscures subtle transcriptomic differences between closely related cell states. Finally, the inherently high-dimensional nature of these datasets [57] exponentially increases the computational complexity of analysis, underscoring the absolute necessity for advanced, noise-tolerant machine learning algorithms.

To robustly identify cell types across novel samples, highly accurate computational methods are essential. In recent years, numerous machine learning classification algorithms specifically tailored for single-cell data have been proposed [58–65], significantly advancing data processing capabilities. Both supervised algorithms (such as SVM and logistic regression) and unsupervised approaches (like K-means clustering) are routinely deployed for cell-type identification. For instance, CIDR [66] tackles data sparsity through dropout-aware implicit imputation to adjust cell–cell dissimilarity matrices, which are subsequently processed using principal coordinate analysis (PCoA)—a classical multidimensional scaling technique—followed by hierarchical clustering, rather than relying on standard PCA. SC3 [67] integrates multiple distance metrics (Euclidean, Pearson, and Spearman) and transformations (PCA and graph Laplacian), executes parallel K-means clustering across multiple dimensions (d), and performs hierarchical clustering on the resulting consensus matrix to effectively decode cellular heterogeneity. SIMLR [68] employs multiple kernel

learning (MKL) to capture complex intercellular similarities based on gene expression profiles. Although these methods have substantially improved single-cell clustering efficiency, they often struggle to provide optimal solutions when processing highly sparse scRNA-seq matrices. Furthermore, many of these classical algorithms incur prohibitive computational and storage overhead, limiting their scalability to modern, million-cell atlases.

To overcome these limitations, the integration of deep learning with scRNA-seq analysis has emerged as a transformative frontier in the life sciences. Numerous innovative deep neural network architectures have been developed, providing powerful new paradigms for single-cell data representation. For example, DESC [69] is built upon a stacked Autoencoder that utilizes mean squared error (MSE) reconstruction loss paired with Kullback–Leibler (KL) divergence-based self-training within a deep embedded clustering (DEC) framework. Notably, it iteratively reduces batch effects implicitly without requiring prior batch labels or explicit zero-inflated negative binomial (ZINB) likelihood assumptions. Similarly, scBERT [70] strictly employs a Performer-based architecture leveraging FAVOR+ linear attention to efficiently process long gene sequences (over 16,000 genes). It bins gene expression values, embeds gene identities via gene2vec, and utilizes masked expression value prediction during its pre-training phase. Other models focus on specific clustering optimization: ADCluster [71] simultaneously performs anomaly detection and cluster analysis to isolate outliers while grouping normal cells, whereas scCAEs [72] utilizes a convolutional Autoencoder paired with a soft K-means deep embedding layer to identify latent cellular subpopulations. Furthermore, scDCCA [73] integrates denoising Autoencoders (DAEs) with contrastive learning modules to extract robust biological features.

Despite these advancements, many foundational deep learning algorithms do not explicitly model gene–gene regulatory mechanisms or intrinsic cellular topologies. Addressing this, the *DeepGRNCS* [74] framework utilizes deep neural networks to predict target gene expression based solely on the expression of transcription factors (TFs), bypassing the need for prior topological information in Gene Regulatory Network (GRN) inference. When benchmarked against existing bulk-based (e.g., PIDC, GENIE3) and single-cell-specific (e.g., NETI2, JEGN) GRN inference tools using simulated datasets, *DeepGRNCS* demonstrated superior clustering and network prediction performance. Its subsequent application to real scRNA-seq tumor datasets effectively mapped complex intra-tumoral heterogeneity. Finally, leveraging the power of large language models (LLMs), the CelltypeGPT [75] software package functions as a GPT-4 API-based prompting wrapper. Accessible as an R package, it efficiently annotates cell types by taking lists of marker or differentially expressed genes as input, operating as an intelligent annotation conduit rather than a standalone pre-trained foundation model.

Driven by the success of self-supervised representation learning, the single-cell community has witnessed an explosion of foundational models pre-trained on tens of millions of cells. The defining advantage of these Transformer-based foundation models lies in their robust cross-dataset generalization. Beyond Geneformer and scGPT, recent architectures such as scFoundation [21], UCE (Universal Cell Embedding) [76], and CellPLM [77] have significantly advanced cross-tissue and cross-species zero-shot annotations by learning universal cellular representations, allowing them to overcome batch effects and accurately perform cell-type annotation across novel, unseen biological contexts. Furthermore, beyond standard clustering and annotation, this generative capacity has unlocked entirely new predictive frontiers, most notably in perturbation prediction. Early VAE-based models like scGen [78] and CPA [79] laid the groundwork, while more recent models like GEARS [80] integrate Graph Neural Networks and deep learning to accurately simulate the transcriptional outcomes of combinatorial genetic perturbations in silico. This signifi-

cantly accelerates drug discovery and functional genomics without the immediate need for exhaustive laboratory screening.

It should be noted that most of the existing algorithms rely on gene expression information during characterization learning while not implementing sharing topological information between cells. In order to further capture the complex relationships between cells and their intrinsic properties, several algorithms based on Graph Neural Networks (GNNs) have been proposed. The GNN can reveal the connections between the target node and its surrounding nodes, thereby enhancing the representation of the node features. For example, scGAC [81] specifically relies on Graph Attention Networks (GATs) coupled with Network Enhancement for graph denoising and DEC-style self-training. scDFC [82] uses attribute information and structural information based on attention mechanisms to precisely construct intercellular maps to address complex biological situations. GNN can capture and represent complex higher-order structural relationships between cells, but existing GNN-based clustering methods easily map different classes of nodes to similar representations during encoding and cannot effectively distinguish different types of cell. The deep learning framework scZAG [83] combines APPNP-style personalized PageRank propagation with a ZINB Autoencoder, graph contrastive learning, and self-optimizing clustering to map intricate cellular relationships. The iterative propagation mechanism of APPNPGCN helps the scZAG model to better capture the complex relationships between cells. By comparing with other state-of-the-art scRNA-seq clustering methods in 10 real datasets, the results show that scZAG clustering performs better and can more effectively distinguish and isolate various cell populations. Another deep framework, DCRELM [84], is not a conventional single-layer classifier; rather, it is a hybrid deep model that integrates Extreme Learning Machine (ELM)-based random mapping, dual-view perturbation, Autoencoder fusion, and triplet self-supervised learning to stabilize feature representation for clustering. The learning machine performs highly by clustering comparison with the six latest single-cell clustering methods on 12 real scRNA-seq datasets.

While complex deep learning architectures have gained significant traction, a scientifically balanced perspective requires acknowledging that standard, highly optimized baseline methods remain the workhorses of practical single-cell analysis. For batch correction and data integration, classical frameworks often perform exceptionally well. Methods such as Harmony [85], Seurat v3/v4 (utilizing integration anchors and Weighted Nearest Neighbor (WNN) analysis), MNN/fastMNN (Mutual Nearest Neighbors) [86], LIGER (based on integrative non-negative matrix factorization) [87], and BBKNN are widely established standards. Similarly, for automated cell-type annotation, classical machine learning approaches frequently provide robust, lightweight, and highly interpretable alternatives to deep learning. Widely used standard tools, including SingleR [88], scmap, scPred, CellTypist, and Azimuth, leverage reference datasets and regularized classifiers to achieve highly accurate annotations, often serving as critical baselines against which newer foundation models must be rigorously benchmarked. Among deep generative models, the Variational Autoencoder (VAE) framework has become truly foundational for single-cell analysis. The introduction of scVI [17] established a robust generative probabilistic model that serves as the baseline for numerous modern pipelines, efficiently handling dropout, batch effects, and differential expression. The scvi-tools [22] ecosystem has since expanded into a comprehensive framework, providing highly specialized architectures: scANVI [89] leverages cell-type labels for semi-supervised annotation, totalVI [90] facilitates the integration of multimodal data (such as joint RNA and surface protein measurements), and scArches [23] utilizes transfer learning to map query datasets onto massive reference atlases without the computationally prohibitive need to retrain the entire model.

Further studies of individual cells will help in the personalized medicine field for a deeper understanding of the underlying processes in various developmental physiologies and disease systems. One study proposed DeepGSEA [91], which utilizes a prototype-based interpretable deep neural network combined with classification-based significance testing to capture complex gene set distributions, distinguishing it from traditional ranked GSEA. Compared with commonly used existing methods, DeepGSEA has a high ability to capture gene profile differences between phenotype groups. DeepIMAGER [92] converts gene-pair co-expression patterns into image-like histograms, subsequently applying Convolutional Neural Networks (CNNs) for supervised Gene Regulatory Network (GRN) inference, continuing the lineage of models like CNNC. Tested in six real scRNA-seq datasets, DeepIMAGER showed superior performance in comparison with 10 popular GRN inference tools. DeepIMAGER was applied to the scRNA-seq dataset of multiple myeloma and successfully detected a potential GRN of the TF of *RORC*, *MITF* and *FOXD2* in MM dendritic cells. Detailed test information is given in Table 2.

Table 2. Application of partial machine learning in the single-cell transcriptome.

Assignment	Model Name	Machine Learning Algorithm	Validation (Methods, Data Set, Results)	Published Time	Reference
Tackles sparsity through dropout-aware implicit imputation to adjust cell–cell dissimilarity for subsequent dimensionality reduction and clustering.	CIDR	PCoA (Principal Coordinate Analysis), Hierarchical clustering	Using simulation data and biological data, and compared with the standard PCA, t-SNE, ZIFA and RaceID algorithms, CIDR achieved good clustering results and ran time much faster than the other algorithms.	2017	[66]
Integrates multiple metrics and transformations to generate a robust consensus matrix for downstream hierarchical clustering.	SC3	Multiple distance metrics, PCA/graph Laplacian, Parallel k-means, Hierarchical consensus clustering	Using six published datasets compared with five commonly used scRNA-seq data clustering methods including tSNE + kmeans, pcaReduce, SNN-Cliq, SINCERA and SEURAT, SC3 had higher ARI values on all datasets and SC3 was able to identify rare cell types containing 1% or 10% cells.	2017	[67]
Learning the similarity measures between cells	SIMLR	Multi-kernel Learning, Rank Constraint, Graph Diffusion, K-means, t-SNE	Tested in human PBMC, mouse cortex and hippocampal datasets, compared with PCA + K-means, t-SNE + K-means and single-cell data analysis tool Seurat, showed higher ARI values and completed analysis of datasets containing hundreds of thousands of samples within minutes.	2017	[68]
Avoids explicit ZINB assumptions to iteratively reduce batch effects implicitly without requiring prior batch labels.	DESC	Stacked Autoencoder (MSE loss), KL-divergence self-training (DEC framework)	In the five datasets, compared with methods such as CCA, MNN, Seurat 3.0, scVI, BERMUDA, and scanorama, the ARI values of DESC were above 0.9, exceeding the remaining methods. And it can still maintain a high ARI when dealing with complex batch effects and large-scale datasets.	2019	[69]

Table 2. *Cont.*

Assignment	Model Name	Machine Learning Algorithm	Validation (Methods, Data Set, Results)	Published Time	Reference
Employs a Performer architecture to efficiently handle long gene sequences with masked expression-value prediction pretraining.	scBERT	Performer-based architecture (FAVOR+ linear attention), gene2vec	The scBERT was evaluated on six datasets, including cell type annotation, across batches, robustness on organ dataset, ability to find new cells, performance on large-scale and hyperparameter sensitivity on six datasets. In the cross-organ dataset, scBERT performance was comparable to the remaining methods, and in five additional ways, scBERT accuracy and macro F1 scores outperformed the other algorithms.	2022	[70]
Iterative optimization, merging microclusters with similar structure, yields high-quality clustering results and effective low-dimensional features	ADCluster	Deep Embedded Clustering	Testing with 11 real and 5 simulated scRNA-seq datasets showed that ADClust has superior cluster performance measures over other comparison methods (ARI, NMI, CA, FMI, SC) (including MultiK, SIMLR, scDeepCluster, SC3, scQcut, IKAP, CIDR, Seurat, SHARP, scGMAI, and DESC) with high scalability for large-scale datasets.	2022	[71]
Address the clustering of high missing rate or noisy scRNA-seq datasets	scCAEs	convolutional Auto-Encoders (CAE), Soft K-means clustering algorithm, KL divergence	Testing in eight real single-cell RNA-seq datasets covering different cell types, test results showed that scCAEs had higher ARI and NMI values than SIMLR, Seurat, SC3, SOUP and scziDesk in all datasets.	2022	[72]
Comparing the similarities and differences between positive and negative sample pairs, we learn more discriminatory feature representations and achieve more accurate clustering	scDCCA	AE, Contrastive learning, Deep embedded clustering, DEC	Fourteen real datasets reflect that scDCCA outperforms the eight state-of-the-art clustering methods in accuracy, generalization ability, scalability and efficiency, and scDCCA achieves the best or near best performance on CA, NMI and ARI	2023	[73]
Predicted for the expression of the target genes	DeepGRNCS	DNN	The model was validated using simulated datasets (based on Gaussian distribution, Boolean network and synthetic network) and real datasets (6 scRNA-seq datasets of BEELINE libraries and scRNA-seq datasets for lung cancer patients). The results indicate excellent AUROC and AUPRC values in all datasets.	2024	[74]
Functions as a prompting wrapper that efficiently annotates cell types by taking lists of marker or differentially expressed genes as input.	CelltypeGPT	GPT-4-API-based prompting wrapper (R package)	Using scRNA-seq data from different human and mouse tissues, GPTCelltype showed high agreement with higher average consistency scores than faster and lower cost of other automated annotation methods. It also showed good robustness and reproducibility in the simulation experiments	2024	[75]

Table 2. *Cont.*

Assignment	Model Name	Machine Learning Algorithm	Validation (Methods, Data Set, Results)	Published Time	Reference
Relies on GAT coupled with Network Enhancement for graph denoising to improve clustering performance.	scGAC	Graph Attention Networks (GAT), Network Enhancement, DEC-style self-training	Model clustering performance was evaluated using 16 real datasets, showing that the model achieved the highest ARI and NMI scores on multiple datasets, clustering results more similar to real labels, embedding cells in two-dimensional space using t-SNE or other dimension reduction methods, and clearer visualization results.	2022	[81]
Integrate information from attribute information AE and structural information AE based on attention mechanism to improve clustering performance	scDFC	Fusion Network	Eated using ten real single-cell RNA-sequencing datasets of varying sizes covering different species such as human, mouse and sponge, scDFC had the highest ARI and NMI scores, and the other four datasets also ranked the top three.	2023	[82]
Combines PageRank propagation with a ZINB Autoencoder and contrastive learning to map intricate cellular relationships.	scZAG	APPNP-style personalized PageRank propagation, ZINB Autoencoder, Graph contrastive learning Figure contrast learning, Self-optimized deep embedding clustering	Real scRNA-seq datasets from 10 different sequencing platforms were used to validate the performance of scZAG. The scZAG achieved the highest ARI scores on nine datasets and the highest NMI score on eight datasets. The Mann-Whitney U test results indicated significant differences between scZAG and other methods in ARI and NMI assessment indicators, further demonstrating the superior performance of scZAG.	2024	[83]
A hybrid deep framework integrating ELM mapping and triplet self-supervised learning to stabilize feature representation for clustering.	DCRELM	ELM-based random mapping, Dual-view perturbation, Autoencoder fusion, Triplet self-supervised learning	Using 12 real datasets with multiple cell types and states and six other state-of-the-art single cell clustering methods (scDeepCluster, GraphSCC, scGNN, DREAM, scDCCA, and scDFC), DCRELM outperformed the other methods in NMI, ARI, and F1 in most datasets.	2024	[84]
Utilizes a prototype-based deep neural network combined with classification-based testing to capture complex gene set distributions.	DeepGSEA	Prototype-based interpretable DNN, Classification-based significance testing	The ability of DeepGSEA was evaluated using four simulated scRNA-seq-seq datasets and three real scRNA-multiple phenoseq datasets. The results showed that DeepGSEA exhibited high sensitivity, interpretability and effective ability to process cellular heterogeneity.	2024	[91]
Converts gene-pair co-expression patterns into image-like histograms for supervised Gene Regulatory Network (GRN) inference.	DeepIMAGER	Convolutional Neural Networks (CNNs)	Compared with 10 existing Gene Regulatory Network inference methods (5 unsupervised and 5 supervised methods) on 6 different types of scRNA-seq and the corresponding ChIP-seq. DeepIMAGER achieves the highest AUROC score on all six datasets and has the minimum loss value and variance, outperforming other methods in prediction accuracy and robustness.	2024	[92]

It is crucial to emphasize that advanced downstream analyses, such as Gene Regulatory Network (GRN) inference and cell–cell communication analysis, should not be simply equated with a single class of machine learning methods. Rather, they are complex, multi-step analytical frameworks that rely heavily on prior biological assumptions and discrete model choices. For instance, GRN inference typically requires preprocessing steps to handle dropout, followed by the integration of prior knowledge (such as transcription factor binding motifs) to constrain the search space. Within this framework, machine learning models (such as the DNNs in *DeepGRNCS* or the CNNs in DeepIMAGER) do not operate as monolithic solutions; instead, they contribute to specific components, such as performing non-linear regression for link prediction or estimating the statistical weights of regulatory edges. Similarly, cell–cell communication analysis is fundamentally anchored in curated ligand–receptor interaction databases. When advanced algorithms (like spatially aware networks) are applied to this domain, their specific role is usually to calculate interaction probabilities, infer spatial proximity constraints, and denoise the underlying expression data. By recognizing machine learning as a specialized computational engine within a broader biological systems framework, researchers can avoid the "black box" trap and make more informed, interpretable analytical choices.

In summary, selecting an analytical framework for single-cell transcriptomics requires carefully balancing computational cost with biological necessity. While Variational Autoencoders (VAEs) excel at probabilistic integration and batch correction, they often lack the explicit ability to model intricate spatial or communication networks. Conversely, Graph Neural Networks (GNNs) are highly specialized for topology and trajectory mapping but can be sensitive to graph construction parameters. Meanwhile, the recent surge of Transformer-based foundation models offers unprecedented capabilities for cross-tissue zero-shot annotation; however, they require massive computational resources and, as recent evaluations suggest [93,94], may not inherently outperform classical linear baselines in simpler, localized tasks. Therefore, future pipeline development must prioritize algorithmic efficiency and rigorous, task-specific benchmarking [95] rather than merely increasing model parameters.

3.3. Machine Learning for Spatial Transcriptome Sequencing

The evolution of transcriptomic technologies has progressed through three pivotal stages. Initially, bulk RNA-sequencing profiled massive admixtures of cells, capturing only average gene expression levels and fundamentally masking cell-specific transcriptional dynamics. The subsequent advent of single-cell transcriptomics extended this resolution to individual cells, enabling researchers to decode organ function and disease progression at unprecedented cellular and subcellular levels [96]. However, traditional scRNA-seq requires the dissociation of solid tissues into single-cell suspensions, inevitably destroying the native spatial coordinates of each cell [97]. Preserving this spatial context is imperative for deeply understanding tissue structure–function relationships, localized cell–cell interactions, and microenvironmental regulation. While early in situ hybridization methods (e.g., FISH) laid the pioneering groundwork for spatial mapping, the formal term "spatial transcriptomics" (ST) was coined in 2016 when Lundeberg and colleagues introduced the first high-throughput, array-based technique for capturing RNA in situ [18]. By integrating ST with conventional scRNA-seq and other multi-omic modalities, researchers can comprehensively map cellular heterogeneity back to precise histological coordinates, offering profound insights into disease pathogenesis and targeted therapeutic design. Recognizing its transformative capacity—driven by continuous improvements in cellular throughput, transcript capture efficiency, and spatial resolution—Nature Methods crowned spatially resolved transcriptomics as the "Method of the Year" in 2020 [98].

Methodologically, scRNA-seq and ST data are highly complementary. While ST inherits several technical limitations characteristic of scRNA-seq—most notably low capture efficiency and severe dropout rates—it introduces entirely new dimensions of complexity. Although many computational pipelines originally developed for scRNA-seq can theoretically be adapted for ST analysis [20], such direct translations must be approached with caution. ST transcript pools do not necessarily follow the same statistical distributions as scRNA-seq data, primarily because standard ST capture spots often contain mixtures of multiple cells rather than pure, isolated single cells. Consequently, standard single-cell analytical assumptions must be rigorously reassessed. Furthermore, the sheer volume and multidimensional nature of ST data present formidable computational bottlenecks, underscoring the necessity for advanced machine learning interventions. Detailed benchmarking of these ML applications is summarized in Table 3.

Table 3. Application of partial machine learning in spatial transcriptomes.

Assignment	Model Name	Machine Learning Algorithm	Validation (Methods, Data Set, Results)	Published Time	Reference
To assess the degree of sample admixture in the potential space and to guide gene deletion value filling	gimVI	Variational Inference/Deep Generative Model	Validation was performed in 2 real mouse datasets and compared against scVI, Liger, Seurat and CORAL algorithms. The gimVI has the highest entropy mixing value for integrating cells into the joint latent space and the best k-NN purity on the mSMS dataset; for gene deletion filling, the Spearman correlation coefficient of gimVI is more than 38% higher than CORAL for the mSMS dataset.	2019	[99]
Integrating scRNA-seq and spatial transcriptome data to predict the expression patterns of unmeasured genes in the spatial transcriptome	SpaGE	PCA, SVD, KNN	The performance evaluation of SpaGE was conducted using six data sets (including one scRNA-seq data set from the same tissue and one spatial transcriptome data set). Compared with the three algorithms of Seurat, Liger, and gemVI, the correlation between SpaGE's predicted value and the true value (such as Spearman correlation coefficient) was significantly higher than other algorithms, and the prediction performance of marker genes in different cell types was also better than other algorithms. Although SpaGE outperformed other algorithms in most cases, it is possible that because the KNN algorithm used by SpaGE is relatively simple, it may not be fully captured for some complex gene expression patterns	2020	[100]
Predict the expression of the unmeasured genes and effectively fill in the expression information of the measured genes	stPlus	AE, KNN	A large dataset, MERFISH_Moffit, containing 64,373 spatial transcriptomic cells and 31,299 scRNA-seq, was used to verify the model efficiency of stPlus. In comparison with the four baseline methods of SpaGE, Seurat, Liger and gemVI, the gene-level Spearman correlation coefficient, cell-level Spearman correlation coefficient, AMI, ARI, Homo and NMI were calculated for each dataset. The results showed that stPlus significantly outperformed the baseline method in all the metrics.	2021	[101]

Table 3. *Cont.*

Assignment	Model Name	Machine Learning Algorithm	Validation (Methods, Data Set, Results)	Published Time	Reference
The scRNA-seq data were spatially aligned with the other spatial data	Tangram	DL	Five snRNA-seq data from human and mice were used to evaluate the performance of Tangram in terms of cell type distribution consistency, gene expression prediction accuracy, low quality data correction, compared with STARmap, Visium, Allen brain map data and SHARE-seq, which showed that Tangram showed good performance on different datasets. Although the performance of Tangram is influenced by the dataset characteristics and the data quality, for most datasets, Tangram is able to generate reliable prediction results.	2021	[102]
Convolution of spatial transcriptomic data, cell type localization, cell-state analysis and analysis of tumor microenvironment	SPOTlight	non-negative matrix factorization (NMF), Non-negative Least Squares (NNLS)	Synthetic datasets and human pancreatic cancer, mouse scRNA-seq brain and ST real datasets were used to evaluate the performance of SPOTlight. Compared with other bulk data and convolution tools (MuSiC, CIBERSORTx, DeconRNA-seq, SCDC, RCTD, NMFreg) and CoGAPS, SPOTlight showed high prediction accuracy on shallow sequencing or small-scale scRNA-seq reference datasets, with JSD values (Jensen-Shannon divergence) close to 0, and the predicted cell type proportion highly similar to the true proportion.	2021	[103]
The graph structure information in the link graph is used to learn the composition of different cell types at each spatial point to improve the prediction accuracy of the model	DSTG	GCN	In two synthetic spatial datasets, the JSD value of DSTG was significantly lower than the SPOTlight method; in three real datasets, DSTG accurately predicted the cellular composition of each spatial point in the spatial transcriptome data and revealed the cellular spatial structure of tissues.	2021	[104]
The proportion of each cell type was inferred	SpatialDWLS	DWLS	A low-resolution simulated spatial transcriptome dataset, mouse brain spatial transcriptome dataset and human heart ST dataset were used for the validation of the model. Compared with other deconvolution methods (MuSiC, RCTD, SPOTlight and stereoscope) on the simulated data set, spatialDWLS performed well in both accuracy and computational efficiency. On real datasets, spatialDWLS can quickly analyze large spatial transcriptomic datasets and reveal dynamic changes in the spatial distribution of cell types.	2021	[104,105]
Decdown cell type mixtures and accurately predict the proportion of each cell type in each pixel	RCTD	MLE, LR	The comparison with unsupervised clustering, NMFreg, and DWLS, on 2 scRNA-seq datasets and three spatial transcriptomic data sets showed that RCTD can accurately distinguish cell types from unsupervised clustering and NMFreg, and can handle platform effects.	2022	[106]

Table 3. *Cont.*

Assignment	Model Name	Machine Learning Algorithm	Validation (Methods, Data Set, Results)	Published Time	Reference
Using features of spatial omics data to improve the accuracy of spatial domain identification and batch effect correction	DeepST	GNN	Eight datasets with different types and spatial resolution are used to verify the model, and compared with six spatial and identification algorithms. DeepST uses deep learning model for feature extraction and spatial information modeling, which can effectively process spatial omics data from different platforms and organizations and improve the accuracy of spatial domain identification.	2022	[107]
Classification cell types to generate more reliable single-cell spatial gene expression profiles	STCellbin	Leiden clustering, Bi-Directional ConvLSTM U-Net, Cellpose 2.0	Using two containing high quality images of nucleus and cell membrane/cell wall staining, and the corresponding spatial gene expression data set in comparing cell segmentation performance with three algorithms and compared with Baysor tool in terms of performance in downstream analysis, STCellbin outperforms other algorithms, especially in dividing cell membrane/cell wall.	2024	[108]

What fundamentally distinguishes spatial transcriptomics from both bulk and single-cell data is the introduction of a new mathematical dimension: physical (x,y) coordinates and tissue histology. Algorithms designed exclusively for scRNA-seq treat cells as isolated entities and cannot process physical topology. To bridge this gap, Graph Neural Networks (GNNs) and spatially aware Convolutional Neural Networks (CNNs) were specifically designed for spatial data. By treating individual capture spots as nodes and their physical or histological proximity as edges, these specialized architectures can jointly optimize gene expression embeddings alongside physical tissue morphology, enabling the precise decoding of spatial domains and cell–cell communication microenvironments.

To enhance data resolution and overcome the inherent sparsity of spatial transcriptomics (ST), computational imputation—driven by deep learning, dimensionality reduction, or other machine learning algorithms—is frequently employed to predict missing gene expression profiles by leveraging matched single-cell RNA-sequencing (scRNA-seq) references. For example, the VAE-based generative model gimVI [99] integrates ST data with unlabeled scRNA-seq matrices. By utilizing conditional distributions, it effectively accounts for platform-specific technical effects and seamlessly processes multi-batch datasets. Alternatively, SpaGE [100] employs principal component analysis (PCA) and singular-value decomposition (SVD) to project both modalities into a shared latent space, subsequently utilizing *k*-Nearest Neighbor (KNN) interpolation to impute unmeasured genes. SpaGE demonstrated superior predictive accuracy over contemporary baselines (Seurat, gimVI, and LIGER) across diverse regions of the mouse brain. Furthermore, it exhibits exceptional scalability, significantly reducing computational and memory overhead when applied to massive spatial cohorts, such as MERFISH datasets comprising over 60,000 spatial measurement units. Building upon this shared-space paradigm, the Autoencoder-based model stPlus [101] leverages weighted KNN clustering to forecast spatial gene expression. Empirical evaluations utilizing Spearman correlation coefficients indicated that stPlus consistently outperformed mainstream predictive methods. Another prominent deep learning framework, Tangram [102], aligns scRNA-seq or single-nucleus RNA-seq (snRNA-seq) profiles directly to spatial coordinates. By correlating these profiles with matched histological and anatomical imaging from the same sample, Tangram resolves transcriptome-wide spatial

expression at true single-cell resolution, effectively overcoming the inherent throughput limitations of primary ST platforms.

Beyond gene imputation, machine learning is indispensable for spatial deconvolution—the computational estimation of discrete cell-type proportions within macroscopic ST capture spots. For instance, SPOTlight [103] integrates ST and scRNA-seq data via non-negative matrix factorization (NMF). Initialized with cell-type marker genes and non-negative least squares (NNLS), it deconvolves ST capture locations to map complex tissue architectures, successfully reconstructing the hierarchical layers of the mouse brain and detailing clinically relevant immune microenvironments in pancreatic adenocarcinoma. However, classical NMF approaches often fail to incorporate the intrinsic spatial topology of neighboring spots. To address this, DSTG (Deconvoluting Spatial Transcriptomics Data through Graph-based Artificial Intelligence) [104] deploys a semi-supervised graph convolutional network (GCN) that explicitly incorporates spatial neighborhood topologies alongside scRNA-seq references, significantly enhancing deconvolution accuracy in heterogeneous tissues. Alternatively, SpatialDWLS [105] utilizes dampened weighted least squares (DWLS) on enriched gene signatures to estimate local cell-type proportions, rigorously accounting for basal expression variances across different lineages. It has demonstrated exceptional accuracy and practicality across both simulated platforms and real Visium datasets. Similarly, RCTD (Robust Cell-Type Decomposition) [106] applies maximum likelihood estimation (MLE) to scRNA-seq profiles to resolve ST mixtures, introducing critical statistical corrections for cross-platform sequencing biases and enabling the precise mapping of subtle cellular subtypes.

Expanding beyond pure deconvolution, the multimodal deep learning framework DeepST [107] jointly embeds spatial coordinates, histological morphology, and gene expression to delineate distinct spatial domains. DeepST not only corrects massive batch effects but also generalizes across diverse ST platforms (e.g., MERFISH, Slide-seq, Stereo-seq). Notably, when applied to breast cancer ST datasets, it identified highly heterogeneous intra-tumoral regions previously undetected by conventional transcriptomic analyses. Most recently, the integration of Transformer-based architectures with spatial awareness has catalyzed the emergence of spatially resolved foundation models. For example, Nicheformer leverages spatially contextualized, self-supervised learning to construct comprehensive representations of tissue microenvironments, thereby unlocking unprecedented nuances in cell–cell communication modeling.

3.4. Trajectory Inference and Cellular Dynamics

A critical pillar of single-cell analysis, alongside clustering and integration, is the modeling of cellular dynamics, state transitions, and differentiation trajectories. Traditional trajectory inference methods construct pseudotime orderings to map developmental lineages; foundational tools in this space include the Monocle suite (up to Monocle 3) and PAGA (Partition-based Graph Abstraction), which efficiently map complex, disconnected topologies within massive datasets.

The conceptualization of RNA velocity [109] revolutionized this domain by utilizing the ratio of spliced to unspliced transcripts to predict the future transcriptional states of individual cells. This was rapidly advanced by machine learning frameworks such as scVelo, which introduced a likelihood-based dynamical model to resolve transient cellular states and complex kinetics. Building upon this foundation, sophisticated tools like CellRank and its iteration CellRank 2 [110] combine RNA velocity with robust graph-based machine learning (e.g., Markov chains) to uncover probabilistic cell fate decisions. These algorithms effectively map complex multi-lineage trajectories, seamlessly identifying initial, terminal, and intermediate cellular states.

3.5. Machine Learning for Spatiotemporal Transcriptomics

While single-cell and spatial transcriptomics have fundamentally resolved cellular heterogeneity and spatial organization, respectively, early iterations of these technologies lacked a temporal dimension. Spatiotemporal transcriptomics has rapidly emerged as a transformative technology that concurrently quantifies gene expression, precisely maps cellular spatial coordinates, and tracks these variables across discrete developmental or disease time points. By sequencing consecutive tissue sections across distinct temporal stages at single-cell resolution, researchers can dynamically trace the spatial trajectories of specific cell lineages. This multidimensional approach facilitates the discovery of rare cell types and elucidates the regulatory mechanisms governing cell differentiation and organogenesis [111]. Highlighting its immense research potential, spatiotemporal omics was prominently featured as a transformative technology by *Nature* in 2022 [112], coinciding with the establishment of the international Spatio-Temporal Omics Consortium (STOC) in Shenzhen, China.

As a high-throughput methodology integrating space and time, spatiotemporal transcriptomics is being extensively applied across diverse biological kingdoms. In animal models, it is pivotal for tracking tumor microenvironment evolution, mapping nervous system development, and characterizing immune responses. In botany, its applications span plant organ development, morphological evolution, and stress resistance. Because plant cells possess unique, rigid structures, such as cell walls, vacuoles, and chloroplasts, standard cell segmentation and clustering algorithms initially struggled to perform accurately. Recently, however, ML algorithms utilizing specialized convolutional architectures and deep watershed models are being explicitly adapted to recognize plant-specific morphological boundaries, substantially expanding spatiotemporal applications in plant biology. Similarly, while applications in microbiology remain nascent, deep learning networks are increasingly deployed to model host–pathogen spatiotemporal interactions. For instance, spatially aware graph algorithms have demonstrated strong potential in tracking infectious diseases, enabling the decoding of virus-specific host responses and the precise spatial resolution of microenvironmental damage caused by SARS-CoV-2 [113,114].

A monumental milestone in spatiotemporal omics was the 2021 introduction of Stereo-seq, developed by BGI-Research (Shenzhen, China) in conjunction with international collaborators. Stereo-seq is a pioneering spatial transcriptomic technology that achieves unprecedented subcellular resolution (500 nm) coupled with a centimeter-scale panoramic field of view (up to 13 cm × 13 cm). By integrating DNA nanoball (DNB) patterned arrays with in situ RNA capture technologies, Stereo-seq decodes the spatiotemporal dynamics of gene expression across massive tissue areas without sacrificing high throughput. This technology has demonstrated exceptionally broad applicability, having been successfully utilized to map the adult mouse brain and track sagittal sections of mouse embryos across various developmental stages, thereby validating its adaptability for profound biological discovery. A detailed workflow of the Stereo-seq methodology is illustrated in Figure 4 [115].

To alleviate the analytical bottlenecks associated with the highly complex, multidimensional spatial data generated by Stereo-seq, BGI-Research developed the Spatial Analysis Workflow (SAW). This comprehensive software suite integrates multiple machine learning algorithms to streamline data processing and assist researchers in deciphering intricate biological structures. The SAW pipeline encompasses a complete suite of standard procedures, including raw data preprocessing, quality assessment, batch effect correction, cell clustering, cell-type identification, and spatial network analysis. Notably, during the preprocessing phase, the STCellbin [108] module deploys the deep learning-based universal

segmentation algorithm, Cellpose 2.0, to achieve highly accurate, morphology-aware cell segmentation.

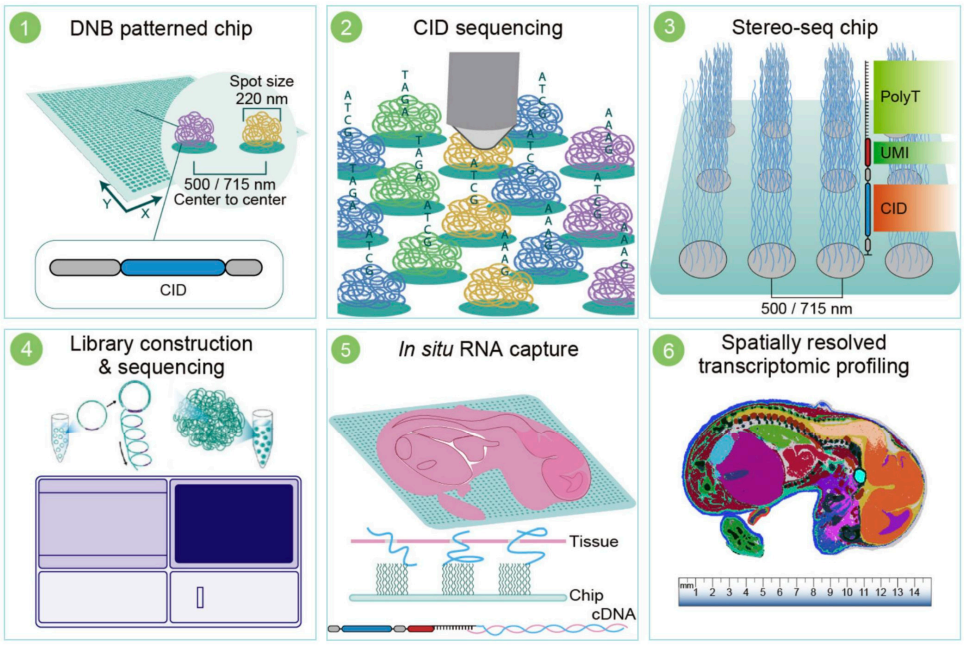

Figure 4. Stereo-seq pipeline (Adapted from Chen et al. [115] under the terms of the Creative Commons CC-BY license.) Step 1, design of the DNB mode chip. DNB libraries containing random barcodes were designed and their templates were deposited on the chip. The chips were sequenced in situ using primers to determine the spatial coordinates (CID) of each DNB. Step 2, in situ sequencing to determine the spatial coordinates of uniquely barcoded oligonucleotides. Step 3, preparation of capture probes by ligating the UMI-polyT containing oligonucleotides to each spot. Step 4, in situ RNA capture from tissue. Step 5, Libraries were constructed and sequenced. The amplified cDNA was used to construct the sequencing libraries and was sequenced together with the CID. Step 6, Data analysis of the sequencing data was performed using the computational analysis tool.

Ultimately, the defining methodological distinction between single-cell and spatial transcriptomics analysis lies in the incorporation of physical coordinates. While traditional single-cell machine learning focuses exclusively on transcriptional similarity, spatial algorithms, particularly those leveraging spatially aware GNNs and spatial Transformers, must jointly optimize gene expression embeddings alongside physical tissue topology. A critical synthesis of these methodologies reveals a clear consensus: integrating multimodal inputs [4] through hybrid neural architectures represents the most robust computational strategy for uncovering genuine tissue microenvironments [4,116].

3.6. Machine Learning for Multimodal Data Integration

Modern transcriptomics increasingly relies on multimodal data to capture comprehensive cellular states. Machine learning is uniquely positioned to bridge scRNA-seq with epigenomic (e.g., scATAC-seq), proteomic (e.g., CITE-seq), and histological imaging modalities. Early probabilistic frameworks, such as totalVI [90], established the foundation by enabling the joint modeling of RNA and surface proteins. More recently, deep learning has driven sophisticated cross-modality translations. For instance, the BABEL algorithm utilizes a deep neural network (DNN) to integrate chromatin profiles with transcriptomes, accurately predicting single-cell gene expression directly from chromatin accessibility (scATAC-seq) data [117]. Similarly, Hist2ST [118] is a specialized deep learning model capable of generating spatial transcriptomic maps directly from standard histological images. Furthermore, ConGI [119] employs CNN- and MLP-based architectures

via contrastive learning to seamlessly integrate histopathological imaging with spatial transcriptomics. This approach effectively decodes spatial domains by learning common cross-modal embeddings while rigorously isolating modality-specific noise.

4. Existing Challenges and Future Prospects

Machine learning (ML) techniques have been applied to transcriptomic research for nearly two decades, fundamentally focusing on extracting biological insights from RNA sequences, epigenetic modification sites, and gene expression profiles. The massive scale and inherent complexity of transcriptome sequencing data pose significant challenges for traditional analytical methods, particularly regarding high-throughput processing, pattern recognition, and multi-omic integration. However, these bottlenecks can be navigated much more efficiently and robustly utilizing ML technologies. While ML has not entirely replaced classical statistical frameworks, the exponential increase in ML-focused transcriptomic publications underscores its profound and expanding potential. Indeed, ML has revolutionized transcriptomic data analysis, not only by enhancing computational efficiency and analytical depth but also by providing unprecedented perspectives for biological discovery and clinical application. ML offers several distinct advantages: it automates the processing of massive datasets, significantly reduces the need for manual intervention, and seamlessly integrates principles from biology, statistics, and computer science. By autonomously learning data distributions, extracting latent features, and uncovering hidden biological relationships, ML fundamentally improves both processing efficiency and diagnostic accuracy. Furthermore, ML algorithms excel at integrating transcriptomic data across disparate platforms and experimental conditions, effectively mitigating batch effects to yield highly comprehensive interpretations. Particularly with the rapid evolution of deep learning (DL), a massive influx of novel neural network architectures has been successfully adapted for transcriptomics.

Despite these remarkable advancements, the application of ML to transcriptomic data exposes several critical challenges, primarily stemming from the extreme diversity and biological complexity of the data. First, the high-dimensional nature of sequencing data frequently predisposes complex ML models to severe overfitting. Second, extreme class imbalances (e.g., highly variable sample sizes across different disease states or rare cell types) often bias models toward majority classes, complicating the training of stable algorithms on small or rare cohorts. Third, the profound spatial heterogeneity and dynamic temporal shifts inherent to cellular states require models capable of mapping 4D spatiotemporal trajectories—a formidable algorithmic hurdle. Furthermore, state-of-the-art ML models often require massive computational resources. Deep neural networks containing millions of parameters inherently function as opaque "black boxes," making their biological decision-making processes difficult to understand or validate. To overcome this limitation, the integration of Explainable AI (XAI) techniques, such as SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-agnostic Explanations), and GNNExplainer, is becoming essential to demystify these complex architectures and provide transparent, biologically meaningful feature attributions. Finally, the steep interdisciplinary learning curve poses a practical barrier; biologists and clinicians without extensive computational backgrounds may struggle to deploy these models effectively, hindering their ultimate translation into routine clinical practice. Consequently, future algorithmic development must prioritize high computational efficiency, rigorous interpretability, and low resource consumption.

Despite the rapid proliferation and substantial parameter scale of single-cell foundation models, their genuine utility in specific downstream tasks remains a subject of critical ongoing debate. Recent rigorous evaluations have suggested that in many zero-shot set-

tings and specific batch-integration scenarios, these resource-intensive foundation models do not consistently outperform classical, simple statistical baselines. For example, critical analyses by Kedzierska et al. (2025) [120] demonstrated that basic logistic regression or non-deep-learning Nearest Neighbor approaches often match or even exceed the performance of pre-trained Transformers on out-of-the-box cell-type annotation and embedding tasks. These findings highlight a significant current challenge: while representation learning offers vast potential, there is an urgent need to establish standardized, rigorous benchmarking frameworks to ensure that the deployment of complex machine learning models translates to tangible biological utility rather than mere computational overhead.

5. Conclusions

Machine learning (ML) has demonstrated immense potential and catalyzed significant advancements within the field of transcriptomics. It consistently outperforms traditional analytical frameworks by excelling in high-throughput data processing, robust latent feature extraction, noise tolerance, and cross-dataset generalization. Nonetheless, formidable challenges persist. The extreme dimensionality of transcriptomic profiles, pervasive technical batch effects, the inherent "black-box" opacity of complex neural networks, and the substantial computational overhead required for model training continue to limit broader deployment. Addressing these bottlenecks will necessitate the rigorous refinement of current algorithms, prioritizing computational efficiency and robust interpretability. As spatial and single-cell sequencing technologies continue to evolve and data acquisition costs plummet, the full transformative capacity of ML in transcriptomics will be realized. Ultimately, these intelligent analytical pipelines will drive profound breakthroughs across precision medicine, agricultural genetics, fundamental biological sciences, and environmental protection.

Author Contributions: Conceptualization, Y.Z., Z.H., L.Z. and L.S.; writing—original draft preparation, Y.Z. and Z.H.; writing—review and editing, Y.Z., Z.H., Y.L. and M.G. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by China Agriculture Research System (CARS-35); Sichuan Science and Technology Program (2021YFYZ0007, 2021ZDZX0008, 2021YFYZ0030); The Program for Pig Industry Technology System Innovation Team of Sichuan Province (SCCXTD-2024-8); National Center of Technology Innovation for Pigs (NCTIP-XD/C13).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

NGS	Second-generation sequencing
ML	Machine Learning
SL	Supervised Learning
USL	Unsupervised Learning
DL	Deep Learning
SVM	Support Vector Machine
DT	Decision Tree
RF	Random Forest

t-SNE	T-distributed random neighborhood embedding
AE	Autoencoder
SAE	Stacked Autoencoder
RNN	Recurrent Neural Network
CNN	Convolutional Neural Networks
RNA-seq	RNA-Sequencing
RPKM	Reads Per Kilobase per Million mapped reads
FPKM	Fragments Per Kilobase per Million
TPM	Transcripts Per Million
KNN	K-Nearest Neighbor
RA	Rheumatoid arthritis
CART	Classification and regression tree
GBM	Gradient Boosting Machine
NNET	Neural Network
PLS	Partial Least Squares
scRNA-seq	single-cell RNA-sequencing
DAEs	Denoising Autoencoders
TF	Transcription Factor
GRN	Gene regulatory Network
GPT-4	Generative pre-trained transformer 4
GNN	Graph Neural Network
ELM	Extreme Learning Machines
CAE	Convolutional Autoencoders
VI	Variational Inference
GAN	Generative Adversarial Network
NMF	Non-negative matrix factorization
NNLS	Non-negative Least Squares
DSTG	Deconvoluting Spatial Transcriptomics Data Through Graph-based Artificial Intelligence
DWLS	Dampened Weighted Least Squares
RCTD	Robust Decomposition of Cell Type Mixtures
SAW	Spatial Analysis Workflow

References

1. Stark, R.; Grzelak, M.; Hadfield, J. RNA sequencing: The teenage years. *Nat. Rev. Genet.* **2019**, *20*, 631–656. [CrossRef] [PubMed]
2. Green, E.D.; Rubin, E.M.; Olson, M.V. The future of DNA sequencing. *Nature* **2017**, *550*, 179–181. [CrossRef]
3. Kharchenko, P.V. The triumphs and limitations of computational methods for scRNA-seq. *Nat. Methods* **2021**, *18*, 723–732. [CrossRef]
4. Vandereyken, K.; Sifrim, A.; Thienpont, B.; Voet, T. Methods and applications for single-cell and spatial multi-omics. *Nat. Rev. Genet.* **2023**, *24*, 494–515. [CrossRef]
5. Greener, J.G.; Kandathil, S.M.; Moffat, L.; Jones, D.T. A guide to machine learning for biologists. *Nat. Rev. Mol. Cell Biol.* **2022**, *23*, 40–55. [CrossRef]
6. Hu, T.; Chitnis, N.; Monos, D.; Dinh, A. Next-generation sequencing technologies: An overview. *Hum. Immunol.* **2021**, *82*, 801–811. [CrossRef]
7. Bhandari, N.; Walambe, R.; Kotecha, K.; Khare, S.P. A comprehensive survey on computational learning methods for analysis of gene expression data. *Front. Mol. Biosci.* **2022**, *9*, 907150. [CrossRef]
8. Xue, B.; Zhang, M.; Browne, W.N.; Yao, X. A survey on evolutionary computation approaches to feature selection. *IEEE Trans. Evol. Comput.* **2015**, *20*, 606–626. [CrossRef]
9. Xu, Y.; Yan, Y.; Zhou, T.; Chun, J.; Tu, Y.; Yang, X.; Qin, J.; Ou, L.; Ye, L.; Liu, F. Genome-wide transcriptome and gene family analysis reveal candidate genes associated with potassium uptake of maize colonized by arbuscular mycorrhizal fungi. *BMC Plant Biol.* **2024**, *24*, 838. [CrossRef] [PubMed]
10. Zhao, X.; Dou, J.; Cao, J.; Wang, Y.; Gao, Q.; Zeng, Q.; Liu, W.; Liu, B.; Cui, Z.; Teng, L.; et al. Uncovering the potential differentially expressed miRNAs as diagnostic biomarkers for hepatocellular carcinoma based on machine learning in The Cancer Genome Atlas database. *Oncol. Rep.* **2020**, *43*, 1771–1784. [CrossRef]

11. Cui, Q.; Yang, H.; Gu, Y.; Zong, C.; Chen, X.; Lin, Y.; Sun, H.; Shen, Y.; Zhu, J. RNA sequencing (RNA-seq) analysis of gene expression provides new insights into hindlimb unloading-induced skeletal muscle atrophy. *Ann. Transl. Med.* **2020**, *8*, 1595. [CrossRef]
12. Štorchová, H.; Krüger, M. Methods for assembling complex mitochondrial genomes in land plants. *J. Exp. Bot.* **2024**, *75*, 5169–5174. [CrossRef] [PubMed]
13. Reel, P.S.; Reel, S.; Pearson, E.; Trucco, E.; Jefferson, E. Using machine learning approaches for multi-omics data analysis: A review. *Biotechnol. Adv.* **2021**, *49*, 107739. [CrossRef]
14. Stathopoulou, K.M.; Georgakopoulos, S.; Tasoulis, S.; Plagianakos, V.P. Investigating the overlap of machine learning algorithms in the final results of RNA-seq analysis on gene expression estimation. *Health Inf. Sci. Syst.* **2024**, *12*, 14. [CrossRef]
15. Bzdok, D.; Altman, N.; Krzywinski, M. Statistics versus machine learning. *Nat. Methods* **2018**, *15*, 233–234. [CrossRef] [PubMed]
16. Heumos, L.; Schaar, A.C.; Lance, C.; Litinetskaya, A.; Drost, F.; Zappia, L.; Lücken, M.D.; Strobl, D.C.; Henao, J.; Curion, F.; et al. Best practices for single-cell analysis across modalities. *Nat. Rev. Genet.* **2023**, *24*, 550–572. [CrossRef]
17. Lopez, R.; Regier, J.; Cole, M.B.; Jordan, M.I.; Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **2018**, *15*, 1053–1058. [CrossRef] [PubMed]
18. Ståhl, P.L.; Salmén, F.; Vickovic, S.; Lundmark, A.; Navarro, J.F.; Magnusson, J.; Giacomello, S.; Asp, M.; Westholm, J.O.; Huss, M.; et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **2016**, *353*, 78–82. [CrossRef]
19. Hu, J.; Li, X.; Coleman, K.; Schroeder, A.; Ma, N.; Irwin, D.J.; Lee, E.B.; Shinohara, R.T.; Li, M. SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nat. Methods* **2021**, *18*, 1342–1351. [CrossRef]
20. Zeng, Z.; Li, Y.; Li, Y.; Luo, Y. Statistical and machine learning methods for spatially resolved transcriptomics data analysis. *Genome Biol.* **2022**, *23*, 83. [CrossRef]
21. Hao, M.; Gong, J.; Zeng, X.; Liu, C.; Guo, Y.; Cheng, X.; Wang, T.; Ma, J.; Zhang, X.; Song, L. Large-scale foundation model on single-cell transcriptomics. *Nat. Methods* **2024**, *21*, 1481–1491. [CrossRef]
22. Gayoso, A.; Lopez, R.; Xing, G.; Boyeau, P.; Valiollah Pour Amiri, V.; Hong, J.; Wu, K.; Jayasuriya, M.; Mehlman, E.; Langevin, M.; et al. A Python library for probabilistic analysis of single-cell omics data. *Nat. Biotechnol.* **2022**, *40*, 163–166. [CrossRef] [PubMed]
23. Lotfollahi, M.; Naghipourfar, M.; Luecken, M.D.; Khajavi, M.; Büttner, M.; Wagenstetter, M.; Avsec, Ž.; Gayoso, A.; Yosef, N.; Interlandi, M.; et al. Mapping single-cell data to reference atlases by transfer learning. *Nat. Biotechnol.* **2022**, *40*, 121–130. [CrossRef] [PubMed]
24. Long, Y.; Ang, K.S.; Li, M.; Chong, K.L.K.; Sethi, R.; Zhong, C.; Xu, H.; Ong, Z.; Sachaphibulkij, K.; Chen, A.; et al. Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with GraphST. *Nat. Commun.* **2023**, *14*, 1155. [CrossRef] [PubMed]
25. Theodoris, C.V.; Xiao, L.; Chopra, A.; Chaffin, M.D.; Al Sayed, Z.R.; Hill, M.C.; Mantineo, H.; Brydon, E.M.; Zeng, Z.; Liu, X.S.; et al. Transfer learning enables predictions in network biology. *Nature* **2023**, *618*, 616–624. [CrossRef]
26. Cui, H.; Wang, C.; Maan, H.; Pang, K.; Luo, F.; Duan, N.; Wang, B. scGPT: Toward building a foundation model for single-cell multi-omics using generative AI. *Nat. Methods* **2024**, *21*, 1470–1480. [CrossRef]
27. Hao, Y.; Hao, S.; Andersen-Nissen, E.; Mauck, W.M., 3rd; Zheng, S.; Butler, A.; Lee, M.J.; Wilk, A.J.; Darby, C.; Zager, M.; et al. Integrated analysis of multimodal single-cell data. *Cell* **2021**, *184*, 3573–3587.e29. [CrossRef]
28. Eraslan, G.; Avsec, Ž.; Gagneur, J.; Theis, F.J. Deep learning: New computational modelling techniques for genomics. *Nat. Rev. Genet.* **2019**, *20*, 389–403. [CrossRef]
29. Wang, Z.; Gerstein, M.; Snyder, M. RNA-Seq: A revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **2009**, *10*, 57–63. [CrossRef]
30. Chen, C.; Wang, J.; Pan, D.; Wang, X.; Xu, Y.; Yan, J.; Wang, L.; Yang, X.; Yang, M.; Liu, G.P. Applications of multi-omics analysis in human diseases. *MedComm* **2023**, *4*, e315. [CrossRef]
31. Wiens, J.; Saria, S.; Sendak, M.; Ghassemi, M.; Liu, V.X.; Doshi-Velez, F.; Jung, K.; Heller, K.; Kale, D.; Saeed, M.; et al. Do no harm: A roadmap for responsible machine learning for health care. *Nat. Med.* **2019**, *25*, 1337–1340. [CrossRef]
32. Liu, H.; Wang, W. Genetic algorithm and support vector machine-based gene microarray analysis. *J. Clin. Rehabil. Tissue Eng. Res.* **2010**, *14*, 3099–3103. [CrossRef]
33. Guo, M.; Yang, S.; Zhao, L. Transcriptome Analysis Method Based on RNA-Seq. *Comput. Sci.* **2020**, *47*, 35–39.
34. Su, C.; Tong, J.; Wang, F. Mining genetic and transcriptomic data using machine learning approaches in Parkinson's disease. *NPJ Park. Dis.* **2020**, *6*, 24. [CrossRef]
35. Potashkin, J.A.; Santiago, J.A.; Ravina, B.M.; Watts, A.; Leontovich, A.A. Biosignatures for Parkinson's disease and atypical parkinsonian disorders patients. *PLoS ONE* **2012**, *7*, e43595. [CrossRef] [PubMed]
36. Shamir, R.; Klein, C.; Amar, D.; Vollstedt, E.J.; Bonin, M.; Usenovic, M.; Wong, Y.C.; Maver, A.; Poths, S.; Safer, H.; et al. Analysis of blood-based gene expression in idiopathic Parkinson disease. *Neurology* **2017**, *89*, 1676–1683. [CrossRef] [PubMed]

37. Greene, C.S.; Krishnan, A.; Wong, A.K.; Ricciotti, E.; Zelaya, R.A.; Himmelstein, D.S.; Zhang, R.; Hartmann, B.M.; Zaslavsky, E.; Sealfon, S.C.; et al. Understanding multicellular function and disease with human tissue-specific networks. *Nat. Genet.* **2015**, *47*, 569–576. [CrossRef]
38. Català, P.; Groen, N.; LaPointe, V.L.S.; Dickman, M.M. A single-cell RNA-seq analysis unravels the heterogeneity of primary cultured human corneal endothelial cells. *Sci. Rep.* **2023**, *13*, 9361. [CrossRef]
39. Štancl, P.; Karlić, R. Machine learning for pan-cancer classification based on RNA sequencing data. *Front. Mol. Biosci.* **2023**, *10*, 1285795. [CrossRef] [PubMed]
40. Conway, A.M.; Mitchell, C.; Kilgour, E.; Brady, G.; Dive, C.; Cook, N. Molecular characterisation and liquid biomarkers in Carcinoma of Unknown Primary (CUP): Taking the 'U' out of 'CUP'. *Br. J. Cancer* **2019**, *120*, 141–153. [CrossRef]
41. Ng, S.; Masarone, S.; Watson, D.; Barnes, M.R. The benefits and pitfalls of machine learning for biomarker discovery. *Cell Tissue Res.* **2023**, *394*, 17–31. [CrossRef]
42. Andersson, B.; Langen, B.; Liu, P.; Dávila López, M. Development of a machine learning framework for radiation biomarker discovery and absorbed dose prediction. *Front. Oncol.* **2023**, *13*, 1156009. [CrossRef]
43. Li, L.; Weinberg, C.R.; Darden, T.A.; Pedersen, L.G. Gene selection for sample classification based on gene expression data: Study of sensitivity to choice of parameters of the GA/KNN method. *Bioinformatics* **2001**, *17*, 1131–1142. [CrossRef]
44. Sethi, S.; Shakyawar, S.; Reddy, A.S.; Patel, J.C.; Guda, C. A Machine Learning Model for the Prediction of COVID-19 Severity Using RNA-Seq, Clinical, and Co-Morbidity Data. *Diagnostics* **2024**, *14*, 1284. [CrossRef]
45. Wolf, F.A.; Angerer, P.; Theis, F.J. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol.* **2018**, *19*, 15. [CrossRef] [PubMed]
46. Asada, K.; Takasawa, K.; Machino, H.; Takahashi, S.; Shinkai, N.; Bolatkan, A.; Kobayashi, K.; Komatsu, M.; Kaneko, S.; Okamoto, K.; et al. Single-Cell Analysis Using Machine Learning Techniques and Its Application to Medical Research. *Biomedicines* **2021**, *9*, 1513. [CrossRef] [PubMed]
47. Tang, F.; Barbacioru, C.; Wang, Y.; Nordman, E.; Lee, C.; Xu, N.; Wang, X.; Bodeau, J.; Tuch, B.B.; Siddiqui, A.; et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **2009**, *6*, 377–382. [CrossRef]
48. Kan, T.; Zhang, S.; Zhou, S.; Zhang, Y.; Zhao, Y.; Gao, Y.; Zhang, T.; Gao, F.; Wang, X.; Zhao, L.; et al. Single-cell RNA-seq recognized the initiator of epithelial ovarian cancer recurrence. *Oncogene* **2022**, *41*, 895–906. [CrossRef] [PubMed]
49. Anjos-Afonso, F.; Buettner, F.; Mian, S.A.; Rhys, H.; Perez-Lloret, J.; Garcia-Albornoz, M.; Rastogi, N.; Ariza-McNaughton, L.; Bonnet, D. Single cell analyses identify a highly regenerative and homogenous human CD34+ hematopoietic stem cell population. *Nat. Commun.* **2022**, *13*, 2048. [CrossRef]
50. Lister, R.; O'Malley, R.C.; Tonti-Filippini, J.; Gregory, B.D.; Berry, C.C.; Millar, A.H.; Ecker, J.R. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* **2008**, *133*, 523–536. [CrossRef]
51. Wang, D.; Gu, J. VASC: Dimension Reduction and Visualization of Single-cell RNA-seq Data by Deep Variational Autoencoder. *Genom. Proteom. Bioinform.* **2018**, *16*, 320–331. [CrossRef]
52. Yang, L.; Liu, J.; Lu, Q.; Riggs, A.D.; Wu, X. SAIC: An iterative clustering approach for analysis of single cell RNA-seq data. *BMC Genom.* **2017**, *18*, 689. [CrossRef]
53. Kharchenko, P.V.; Silberstein, L.; Scadden, D.T. Bayesian approach to single-cell differential expression analysis. *Nat. Methods* **2014**, *11*, 740–742. [CrossRef]
54. Gong, W.; Kwak, I.Y.; Pota, P.; Koyano-Nakagawa, N.; Garry, D.J. DrImpute: Imputing dropout events in single cell RNA sequencing data. *BMC Bioinform.* **2018**, *19*, 220. [CrossRef]
55. Büttner, M.; Miao, Z.; Wolf, F.A.; Teichmann, S.A.; Theis, F.J. A test metric for assessing single-cell RNA-seq batch correction. *Nat. Methods* **2019**, *16*, 43–49. [CrossRef]
56. Fei, T.; Yu, T. scBatch: Batch-effect correction of RNA-seq data through sample distance matrix adjustment. *Bioinformatics* **2020**, *36*, 3115–3123. [CrossRef]
57. Linderman, G.C. Dimensionality Reduction of Single-Cell RNA-Seq Data. *Methods Mol. Biol.* **2021**, *2284*, 331–342. [CrossRef]
58. Liu, Y.; Zhang, H.; Mao, Y.; Shi, Y.; Wang, X.; Shi, S.; Hu, D.; Liu, S. Bulk and single-cell RNA-sequencing analyses along with abundant machine learning methods identify a novel monocyte signature in SKCM. *Front. Immunol.* **2023**, *14*, 1094042. [CrossRef]
59. Zhang, Z.; Mou, L.; Pu, Z.; Zhuang, X. Construction of a hepatocytes-related and protein kinase-related gene signature in HCC based on ScRNA-Seq analysis and machine learning algorithm. *J. Physiol. Biochem.* **2023**, *79*, 771–785. [CrossRef]
60. Chen, J.; Wang, X.; Ma, A.; Wang, Q.E.; Liu, B.; Li, L.; Xu, D.; Ma, Q. Deep transfer learning of cancer drug responses by integrating bulk and single-cell RNA-seq data. *Nat. Commun.* **2022**, *13*, 6494. [CrossRef]
61. Wang, Y.; Huang, Z.; Xiao, Y.; Wan, W.; Yang, X. The shared biomarkers and pathways of systemic lupus erythematosus and metabolic syndrome analyzed by bioinformatics combining machine learning algorithm and single-cell sequencing analysis. *Front. Immunol.* **2022**, *13*, 1015882. [CrossRef]
62. Cortese, R.; Adams, T.S.; Cataldo, K.H.; Hummel, J.; Kaminski, N.; Kheirandish-Gozal, L.; Gozal, D. Single-cell RNA-seq uncovers cellular heterogeneity and provides a signature for paediatric sleep apnoea. *Eur. Respir. J.* **2023**, *61*, 2201465. [CrossRef] [PubMed]

63. Chi, H.; Zhao, S.; Yang, J.; Gao, X.; Peng, G.; Zhang, J.; Xie, X.; Song, G.; Xu, K.; Xia, Z.; et al. T-cell exhaustion signatures characterize the immune landscape and predict HCC prognosis via integrating single-cell RNA-seq and bulk RNA-sequencing. *Front. Immunol.* **2023**, *14*, 1137025. [CrossRef] [PubMed]
64. Ding, J.; Sharon, N.; Bar-Joseph, Z. Temporal modelling using single-cell transcriptomics. *Nat. Rev. Genet.* **2022**, *23*, 355–368. [CrossRef]
65. Wu, Z.; Uhl, B.; Gires, O.; Reichel, C.A. A transcriptomic pan-cancer signature for survival prognostication and prediction of immunotherapy response based on endothelial senescence. *J. Biomed. Sci.* **2023**, *30*, 21. [CrossRef]
66. Lin, P.; Troup, M.; Ho, J.W. CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biol.* **2017**, *18*, 59. [CrossRef]
67. Kiselev, V.Y.; Kirschner, K.; Schaub, M.T.; Andrews, T.; Yiu, A.; Chandra, T.; Natarajan, K.N.; Reik, W.; Barahona, M.; Green, A.R.; et al. SC3: Consensus clustering of single-cell RNA-seq data. *Nat. Methods* **2017**, *14*, 483–486. [CrossRef]
68. Wang, B.; Zhu, J.; Pierson, E.; Ramazzotti, D.; Batzoglou, S. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nat. Methods* **2017**, *14*, 414–416. [CrossRef]
69. Li, X.; Wang, K.; Lyu, Y.; Pan, H.; Zhang, J.; Stambolian, D.; Susztak, K.; Reilly, M.P.; Hu, G.; Li, M. Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis. *Nat. Commun.* **2020**, *11*, 2338. [CrossRef] [PubMed]
70. Yang, F.; Wang, W.; Wang, F.; Fang, Y.; Tang, D.; Huang, J.; Lu, H.; Yao, J. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nat. Mach. Intell.* **2022**, *4*, 852–866. [CrossRef]
71. Zeng, Y.; Wei, Z.; Zhong, F.; Pan, Z.; Lu, Y.; Yang, Y. A parameter-free deep embedded clustering method for single-cell RNA-seq data. *Brief. Bioinform.* **2022**, *23*, bbac172. [CrossRef] [PubMed]
72. Hu, H.; Li, Z.; Li, X.; Yu, M.; Pan, X. ScCAEs: Deep clustering of single-cell RNA-seq via convolutional autoencoder embedding and soft K-means. *Brief. Bioinform.* **2022**, *23*, bbab321. [CrossRef]
73. Wang, J.; Xia, J.; Wang, H.; Su, Y.; Zheng, C.H. scDCCA: Deep contrastive clustering for single-cell RNA-seq data based on auto-encoder network. *Brief. Bioinform.* **2023**, *24*, bbac625. [CrossRef] [PubMed]
74. Lei, Y.; Huang, X.T.; Guo, X.; Hang Katie Chan, K.; Gao, L. DeepGRNCS: Deep learning-based framework for jointly inferring gene regulatory networks across cell subpopulations. *Brief. Bioinform.* **2024**, *25*, bbae334. [CrossRef]
75. Hou, W.; Ji, Z. Assessing GPT-4 for cell type annotation in single-cell RNA-seq analysis. *Nat. Methods* **2024**, *21*, 1462–1465. [CrossRef]
76. Rosen, Y.; Roohani, Y.; Agarwal, A.; Samotorčan, L.; Consortium, T.S.; Quake, S.R.; Leskovec, J. Universal cell embeddings: A foundation model for cell biology. *BioRxiv* **2023**. [CrossRef]
77. Wen, H.; Tang, W.; Dai, X.; Ding, J.; Jin, W.; Xie, Y.; Tang, J. CellPLM: Pre-training of cell language model beyond single cells. *BioRxiv* **2023**. [CrossRef]
78. Lotfollahi, M.; Wolf, F.A.; Theis, F.J. scGen predicts single-cell perturbation responses. *Nat. Methods* **2019**, *16*, 715–721. [CrossRef]
79. Lotfollahi, M.; Klimovskaia Susmelj, A.; De Donno, C.; Hetzel, L.; Ji, Y.; Ibarra, I.L.; Srivatsan, S.R.; Naghipourfar, M.; Daza, R.M.; Martin, B. Predicting cellular responses to complex perturbations in high-throughput screens. *Mol. Syst. Biol.* **2023**, *19*, e11517. [CrossRef]
80. Roohani, Y.; Huang, K.; Leskovec, J. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nat. Biotechnol.* **2024**, *42*, 927–935. [CrossRef] [PubMed]
81. Cheng, Y.; Ma, X. scGAC: A graph attentional architecture for clustering single-cell RNA-seq data. *Bioinformatics* **2022**, *38*, 2187–2193. [CrossRef] [PubMed]
82. Hu, D.; Liang, K.; Zhou, S.; Tu, W.; Liu, M.; Liu, X. scDFC: A deep fusion clustering method for single-cell RNA-seq data. *Brief. Bioinform.* **2023**, *24*, bbad216. [CrossRef]
83. Zhang, T.; Ren, J.; Li, L.; Wu, Z.; Zhang, Z.; Dong, G.; Wang, G. scZAG: Integrating ZINB-Based Autoencoder with Adaptive Data Augmentation Graph Contrastive Learning for scRNA-seq Clustering. *Int. J. Mol. Sci.* **2024**, *25*, 5976. [CrossRef]
84. Gao, Q.; Ai, Q. DCRELM: Dual correlation reduction network-based extreme learning machine for single-cell RNA-seq data clustering. *Sci. Rep.* **2024**, *14*, 13541. [CrossRef] [PubMed]
85. Korsunsky, I.; Millard, N.; Fan, J.; Slowikowski, K.; Zhang, F.; Wei, K.; Baglaenko, Y.; Brenner, M.; Loh, P.R.; Raychaudhuri, S. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **2019**, *16*, 1289–1296. [CrossRef]
86. Stuart, T.; Butler, A.; Hoffman, P.; Hafemeister, C.; Papalexi, E.; Mauck, W.M., 3rd; Hao, Y.; Stoeckius, M.; Smibert, P.; Satija, R. Comprehensive Integration of Single-Cell Data. *Cell* **2019**, *177*, 1888–1902.e21. [CrossRef] [PubMed]
87. Welch, J.D.; Kozareva, V.; Ferreira, A.; Vanderburg, C.; Martin, C.; Macosko, E.Z. Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell* **2019**, *177*, 1873–1887.e17. [CrossRef]
88. Aran, D.; Looney, A.P.; Liu, L.; Wu, E.; Fong, V.; Hsu, A.; Chak, S.; Naikawadi, R.P.; Wolters, P.J.; Abate, A.R.; et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol.* **2019**, *20*, 163–172. [CrossRef]
89. Xu, C.; Lopez, R.; Mehlman, E.; Regier, J.; Jordan, M.I.; Yosef, N. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Mol. Syst. Biol.* **2021**, *17*, e9620. [CrossRef]

90. Gayoso, A.; Steier, Z.; Lopez, R.; Regier, J.; Nazor, K.L.; Streets, A.; Yosef, N. Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nat. Methods* **2021**, *18*, 272–282. [CrossRef] [PubMed]
91. Xiong, G.; LeRoy, N.J.; Bekiranov, S.; Sheffield, N.C.; Zhang, A. DeepGSEA: Explainable deep gene set enrichment analysis for single-cell transcriptomic data. *Bioinformatics* **2024**, *40*, btae434. [CrossRef]
92. Zhou, X.; Pan, J.; Chen, L.; Zhang, S.; Chen, Y. DeepIMAGER: Deeply Analyzing Gene Regulatory Networks from scRNA-seq Data. *Biomolecules* **2024**, *14*, 766. [CrossRef]
93. Kedzierska, K.Z.; Crawford, L.; Amini, A.P.; Lu, A.X. Zero-shot evaluation reveals limitations of single-cell foundation models. *Genome Biol.* **2025**, *26*, 101. [CrossRef]
94. Ahlmann-Eltze, C.; Huber, W.; Anders, S. Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. *Nat. Methods* **2025**, *22*, 1657–1661. [CrossRef] [PubMed]
95. Luecken, M.D.; Büttner, M.; Chaichoompu, K.; Danese, A.; Interlandi, M.; Mueller, M.F.; Strobl, D.C.; Zappia, L.; Dugas, M.; Colomé-Tatché, M.; et al. Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **2022**, *19*, 41–50. [CrossRef] [PubMed]
96. Duhan, L.; Kumari, D.; Naime, M.; Parmar, V.S.; Chhillar, A.K.; Dangi, M.; Pasrija, R. Single-cell transcriptomics: Background, technologies, applications, and challenges. *Mol. Biol. Rep.* **2024**, *51*, 600. [CrossRef]
97. Dong, K.; Zhang, S. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nat. Commun.* **2022**, *13*, 1739. [CrossRef]
98. Marx, V. Method of the Year 2020: Spatially resolved transcriptomics. *Nat. Methods* **2021**, *18*, 1. [CrossRef]
99. Lopez, R.; Nazaret, A.; Langevin, M.; Samaran, J.; Regier, J.; Jordan, M.I.; Yosef, N. A joint model of unpaired data from scRNA-seq and spatial transcriptomics for imputing missing gene expression measurements. *arXiv* **2019**, arXiv:1905.02269.
100. Abdelaal, T.; Mourragui, S.; Mahfouz, A.; Reinders, M.J.T. SpaGE: Spatial Gene Enhancement using scRNA-seq. *Nucleic Acids Res.* **2020**, *48*, e107. [CrossRef]
101. Shengquan, C.; Boheng, Z.; Xiaoyang, C.; Xuegong, Z.; Rui, J. stPlus: A reference-based method for the accurate enhancement of spatial transcriptomics. *Bioinformatics* **2021**, *37*, i299–i307. [CrossRef]
102. Biancalani, T.; Scalia, G.; Buffoni, L.; Avasthi, R.; Lu, Z.; Sanger, A.; Tokcan, N.; Vanderburg, C.R.; Segerstolpe, Å.; Zhang, M.; et al. Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. *Nat. Methods* **2021**, *18*, 1352–1362. [CrossRef] [PubMed]
103. Elosua-Bayes, M.; Nieto, P.; Mereu, E.; Gut, I.; Heyn, H. SPOTlight: Seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. *Nucleic Acids Res.* **2021**, *49*, e50. [CrossRef]
104. Song, Q.; Su, J. DSTG: Deconvoluting spatial transcriptomics data through graph-based artificial intelligence. *Brief. Bioinform.* **2021**, *22*, bbaa414. [CrossRef] [PubMed]
105. Dong, R.; Yuan, G.C. SpatialDWLS: Accurate deconvolution of spatial transcriptomic data. *Genome Biol.* **2021**, *22*, 145. [CrossRef]
106. Cable, D.M.; Murray, E.; Zou, L.S.; Goeva, A.; Macosko, E.Z.; Chen, F.; Irizarry, R.A. Robust decomposition of cell type mixtures in spatial transcriptomics. *Nat. Biotechnol.* **2022**, *40*, 517–526. [CrossRef]
107. Xu, C.; Jin, X.; Wei, S.; Wang, P.; Luo, M.; Xu, Z.; Yang, W.; Cai, Y.; Xiao, L.; Lin, X.; et al. DeepST: Identifying spatial domains in spatial transcriptomics by deep learning. *Nucleic Acids Res.* **2022**, *50*, e131. [CrossRef]
108. Zhang, B.; Li, M.; Kang, Q.; Deng, Z.; Qin, H.; Su, K.; Feng, X.; Chen, L.; Liu, H.; Fang, S.; et al. Generating single-cell gene expression profiles for high-resolution spatial transcriptomics based on cell boundary images. *GigaByte* **2024**, *2024*, gigabyte110. [CrossRef]
109. La Manno, G.; Soldatov, R.; Zeisel, A.; Braun, E.; Hochgerner, H.; Petukhov, V.; Lidschreiber, K.; Kastriti, M.E.; Lönnerberg, P.; Furlan, A.; et al. RNA velocity of single cells. *Nature* **2018**, *560*, 494–498. [CrossRef]
110. Lange, M.; Bergen, V.; Klein, M.; Setty, M.; Reuter, B.; Bakhti, M.; Lickert, H.; Ansari, M.; Schniering, J.; Schiller, H.B.; et al. CellRank for directed single-cell fate mapping. *Nat. Methods* **2022**, *19*, 159–170. [CrossRef] [PubMed]
111. Marx, V. Method of the Year: Spatially resolved transcriptomics. *Nat. Methods* **2021**, *18*, 9–14. [CrossRef]
112. Eisenstein, M. Seven technologies to watch in 2022. *Nature* **2022**, *601*, 658–661. [CrossRef]
113. Butler, D.; Mozsary, C.; Meydan, C.; Foox, J.; Rosiene, J.; Shaiber, A.; Danko, D.; Afshinnekoo, E.; MacKay, M.; Sedlazeck, F.J.; et al. Shotgun transcriptome, spatial omics, and isothermal profiling of SARS-CoV-2 infection reveals unique host responses, viral diversification, and drug interactions. *Nat. Commun.* **2021**, *12*, 1660. [CrossRef]
114. Kulasinghe, A.; Tan, C.W.; Ribeiro Dos Santos Miggiolaro, A.F.; Monkman, J.; SadeghiRad, H.; Bhuva, D.D.; Motta Junior, J.D.S.; Busatta Vaz de Paula, C.; Nagashima, S.; Baena, C.P.; et al. Profiling of lung SARS-CoV-2 and influenza virus infection dissects virus-specific host responses and gene signatures. *Eur. Respir. J.* **2022**, *59*, 2101881. [CrossRef]
115. Chen, A.; Liao, S.; Cheng, M.; Ma, K.; Wu, L.; Lai, Y.; Qiu, X.; Yang, J.; Xu, J.; Hao, S.; et al. Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. *Cell* **2022**, *185*, 1777–1792.e21. [CrossRef]
116. Moses, L.; Pachter, L. Museum of spatial transcriptomics. *Nat. Methods* **2022**, *19*, 534–546. [CrossRef]

117. Wu, K.E.; Yost, K.E.; Chang, H.Y.; Zou, J. BABEL enables cross-modality translation between multiomic profiles at single-cell resolution. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2023070118. [CrossRef]
118. Zeng, Y.; Wei, Z.; Yu, W.; Yin, R.; Yuan, Y.; Li, B.; Tang, Z.; Lu, Y.; Yang, Y. Spatial transcriptomics prediction from histology jointly through Transformer and graph neural networks. *Brief. Bioinform.* **2022**, *23*, bbac297. [CrossRef]
119. Zeng, Y.; Yin, R.; Luo, M.; Chen, J.; Pan, Z.; Lu, Y.; Yu, W.; Yang, Y. Identifying spatial domain by adapting transcriptomics with histology through contrastive learning. *Brief. Bioinform.* **2023**, *24*, bbad048. [CrossRef]
120. Kedzierska, K.Z.; Crawford, L.; Amini, A.P.; Lu, A.X. Assessing the limits of zero-shot foundation models in single-cell biology. *BioRxiv* **2023**. [CrossRef]