LLM4Cell: A Survey of Large Language and Agentic Models for Single-Cell Biology

Anonymous ACL submission

Abstract

Large language models (LLMs) and emerging agentic frameworks are beginning to influence single-cell biology by enabling natural-language interfaces, generative annotation, and multimodal data integration. However, progress remains fragmented across data modalities, model families, and evaluation practices. LLM4Cell presents a unified survey of 58 foundation and agentic models developed for single-cell research, spanning RNA, ATAC, multi-omic, and spatial modalities. We organize these methods into five families foundation, text-bridge, spatial/multimodal, epigenomic, and agentic and map them to eight key analytical tasks, including annotation, trajectory inference, perturbation modeling, and drug-response prediction. Drawing on over 40 public datasets, we analyze benchmark coverage, data diversity, and ethical or scalability constraints, and synthesize reported capabilities across ten domain-level dimensions related to biological grounding, multimodal alignment, fairness, privacy, and interpretability. By explicitly linking datasets, modeling paradigms, and evaluation domains, LLM4Cell provides an integrated perspective on language-driven single-cell analysis and highlights open challenges in standardization, interpretability, and trustworthy model development.

1 Introduction

Large language models (LLMs) are transforming biomedical discovery by linking molecular patterns with knowledge encoded in text (Zhang et al., 2025a; Aljohani et al., 2025; Bi et al., 2024). Prior work on computational workflows and biological prediction tasks demonstrates how learned representations capture complex molecular relationships and enhance practical performance (Lam et al., 2024; Bhattacharya et al., 2024; Dip et al., 2024; Nam et al., 2024; Shuvo et al., 2024). In single-cell biology, where each experiment measures thousands of genes across millions of cells, LLMs promise to unify gene expression, chromatin accessibility, and spatial organization under a shared, interpretable framework. Early foundation models such as scGPT (Cui et al., 2024), Geneformer (Theodoris et al., 2023), and scFoundation (Hao et al., 2024) learn transferable representations directly from single-cell data, while emerging systems like CellLM (Zhao et al., 2023), scAgent (Mao et al., 2025), and CellVerse (Zhang et al., 2025b) extend these capabilities toward reasoning, dialogue, and autonomous analysis.

Despite recent progress, the single-cell LLM landscape remains fragmented. Models vary widely in data modality, supervision, and evaluation standards from scRNA-seq to ATAC, multi-omic, and spatial data making cross-model comparison and reproducibility difficult. Benchmarking is inconsistent, datasets are unevenly distributed across modalities, and agentic frameworks lack standardized ways to evaluate biological grounding or reasoning correctness. As a result, it remains unclear what current single-cell LLMs truly understand, and where their limitations lie.

To address this, we introduce LLM4Cell, a unified survey of large language and agentic models for single-cell biology. Beyond cataloging models, LLM4Cell reveals how model design, dataset availability, and evaluation practices jointly shape biological understanding and reasoning claims an aspect largely missing from prior surveys. We analyze 58 representative methods across five methodological families Foundation, Text-Bridge, Spatial/Multimodal, Epigenomic, and Agentic and organize them across eight core tasks, including annotation, trajectory inference, perturbation modeling, and drug-response prediction. We further curate over 40 public datasets spanning RNA, ATAC,

multi-omic, spatial, perturbation, and plant domains, and assess models using a ten-dimension rubric that highlights gaps in grounding, generalization, interpretability, and scalability. Prior surveys (Zhang et al., 2025a; Lan et al., 2025; Bian et al., 2024b) primarily organize the literature around model architectures or prompting strategies ; in contrast, LLM4Cell is structured around datasets, tasks, and evaluation practice, explicitly analyzing how data availability and benchmarking constraints shape reported model capabilities.

Contributions. LLM4Cell offers (1) a modality-balanced registry of ∼40 benchmark datasets, (2) a unified taxonomy of 58 foundation and agentic models spanning eight analytical tasks, (3) a ten-dimension rubric assessing grounding, generalization, and ethical considerations, and (4) a focused discussion of open challenges in cross-modal alignment, reasoning, and trustworthy single-cell AI.

Beyond coverage, LLM4Cell reveals structural patterns that prior surveys leave implicit. We show that claims of reasoning and autonomy often outpace evaluation, particularly for agentic systems lacking standardized benchmarks. Across tasks, reported gains correlate more strongly with data scale and pairing than with architectural novelty. While language grounding improves interpretability, it does not reliably yield biological causality without explicit constraints, and agentic methods remain most effective for annotation, with limited generalization beyond this setting.

2 Related Work

Recent work has explored the use of large language models (LLMs) for single-cell biology, including surveys of foundation models and benchmarks for annotation and question answering. The ACL 2025 survey (Zhang et al., 2025a) and related reviews on large cellular models (LCMs) (Lan et al., 2025; Bian et al., 2024b) primarily organize the literature around model architectures, scaling, and prompting strategies, while benchmark efforts such as Single-Cell Omics Arena and CellVerse (Liu et al., 2024; Zhang et al., 2025b) focus on task-level performance. A side-by-side comparison with transformer-centric single-cell reviews and broader multimodal oncology or multi-omics surveys is provided in Appendix Table 1.

In contrast, LLM4Cell adopts a data-centric perspective that explicitly links models to the datasets, tasks, and evaluation practices underlying their claims. We further treat agentic and tool-augmented systems as a first-class methodological category and analyze trust-related dimensions such as grounding, fairness, and privacy, highlighting structural gaps that are not apparent in architecture-centric surveys.

3 Datasets

Progress in large language models for single-cell biology relies on the rapid growth of high-quality datasets across transcriptomic, epigenomic, multimodal, and spatial domains. Our survey compiles over forty public resources spanning five major modalities RNA, ATAC, multi-omic, spatial, and perturbation/drug-response plus emerging plant single-cell atlases. These datasets underpin model pretraining, evaluation, and cross-modal reasoning for annotation, integration, trajectory inference, perturbation modeling, and spatial mapping.

Transcriptomic atlases such as Tabula Sapiens, Tabula Muris (Consortium* et al., 2022), and the Human Cell Atlas (Travaglini et al., 2020) remain dominant for foundation-model training and ontology-aware annotation. Chromatin-accessibility data (e.g. Cusanovich mouse atlas (Cusanovich et al., 2018), human adult/fetal scATAC (Domcke et al., 2020)) map regulatory states but remain sparse and heterogeneous. Multi-omic resources (e.g. TEA-seq (Swanson et al., 2021), DOGMA-seq (Mimitou et al., 2021), CITE-seq (Stoeckius et al., 2017)) link RNA, ATAC, and protein modalities, providing supervision for cross-view alignment. Spatial technologies (e.g. Visium (Oliveira et al., 2025), Slide-seqV2 (Stickels et al., 2021), MERFISH (Chen et al., 2015), Stereo-seq (Chen et al., 2022)) connect molecular profiles to tissue architecture, while functional datasets (e.g. Perturb-seq (Replogle et al., 2022)) benchmark causal reasoning. Plant resources (e.g. scPlantDB (He et al., 2024), Arabidopsis E-CURD-4 (Shulse et al., 2019)) extend modeling beyond animal systems, opening cross-kingdom evaluation.

Despite this diversity, major gaps persist. (i) RNA atlases vastly outscale other modalities. (ii) Benchmark fragmentation and inconsistent metadata hinder reproducibility. (iii) Privacy constraints limit access to clinical spatial datasets. (iv) Non-human and plant data remain scarce. (v) Paired and tri-modal resources provide ideal testbeds for next-generation multimodal and agentic LLMs. Table 2, 3, 4, 5, 6, 7 (Appendix G) list representative datasets with tasks, scale, and source links for re-

producible benchmarking.

4 Model Taxonomy

Large language models are rapidly reshaping single-cell biology, producing diverse architectures, pretraining schemes, and reasoning frameworks. We categorize 58 representative methods into five methodological families Foundation Models, Text-Bridge LLMs, Spatial and Multimodal Models, Epigenomic Models, and Agentic Frameworks based on their core design and data modality. These span the progression from gene-level embeddings to multimodal and autonomous systems capable of biological reasoning.

Our taxonomy is organized along five orthogonal dimensions: (i) Modality (RNA, ATAC, multiomic, spatial, or hybrid), (ii) Grounding type (atlas, ontology, or marker-based), (iii) Agentic capability (multi-step reasoning or autonomous orchestration), (iv) Primary task (annotation, trajectory, perturbation, integration, etc.), and (v) Domain quality, represented by ten quantitative scores for grounding, fairness, scalability, and interpretability. Figure 1 summarizes this taxonomy, with detailed attributes in Appendix Table 8. Foundation models currently dominate in scope and adoption, while emerging spatial, epigenomic, and agentic frameworks signal a shift toward contextual, reasoning-driven single-cell intelligence.

4.1 Foundation Models

Foundation models learn transferable cell and gene embeddings directly from large-scale scRNA-seq without explicit labels. Representative systems such as scGPT (Cui et al., 2024), Geneformer (Theodoris et al., 2023), and scFoundation (Hao et al., 2024) pretrain on multi-tissue atlases exceeding one million cells, using masked-gene or rank-based reconstruction to capture expression context. Variants like tGPT (Shen et al., 2023), scBERT (Yang et al., 2022), and scGraphformer (Fan et al., 2024) treat genes as tokens or graph nodes, while cross-species models (UCE (Rosen et al., 2023), GeneCompass (Yang et al., 2024b)) apply contrastive alignment across homologous genes. These models underpin most single-cell LLM pipelines, offering strong transfer for annotation and integration but limited ontology grounding and explainability, motivating later text-bridged and reasoning frameworks.

4.2 Text-Bridge LLMs

Text-bridge models couple molecular embeddings with biomedical language to ground single-cell representations in semantics and ontology. In this survey, text-bridge models are defined as methods that *explicitly align* molecular representations with natural language or ontologies during training or inference; by contrast, foundation models (e.g., *scGPT*, *Geneformer*) operate purely on molecular tokens without language alignment, and models that use text only as weak supervision or metadata are classified as multimodal. Representative text-bridge systems include *scELMo* (Liu et al., 2023), *CellLM* (Zhao et al., 2023), and *GenePT* (Chen and Zou, 2024), which align gene or cell embeddings with textual descriptors, as well as *Cell2Sentence* (Levine et al., 2024) and *Cell2Text* (Kharouiche et al., 2025), which translate expression profiles into natural-language summaries. Architecturally, these models use dual-encoder or encoder–decoder designs with contrastive or prompt-based alignment between molecular encoders and domain-tuned language models (e.g., BioBERT). Text-bridge approaches enhance interpretability and enable zero-shot annotation but rely on curated vocabularies and remain non-agentic, positioning them between foundation pretraining and downstream reasoning frameworks. Boundary cases and hybrid designs are discussed in the appendix.

4.3 Spatial and Multimodal Models

Spatial and multimodal frameworks integrate gene expression with spatial coordinates, histology, or additional omics to capture tissue architecture. While some multimodal models may include textual metadata, they are grouped here when language is not used as a first-class semantic alignment target but rather as auxiliary input alongside spatial or molecular features. Representative models include TransformerST (Lu et al., 2024), spaLLM (Ji et al., 2024), and OmiCLIP (Cui et al., 2025) for spatial mapping, and scMMGPT (Shi et al., 2025) and FmH2ST for RNA–ATAC–protein integration. They use multi-branch Transformers with modality-specific encoders and cross-attention fusion, trained on datasets such as Visium DLPFC and MERFISH. These models achieve strong spatial alignment and biological realism but face heterogeneous resolutions, limited open benchmarks, and high computational cost.

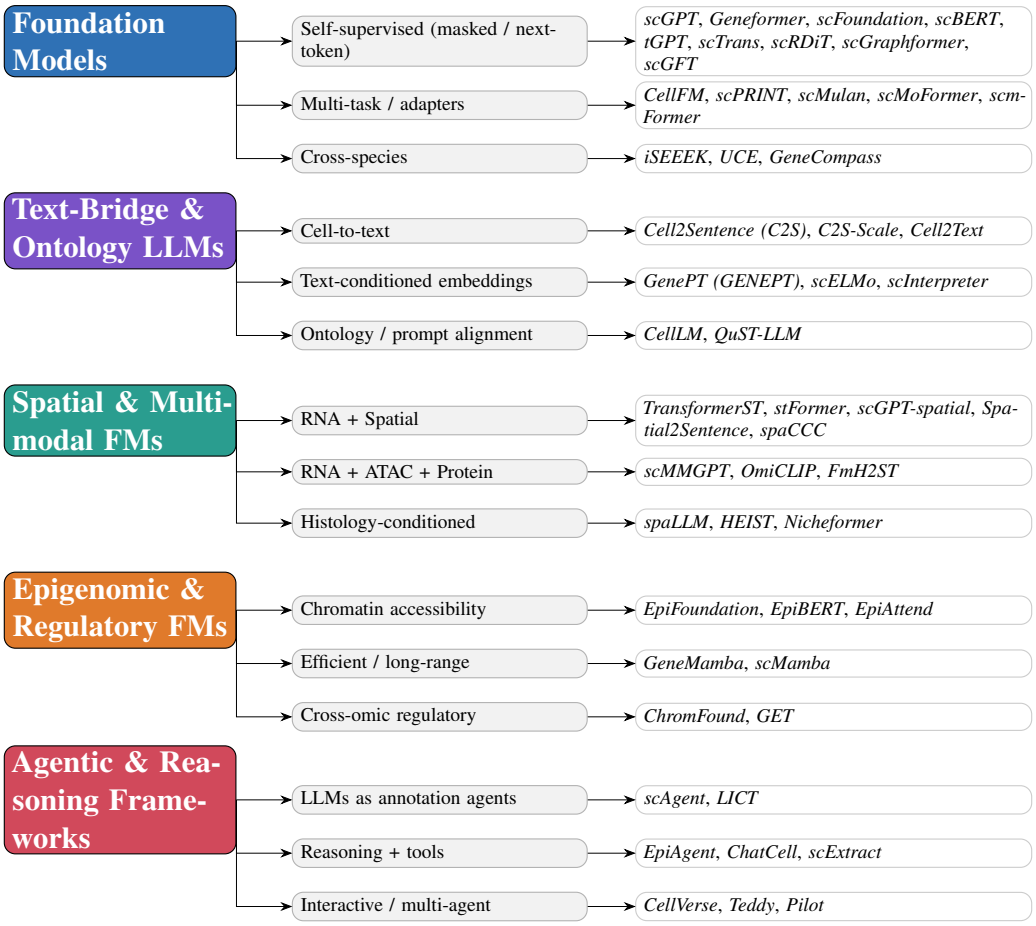

Figure 1: Hierarchical taxonomy for *LLM4Cell*. Color-coded families expand into sub-branches and representative models.

4.4 Epigenomic Models

Epigenomic foundation models extend LLM concepts to chromatin-accessibility and regulatory data such as scATAC-seq. EpiFoundation (Wu et al., 2025), EpiBERT (Javed et al., 2025), and EpiAttend (Li et al., 2022) learn cis-regulatory patterns from ENCODE and other compendia, while GeneMamba and scMamba use efficient state-space layers for long-range dependency modeling. Cross-omic variants (ChromFound (Jiao et al., 2025), GET (Fu et al., 2025)) jointly embed RNA and ATAC to infer gene-regulatory networks. These models improve biological grounding but remain constrained by sparse data and lack unified benchmarks across regulatory modalities.

4.5 Agentic Frameworks

Agentic systems integrate pretrained models with reasoning modules for autonomous single-cell analysis. Frameworks such as scAgent (Mao et al., 2025), CellVerse (Zhang et al., 2025b), and Teddy (Chevalier et al., 2025) combine domain-specific encoders with LLM controllers that plan tasks, query ontologies, and interface with tools or APIs. EpiAgent (Chen et al., 2025b) extends this paradigm to regulatory genomics. These systems enable dialogue-based annotation and multi-step reasoning but lack standardized benchmarks for reasoning fidelity and rely on underlying LLM reliability.

Failure modes and limitations. In practice, reported agentic gains are most consistent for annotation-centric workflows, while multi-step planning and reasoning remain fragile. Common limitations include sensitivity to prompt phrasing, brittle tool sequencing, and ontology drift that can lead to over-specific or unsupported cell-type predictions. Most frameworks do not explicitly model uncertainty or compare against strong non-agentic baselines, making it difficult to isolate when agentic control provides benefits beyond prompting or rule-based pipelines. These observations underscore the need for principled evaluation of reasoning fidelity rather than task-level accuracy alone.

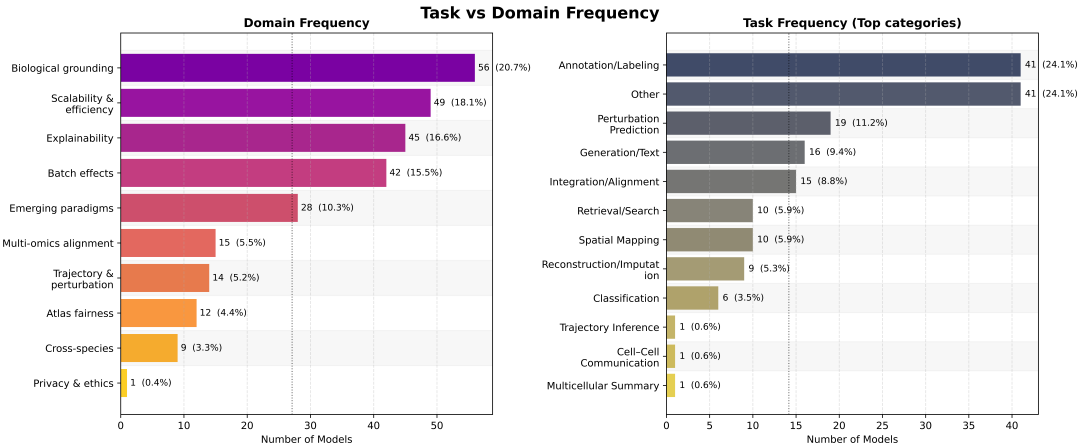

Figure 2: Domain and task coverage of surveyed LLM-based single-cell models. Left: Number of models reporting partial or full support for each evaluation domain (e.g., biological grounding, scalability, explainability). Right: Frequency of primary analytical tasks addressed. Counts are computed over all 58 surveyed models; a model may contribute to multiple domains or tasks if explicitly reported. Percentages are relative to the total number of models and indicate coverage frequency rather than exclusive categories or comparative performance.

5 Task-specific Applications

While the previous section categorized models by architectural family, here we analyze them through the lens of biological objectives shown in Figure 2 and Appendix Figure 3. Across the 58 methods surveyed, we identify eight recurring tasks that define single-cell modeling pipelines: (1) annotation and ontology mapping, (2) trajectory and perturbation modeling, (3) multi-omic integration, (4) spatial mapping and deconvolution, (5) regulatory-network and pathway inference, (6) cross-species translation, (7) generative simulation, and (8) drug-response prediction. Most models span multiple tasks-for example, scGPT and Geneformer support both annotation and perturbation prediction-reflecting the convergence between representation learning and functional reasoning.

This task-centric view complements the architectural taxonomy by revealing how foundation and agentic systems differ in operational scope. Foundation models dominate annotation, integration, and trajectory inference, whereas emerging text-bridge and agentic systems extend to knowledge-grounded reasoning and dynamic planning. Spatial and epigenomic frameworks contribute specialized capabilities in tissue mapping and chromatin-level interpretation.

5.1 Annotation and Ontology Mapping

Annotation is the most common single-cell task, spanning automated cell-type labeling, ontology alignment, and cross-dataset harmonization. Foundation models such as scGPT, scFoundation, and scBERT achieve high accuracy on atlases like Tabula Sapiens and the Human Cell Atlas. Text-bridge models (CellLM (Zhao et al., 2023), scELMo (Liu et al., 2023), GenePT (Chen and Zou, 2024)) add ontology-based interpretability, and agentic systems (scAgent) perform multi-step annotation reasoning via LLM controllers. Models are fine-tuned on curated references (e.g., Azimuth PBMC, HCA Lung) using masked-gene or contrastive objectives. Evaluation uses label accuracy or ARI; key limitations include rare-cell detection, cross-species consistency, and lack of reasoning benchmarks.

5.2 Trajectory and Perturbation Modeling

Trajectory modeling captures dynamic state transitions and causal responses to interventions. Geneformer (Theodoris et al., 2023), scGPT (Cui et al., 2024), and scRDiT (Dong et al., 2025) model temporal or perturbation trajectories using denoising and contrastive objectives, trained on datasets such as Replogle 2022 Perturb-seq and sci-Plex. Epigenomic models (EpiFoundation (Wu et al., 2025) ,EpiAgent (Chen et al., 2025b)) extend to regulatory responses through chromatin context and ontology reasoning. While these models predict single perturbations effectively, performance degrades for combinatorial or long-range effects, underscoring the need for unified causal benchmarks.

5.3 Multi-omic Integration

Multi-omic integration aims to learn unified representations across RNA, ATAC, and protein modalities to capture gene-regulatory and signaling re-

lationships. Representative models include scMMGPT (Shi et al., 2025), GET (Fu et al., 2025), ChromFound (Jiao et al., 2025), and epigenomic-aware variants such as EpiFoundation (Wu et al., 2025) and EpiBERT (Javed et al., 2025). These models are trained on paired datasets including 10x Multiome PBMC, TEA-seq, DOGMA-seq, and ASAP-seq, using cross-modal transformers or contrastive alignment. Most architectures use modality-specific encoders with a shared latent space. scMMGPT applies masked reconstruction across modalities, while GET uses cross-attention between gene and chromatin tokens. Compared to unimodal models, multi-omic methods improve cell-type alignment and batch correction, but remain limited by data imbalance and sparse modality overlap. Evaluation typically relies on modality-matching accuracy and latent-space correlation. Interpretability, standardized benchmarks, and scaling to tri-modal data is still challenging.

5.4 Spatial Mapping and Deconvolution

Spatial mapping links molecular profiles to tissue locations, enabling inference of cellular organization and microenvironmental context. Representative models include TransformerST (Lu et al., 2024), spaLLM, OmiCLIP, FmH2ST, HEIST (Madhu et al., 2025), and Spatial2Sentence (Chen et al., 2025a), trained on datasets such as Visium DLPFC, Slide-seqV2, MERFISH, and Xenium. These approaches integrate spatial coordinates, histology features, and gene-expression embeddings using cross-attention or contrastive alignment to reconstruct cell- or spot-level maps. TransformerST applies axial attention for spot-to-cell deconvolution, while spaLLM and OmiCLIP align histology and transcriptomics via language-guided contrastive learning.

Recent benchmarks including HEST-1k (Jaume et al., 2024), HESCAPE (Gindra et al., 2025), and STimage-1K4M (Chen et al., 2024) enable paired histology–transcriptomics and perturbation-aware evaluation of spatial alignment and robustness, but remain underutilized by current LLM-based models. Evaluation typically relies on spatial correlation, clustering accuracy, or cell-type F1, and heterogeneous resolutions continue to limit systematic comparison across methods.

5.5 Regulatory Network & Pathway Inference

Regulatory and pathway inference aims to uncover gene–gene and enhancer–promoter dependencies underlying transcriptional programs. Representative models include GeneMamba (Qi et al., 2025a), scMamba (Yuan et al., 2025), EpiFoundation (Wu et al., 2025), EpiBERT (Javed et al., 2025), ChromFound (Jiao et al., 2025), and GET (Fu et al., 2025), trained on large-scale ATAC and multiome atlases such as ENCODE and 10x Multiome PBMC. These approaches model genes or chromatin regions as tokens and learn regulatory structure via self- or cross-attention. GeneMamba and scMamba use state-space layers to capture long-range dependencies, while EpiBERT and EpiFoundation apply masked-region reconstruction to model enhancer–promoter coupling. ChromFound and GET integrate RNA and accessibility signals to infer transcription-factor activity and regulatory directionality. Evaluation typically relies on motif enrichment, AUROC for known interactions, or pathway overlap with KEGG or Reactome. Despite improved interpretability, sparse training data and the absence of unified regulatory benchmarks remain key limitations.

5.6 Cross-Species Translation

Cross-species translation transfers cell and gene representations across organisms to support comparative and evolutionary single-cell analysis. Representative models include iSEEEK (Shen et al., 2022), UCE (Rosen et al., 2023), GeneCompass (Yang et al., 2024b), and scPlantLLM (Cao et al., 2025), trained on homolog-mapped atlases such as the Mouse Cell Atlas, Tabula Muris, and scPlantDB. These models align orthologous genes or conserved cell states to enable zero-shot annotation across species. iSEEEK integrates over 11 million human and mouse cells using shared gene-token vocabularies, UCE employs masked autoencoding with contrastive alignment, and GeneCompass incorporates ontology-informed pretraining; scPlantLLM extends this paradigm to 17 plant species. Evaluation typically measures transfer accuracy and cross-species embedding consistency. Limitations stem from incomplete ortholog mappings and domain shifts in sequencing depth and tissue composition.

5.7 Generation and Simulation

Generative models synthesize realistic single-cell profiles, simulate perturbations, and reconstruct missing modalities. Representative methods include scGPT (Qi et al., 2025b), scFoundation (Hao et al., 2024), CellFM (Zeng et al., 2025), and Geneformer (Theodoris et al., 2023), which use

decoder- or diffusion-style architectures to model gene-expression distributions. These models are pretrained on large atlases such as Tabula Sapiens and the Human Cell Atlas, and fine-tuned to generate novel cells or perturbation outcomes. scGPT and scFoundation extend masked-token prediction with generative decoding, while CellFM incorporates multimodal conditioning for cross-tissue synthesis. Evaluation commonly relies on distributional similarity metrics or recovery of cell-type proportions. While enabling scalable *in silico* experimentation, these approaches face challenges in maintaining biological realism and avoiding bias toward dominant cell types.

5.8 Drug-Response Prediction

Drug-response models infer transcriptional changes following chemical or genetic perturbations to support virtual screening and mechanism discovery. Representative approaches include Geneformer (Theodoris et al., 2023), scGPT (Cui et al., 2024), EpiFoundation (Wu et al., 2025), and EpiAgent (Chen et al., 2025b), evaluated on datasets such as sci-Plex, Replogle 2022 Perturb-seq, and the ARC Virtual Cell Challenge. These methods learn conditional embeddings that map control to perturbed states, enabling zero-shot prediction of unseen compounds or targets. Geneformer applies masked-token denoising, scGPT and EpiFoundation incorporate modality-specific conditioning, and EpiAgent adds ontology-guided reasoning linking drugs to gene pathways. Evaluation typically uses Pearson correlation or top-k target recovery. While effective for single-agent perturbations, performance degrades for combinatorial treatments and unseen pathways, motivating richer multimodal benchmarks.

Cross-task observations. Across tasks, foundation models dominate annotation, integration, and generation due to large-scale atlas pretraining, while text-bridge methods improve interpretability through language and ontology grounding. Spatial and epigenomic models add tissue and regulatory context, and agentic frameworks introduce multi-step reasoning and tool use. However, evaluation remains uneven: annotation and integration are well benchmarked, whereas trajectory, drug-response, and agentic reasoning lack standardized protocols, and cross-species and privacy-aware tasks remain underrepresented. Overall, the field is maturing toward general representations, with agentic systems beginning to operationalize biological reasoning.

6 Domain Analysis

We evaluate fifty-eight methods across ten domain dimensions capturing reliability, generalization, and interpretability. Each score reflects published evidence of performance or explicit design alignment, summarized in Appendix Table 8 and Figure 2. Together, these dimensions provide a composite view of model maturity beyond task accuracy.

Biological Grounding Most foundation and multimodal models achieve strong biological grounding by encoding marker genes, pathways, or regulatory regions. scGPT, Geneformer, and EpiFoundation leverage atlas-scale training to recover known cell types and pathway enrichments. Spatial frameworks such as TransformerST and OmiCLIP add tissue-context grounding via histology alignment. However, textual and ontology grounding remain inconsistent, and biological semantics are rarely explicit in model objectives.

Batch Effects and Heterogeneity Robustness to batch effects varies widely. scFoundation (Hao et al., 2024), CellFM (Zeng et al., 2025), and iSEEEK (Shen et al., 2022) mitigate donor and protocol variation through large-scale multi-domain pretraining, while others rely on explicit batch tokens or adversarial alignment. True cross-platform generalization remains limited; even large models show degradation when tested on unseen sequencing chemistries or tissue types.

Multi-omics Alignment Models such as scMMGPT (Shi et al., 2025), GET (Fu et al., 2025), and ChromFound (Jiao et al., 2025) demonstrate consistent alignment across RNA, ATAC, and protein features using cross-attention or contrastive fusion. Spatial hybrids (HEIST, FmH2ST) further connect molecular and visual domains. While alignment quality improves, limited paired datasets and uneven modality coverage constrain reproducibility and benchmarking.

Trajectory and Perturbation Dynamic inference is strongest in generative transformers (Geneformer, scGPT) and epigenomic variants (EpiAgent). They predict post-perturbation expression or accessibility shifts with reasonable fidelity but struggle to model multi-target or time-continuous responses. Few methods quantify causal uncertainty or biological plausibility.

Cross-Species and Cross-Tissue Generalization Cross-organism transfer remains challenging. iSEEEK (Shen et al., 2022), UCE (Rosen

et al., 2023), and GeneCompass (Yang et al., 2024b) align human and mouse cell embeddings via ortholog mapping, while scPlantLLM (Cao et al., 2025) extends this to plants. Performance drops sharply for divergent taxa, reflecting bias toward human datasets.

Atlas Fairness and Representation Balance Dataset imbalance affects nearly all models. Human and immune-cell–dominant atlases overrepresent certain tissues and demographics. Teddy (Chevalier et al., 2025) introduces benchmarking for representation balance, but fairness metrics are rarely adopted. No model yet enforces demographic or tissue-specific parity during training.

Explainability Text-bridge systems (GenePT, CellLM, Cell2Text (Kharouiche et al., 2025)) improve interpretability by linking embeddings to ontology terms. Agentic systems (Mao et al., 2025) extend this via natural-language reasoning and tool explanations. Most foundation models still rely on attention visualization, with limited causal transparency.

Privacy and Ethics Data privacy remains underexplored. Synthetic generation (scGFT (Nouri, 2025)) and open-access chat systems (ChatCell (Fang et al., 2024)) raise concerns over reidentification and data leakage. No single-cell LLM currently implements federated or privacy-preserving learning, and ethical guidance for multiomic sharing remains informal.

Scalability and Efficiency Scaling trends mirror NLP: efficient transformers (xTrimoGene (Gong et al., 2023)) and state-space models (GeneMamba (Qi et al., 2025a), scMamba (Yuan et al., 2025)) reduce memory cost while retaining accuracy. CellFM (Zeng et al., 2025) demonstrates training across 100 M cells, showing feasibility of atlas-scale learning. Still, high compute requirements limit accessibility for most research groups.

Emerging Paradigms and Agentic Behavior Recent frameworks (scAgent (Mao et al., 2025), CellVerse (Zhang et al., 2025b), EpiAgent (Chen et al., 2025b)) introduce reasoning and autonomous decision pipelines, integrating LLM controllers with specialized encoders. These systems achieve the highest scores in explainability and cross-modal planning but lack standardized evaluation of reasoning fidelity or reproducibility.

7 Open Problems

Despite rapid progress, several open challenges remain. Evaluation across modalities is inconsistent, with metrics often favoring reconstruction over biological plausibility and little independent replication, underscoring the need for standardized benchmarks and community-curated validation sets. Training data are heavily skewed toward human and mouse atlases, limiting cross-species and clinical generalization and reinforcing bias against rare-cell, plant, and microbial systems. True integration of RNA, ATAC, spatial, and temporal modalities remains unresolved, as most models handle only pairwise fusion and lack unified tokenization or multimodal pretraining pipelines. Interpretability is also limited: current embeddings capture correlations but rarely provide mechanistic or causal insight without explicit reasoning layers or experimental grounding. Ethical and privacy concerns persist, as few models support consent-aware governance, auditability, or privacy-preserving learning. Finally, agentic frameworks show promise but lack reliable benchmarks, with multi-step reasoning, tool use, and orchestration remaining fragile and difficult to evaluate.

8 Conclusions

LLM4Cell presents a unified survey of large language and agentic models for single-cell biology, linking datasets, model families, and evaluation practices within a common framework. By analyzing 58 models and over 40 public datasets across transcriptomic, epigenomic, spatial, and perturbation modalities, we show how foundation pretraining, multimodal grounding, and emerging agentic approaches are shaping current methods. Our taxonomy and evaluation rubric clarify strengths and gaps, and reveal that many reported gains are driven by data scale and task scope rather than robust reasoning, underscoring the need for grounded and standardized evaluation.

Limitations

Our analysis is limited by heterogeneous reporting across studies and a rubric that captures qualitative trends rather than standardized scores. Access restrictions prevent inclusion of some clinical and proprietary spatial datasets, and non-animal resources remain scarce. We do not evaluate computational efficiency or hyperparameter sensitivity. As the field is rapidly evolving, LLM4Cell represents a snapshot in time, motivating future community benchmarks and reasoning-focused evaluations.

References

Manar Aljohani, Jun Hou, Sindhura Kommu, and Xuan Wang. 2025. A comprehensive survey on the trustworthiness of large language models in healthcare. *arXiv preprint arXiv:2502.15871*.

Manojit Bhattacharya, Soumen Pal, Srijan Chatterjee, Sang-Soo Lee, and Chiranjib Chakraborty. 2024. Large language model to multimodal large language model: A journey to shape the biological macromolecules to biological sciences and medicine. *Molecular Therapy Nucleic Acids*, 35(3).

Zhenyu Bi, Sajib Acharjee Dip, Daniel Hajialigol, Sindhura Kommu, Hanwen Liu, Meng Lu, and Xuan Wang. 2024. Ai for biomedicine in the era of large language models. *arXiv preprint arXiv:2403.15673*.

Haiyang Bian, Yixin Chen, Xiaomin Dong, Chen Li, Minsheng Hao, Sijie Chen, Jinyi Hu, Maosong Sun, Lei Wei, and Xuegong Zhang. 2024a. scmulan: a multitask generative pre-trained language model for single-cell analysis. In *International Conference on Research in Computational Molecular Biology*, pages 479–482. Springer.

Haiyang Bian, Yixin Chen, Erpai Luo, Xinze Wu, Minsheng Hao, Lei Wei, and Xuegong Zhang. 2024b. General-purpose pre-trained large cellular models for single-cell transcriptomics. *National Science Review*, 11(11):nwae340.

Guangshuo Cao, Haoyu Chao, Wenqi Zheng, Yangming Lan, Kaiyan Lu, Yueyi Wang, Ming Chen, He Zhang, and Dijun Chen. 2025. scplantllm: A foundation model for exploring single-cell expression atlases in plants. *Genomics, Proteomics and Bioinformatics*, 23(3):qzaf024.

Shenghao Cao, Kaiyuan Yang, Jiabei Cheng, Jiachen Li, Hong-Bin Shen, Xiaoyong Pan, and Ye Yuan. 2024. stformer: a foundation model for spatial transcriptomics. *bioRxiv*, pages 2024–09.

Ao Chen, Sha Liao, Mengnan Cheng, Kailong Ma, Liang Wu, Yiwei Lai, Xiaojie Qiu, Jin Yang, Jiangshan Xu, Shijie Hao, and 1 others. 2022. Spatiotemporal transcriptomic atlas of mouse organogenesis using dna nanoball-patterned arrays. *Cell*, 185(10):1777–1792.

Chi-Jane Chen, Yuhang Chen, Sukwon Yun, Natalie Stanley, and Tianlong Chen. 2025a. Spatial coordinates as a cell language: A multi-sentence framework for imaging mass cytometry analysis. *arXiv preprint arXiv:2506.01918*.

Jiawen Chen, Muqing Zhou, Wenrong Wu, Jinwei Zhang, Yun Li, and Didong Li. 2024. Stimage-1k4m: A histopathology image-gene expression dataset for spatial transcriptomics. *Advances in Neural Information Processing Systems*, 37:35796–35823.

Kok Hao Chen, Alistair N Boettiger, Jeffrey R Moffitt, Siyuan Wang, and Xiaowei Zhuang. 2015. Spatially resolved, highly multiplexed rna profiling in single cells. *Science*, 348(6233):aaa6090.

Xiaoyang Chen, Keyi Li, Xuejian Cui, Zian Wang, Qun Jiang, Jiacheng Lin, Zhen Li, Zijing Gao, Hairong Lv, and Rui Jiang. 2025b. Epiagent: foundation model for single-cell epigenomics. *Nature Methods*, pages 1–12.

Yiqun Chen and James Zou. 2024. Genept: a simple but effective foundation model for genes and cells built from chatgpt. *bioRxiv*, pages 2023–10.

Alexis Chevalier, Soumya Ghosh, Urvi Awasthi, James Watkins, Julia Bieniewska, Nichita Mitrea, Olga Kotova, Kirill Shkura, Andrew Noble, Michael Steinbaugh, and 1 others. 2025. Teddy: A family of foundation models for understanding single cell biology. *arXiv preprint arXiv:2503.03485*.

The Tabula Sapiens Consortium*, Robert C Jones, Jim Karkanias, Mark A Krasnow, Angela Oliveira Pisco, Stephen R Quake, Julia Salzman, Nir Yosef, Bryan Bulthaup, Phillip Brown, and 1 others. 2022. The tabula sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science*, 376(6594):eabl4896.

Haotian Cui, Alejandro Tejada-Lapuerta, Maria Brbić, Julio Saez-Rodriguez, Simona Cristea, Hani Goodarzi, Mohammad Lotfollahi, Fabian J Theis, and Bo Wang. 2025. Towards multimodal foundation models in molecular cell biology. *Nature*, 640(8059):623–633.

Haotian Cui, Chloe Wang, Hassaan Maan, Nan Duan, and Bo Wang. 2022. scformer: a universal representation learning approach for single-cell data using transformers. *bioRxiv*, pages 2022–11.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. 2024. scgpt: toward building a foundation model for single-cell multiomics using generative ai. *Nature methods*, 21(8):1470–1480.

Darren A Cusanovich, Andrew J Hill, Delasa Aghamirzaie, Riza M Daza, Hannah A Pliner, Joel B Berletch, Galina N Filippova, Xingfan Huang, Lena Christiansen, William S DeWitt, and 1 others. 2018. A single-cell atlas of in vivo mammalian chromatin accessibility. *Cell*, 174(5):1309–1324.

Bharath Dandala, Michael M Danziger, Ella Barkan, Tanwi Biswas, Viatcheslav Gurev, Jianying Hu, Matthew Madgwick, Akira Koseki, Tal Kozlovski, Michal Rosen-Zvi, and 1 others. 2025. Bmfm-rna: An open framework for building and evaluating transcriptomic foundation models. *arXiv preprint arXiv:2506.14861*.

Sajib Acharjee Dip, Da Ma, and Liqing Zhang. 2024. Deepage: Harnessing deep neural network for epigenetic age estimation from dna methylation data of human blood samples. In *Proceedings of the AAAI Symposium Series*, volume 4, pages 267–274.

Silvia Domcke, Andrew J Hill, Riza M Daza, Junyue Cao, Diana R O'Day, Hannah A Pliner, Kimberly A Aldinger, Dmitry Pokholok, Fan Zhang, Jennifer H Milbank, and 1 others. 2020. A human cell atlas of fetal chromatin accessibility. *Science*, 370(6518):eaba7612.

Shengze Dong, Zhuorui Cui, Ding Liu, and Jinzhi Lei. 2025. scrdit: Generating single-cell rna-seq data by diffusion transformers and accelerating sampling. *Interdisciplinary Sciences: Computational Life Sciences*, pages 1–12.

Can Ergen, Valeh Valiollah Pour Amiri, Martin Kim, Ori Kronfeld, Aaron Streets, Adam Gayoso, and Nir Yosef. 2025. Scvi-hub: an actionable repository for model-driven single-cell analysis. *Nature Methods*, pages 1–10.

Xingyu Fan, Jiacheng Liu, Yaodong Yang, Chunbin Gu, Yuqiang Han, Bian Wu, Yirong Jiang, Guangyong Chen, and Pheng-Ann Heng. 2024. scgraphformer: unveiling cellular heterogeneity and interactions in scrna-seq data using a scalable graph transformer network. *Communications Biology*, 7(1):1463.

Yin Fang, Kangwei Liu, Ningyu Zhang, Xinle Deng, Penghui Yang, Zhuo Chen, Xiangru Tang, Mark Gerstein, Xiaohui Fan, and Huajun Chen. 2024. Chatcell: Facilitating single-cell analysis with natural language. *arXiv preprint arXiv:2402.08303*.

Xi Fu, Shentong Mo, Alejandro Buendia, Anouchka P Laurent, Anqi Shao, Maria del Mar Alvarez-Torres, Tianji Yu, Jimin Tan, Jiayu Su, Romella Sagatelian, and 1 others. 2025. A foundation model of transcription across human cell types. *Nature*, 637(8047):965–973.

Adam Gayoso, Romain Lopez, Galen Xing, Pierre Boyeau, Valeh Valiollah Pour Amiri, Justin Hong, Katherine Wu, Michael Jayasuriya, Edouard Mehlman, Maxime Langevin, and 1 others. 2022. A python library for probabilistic analysis of single-cell omics data. *Nature biotechnology*, 40(2):163–166.

Rushin H Gindra, Giovanni Palla, Mathias Nguyen, Sophia J Wagner, Manuel Tran, Fabian J Theis, Dieter Saur, Lorin Crawford, and Tingying Peng. 2025. A large-scale benchmark of cross-modal learning for histology and gene expression in spatial transcriptomics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1182–1192.

Jing Gong, Minsheng Hao, Xingyi Cheng, Xin Zeng, Chiming Liu, Jianzhu Ma, Xuegong Zhang, Taifeng Wang, and Le Song. 2023. xtrimogene: an efficient and scalable representation learner for single-cell rna-seq data. *Advances in Neural Information Processing Systems*, 36:69391–69403.

Zhen-Hao Guo, Yan Wu, Siguo Wang, Qinhu Zhang, Jin-Ming Shi, Yan-Bin Wang, and Zhan-Heng Chen. 2023. scinterpreter: a knowledge-regularized generative model for interpretably integrating scrna-seq data. *BMC bioinformatics*, 24(1):481.

Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, and Le Song. 2024. Large-scale foundation model on single-cell transcriptomics. *Nature methods*, 21(8):1481–1491.

Zhaohui He, Yuting Luo, Xinkai Zhou, Tao Zhu, Yangming Lan, and Dijun Chen. 2024. scplantdb: a comprehensive database for exploring cell types and markers of plant cell atlases. *Nucleic acids research*, 52(D1):D1629–D1638.

Chao Hui Huang. 2024. Qust-llm: Integrating large language models for comprehensive spatial transcriptomics analysis. *arXiv preprint arXiv:2406.14307*.

Guillaume Jaume, Paul Doucet, Andrew Song, Ming Yang Lu, Cristina Almagro Pérez, Sophia Wagner, Anurag Vaidya, Richard Chen, Drew Williamson, Ahrong Kim, and 1 others. 2024. Hest-1k: A dataset for spatial transcriptomics and histology image analysis. *Advances in Neural Information Processing Systems*, 37:53798–53833.

Nauman Javed, Thomas Weingarten, Arijit Sehanobish, Adam Roberts, Avinava Dubey, Krzysztof Choromanski, and Bradley E Bernstein. 2025. A multi-modal transformer for cell type-agnostic regulatory predictions. *Cell Genomics*, 5(2).

Boya Ji, Xiaoqi Wang, Debin Qiao, Liwen Xu, and Shaoliang Peng. 2024. Spaccc: Large language model-based cell-cell communication inference for spatially resolved transcriptomic data. *Big Data Mining and Analytics*, 7(4):1129–1147.

Yifeng Jiao, Yuchen Liu, Yu Zhang, Xin Guo, Yushuai Wu, Chen Jiang, Jiyang Li, Hongwei Zhang, Limei Han, Xin Gao, and 1 others. 2025. Chromfound: Towards a universal foundation model for single-cell chromatin accessibility data. *arXiv preprint arXiv:2505.12638*.

Mehdi Joodaki, Mina Shaigan, Victor Parra, Roman D Bülow, Christoph Kuppe, David L Hölscher, Mingbo Cheng, James S Nagai, Michaël Goedertier, Nassim Bouteldja, and 1 others. 2024. Detection of patient-level distances from single cell genomics and pathomics data with optimal transport (pilot). *Molecular systems biology*, 20(2):57–74.

Jérémie Kalfon, Jules Samaran, Gabriel Peyré, and Laura Cantini. 2025. scprint: pre-training on 50 million cells allows robust gene network predictions. *Nature Communications*, 16(1):3607.

Oussama Kharouiche, Aris Markogiannakis, Xiao Fei, Michail Chatzianastasis, and Michalis Vazirgiannis. 2025. Cell2text: Multimodal llm for generating single-cell descriptions from rna-seq data. *arXiv preprint arXiv:2509.24840*.

Zhenglun Kong, Mufan Qiu, John Boesen, Xiang Lin, Sukwon Yun, Tianlong Chen, Manolis Kellis, and Marinka Zitnik. 2025. Spatia: Multimodal model for prediction and generation of spatial cell phenotypes. *ArXiv*, pages arXiv–2507.

Hilbert Yuen In Lam, Xing Er Ong, and Marek Mutwil. 2024. Large language models in plant biology. *Trends in Plant Science*, 29(10):1145–1155.

Wei Lan, Zhentao Tang, Mingyang Liu, Qingfeng Chen, Wei Peng, Yiping Phoebe Chen, and Yi Pan. 2025. The large language models on biomedical data analysis: a survey. *IEEE Journal of Biomedical and Health Informatics*.

Daniel Levine, Syed A Rizvi, Sacha Lévy, Nazreen Pallikkavaliyaveetil, David Zhang, Xingyu Chen, Sina Ghadermarzi, Ruiming Wu, Zihe Zheng, Ivan Vrkic, Anna Zhong, Daphne Raskin, Insu Han, Antonio Henrique De Oliveira Fonseca, Josue Ortega Caro, Amin Karbasi, Rahul Madhav Dhodapkar, and David Van Dijk. 2024. Cell2Sentence: Teaching large language models the language of biology. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 27299–27325. PMLR.

Russell Li, Heng Xu, and Eran A Mukamel. 2022. Epiattend: A transformer model of gene regulation combining single cell epigenomes with dna sequence. In *NeurIPS 2022 Workshop on Learning Meaningful Representations of Life*.

Junhao Liu, Siwei Xu, Lei Zhang, and Jing Zhang. 2024. Single-cell omics arena: A benchmark study for large language models on cell type annotation using single-cell data. *arXiv preprint arXiv:2412.02915*.

Tianyu Liu, Tianqi Chen, Wangjie Zheng, Xiao Luo, Yiqun Chen, and Hongyu Zhao. 2023. scelmo: Embeddings from language models are good learners for single-cell data analysis. *bioRxiv*, pages 2023–12.

Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. 2018. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058.

L Lu, Wen Xue, Xindian Wei, and 1 others. 2024. Sctrans: multi-scale scrna-seq sub-vector completion transformer for gene-selective cell type annotation. In *33rd International Joint Conference on Artificial Intelligence (IJCAI 2024)*, pages 5954–62. International Joint Conferences on Artificial Intelligence.

Hiren Madhu, João Felipe Rocha, Tinglin Huang, Siddharth Viswanath, Smita Krishnaswamy, and Rex Ying. 2025. Heist: A graph foundation model for spatial transcriptomics and proteomics data. *arXiv preprint arXiv:2506.11152*.

Yuren Mao, Yu Mi, Peigen Liu, Mengfei Zhang, Hanqing Liu, and Yunjun Gao. 2025. scagent: Universal single-cell annotation via a llm agent. *arXiv preprint arXiv:2504.04698*.

Eleni P Mimitou, Caleb A Lareau, Kelvin Y Chen, Andre L Zorzetto-Fernandes, Yuhan Hao, Yusuke Takeshima, Wendy Luo, Tse-Shun Huang, Bertrand Z Yeung, Efthymia Papalexi, and 1 others. 2021. Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. *Nature biotechnology*, 39(10):1246–1258.

Daye Nam, Andrew Macvean, Vincent Hellendoorn, Bogdan Vasilescu, and Brad Myers. 2024. Using an llm to help with code understanding. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, pages 1–13.

Nima Nouri. 2025. Single-cell rna-seq data augmentation using generative fourier transformer. *Communications Biology*, 8(1):113.

Nima Nouri, Ronen Artzi, and Virginia Savova. 2025. An agentic ai framework for ingestion and standardization of single-cell rna-seq data analysis. *bioRxiv*, pages 2025–07.

Michelli Faria de Oliveira, Juan Pablo Romero, Meii Chung, Stephen R Williams, Andrew D Gottscho, Anushka Gupta, Susan E Pilipauskas, Seayar Mohabbat, Nandhini Raman, David J Sukovich, and 1 others. 2025. High-definition spatial transcriptomic profiling of immune cell populations in colorectal cancer. *Nature Genetics*, pages 1–12.

Steven Palayew, Bo Wang, and Gary Bader. 2025. Towards applying large language models to complement single-cell foundation models. *arXiv preprint arXiv:2507.10039*.

Cong Qi, Hanzhang Fang, Tianxing Hu, Siqi Jiang, and Wei Zhi. 2025a. Bidirectional mamba for single-cell data: Efficient context learning with biological fidelity. *arXiv preprint arXiv:2504.16956*.

Yunjing Qi, Yulong Kan, Jing Qi, and Shuilin Jin. 2025b. scgt: I ntegration algorithm for single-cell rna-seq and atac-seq based on graph transforme r. *Bioinformatics*, page btaf357.

Joseph M Replogle, Reuben A Saunders, Angela N Pogson, Jeffrey A Hussmann, Alexander Lenail, Alina Guna, Lauren Mascibroda, Eric J Wagner, Karen Adelman, Gila Lithwick-Yanai, and 1 others. 2022. Mapping information-rich genotype-phenotype landscapes with genome-scale perturb-seq. *Cell*, 185(14):2559–2575.

Syed Asad Rizvi, Daniel Levine, Aakash Patel, Shiyang Zhang, Eric Wang, Sizhuang He, David Zhang, Cerise Tang, Zhuoyang Lyu, Rayyan Darji, Chang Li, Emily Sun, David Jeong, Lawrence Zhao, Jennifer Kwan, David Braun, Brian Hafler, Jeffrey Ishizuka, Rahul M. Dhodapkar, and 4 others. 2025. Scaling large language models for next-generation single-cell analysis. *bioRxiv*.

Yanay Rosen, Yusuf Roohani, Ayush Agarwal, Leon Samotorčan, Tabula Sapiens Consortium, Stephen R Quake, and Jure Leskovec. 2023. Universal cell embeddings: A foundation model for cell biology. *bioRxiv*, pages 2023–11.

Anna C Schaar, Alejandro Tejada-Lapuerta, Giovanni Palla, Robert Gutgesell, Lennard Halle, Mariia Minaeva, Larsen Vornholz, Leander Dony, Francesca Drummer, Mojtaba Bahrami, and 1 others. 2024. Nicheformer: a foundation model for single-cell and spatial omics. *bioRxiv*, pages 2024–04.

Hongru Shen, Jilei Liu, Jiani Hu, Xilin Shen, Chao Zhang, Dan Wu, Mengyao Feng, Meng Yang, Yang Li, Yichen Yang, and 1 others. 2023. Generative pre-training from large-scale transcriptomes for single-cell deciphering. *Iscience*, 26(5).

Hongru Shen, Xilin Shen, Mengyao Feng, Dan Wu, Chao Zhang, Yichen Yang, Meng Yang, Jiani Hu, Jilei Liu, Wei Wang, and 1 others. 2022. A universal approach for integrating super large-scale single-cell transcriptomes by exploring gene rankings. *Briefings in Bioinformatics*, 23(2).

Yaorui Shi, Jiaqi Yang, Sihang Li, Junfeng Fang, Xiang Wang, Zhiyuan Liu, and Yang Zhang. 2025. Multi-modal language modeling for high-accuracy single cell transcriptomics analysis and generation. *arXiv e-prints*, pages arXiv–2503.

Christine N Shulse, Benjamin J Cole, Doina Ciobanu, Junyan Lin, Yuko Yoshinaga, Mona Gouran, Gina M Turco, Yiwen Zhu, Ronan C O'Malley, Siobhan M Brady, and 1 others. 2019. High-throughput single-cell transcriptome profiling of plant cell types. *Cell reports*, 27(7):2241–2247.

Uddip Acharjee Shuvo, Sajib Acharjee Dip, Nirvar Roy Vaskar, and ABM Alim Al Islam. 2024. Assessing chatgpt's code generation capabilities with short vs long context programming problems. In *Proceedings of the 11th International Conference on Networking, Systems, and Security*, pages 32–40.

Robert R Stickels, Evan Murray, Pawan Kumar, Jilong Li, Jamie L Marshall, Daniela J Di Bella, Paola Arlotta, Evan Z Macosko, and Fei Chen. 2021. Highly sensitive spatial transcriptomics at near-cellular resolution with slide-seqv2. *Nature biotechnology*, 39(3):313–319.

Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. 2017. Simultaneous epitope and transcriptome measurement in single cells. *Nature methods*, 14(9):865–868.

Elliott Swanson, Cara Lord, Julian Reading, Alexander T Heubeck, Palak C Genge, Zachary Thomson, Morgan DA Weiss, Xiao-jun Li, Adam K Savage, Richard R Green, and 1 others. 2021. Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using tea-seq. *Elife*, 10:e63632.

Wenzhuo Tang, Hongzhi Wen, Renming Liu, Jiayuan Ding, Wei Jin, Yuying Xie, Hui Liu, and Jiliang Tang. 2023. Single-cell multimodal prediction via transformers. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 2422–2431.

Christina V Theodoris, Ling Xiao, Anant Chopra, Mark D Chaffin, Zeina R Al Sayed, Matthew C Hill, Helene Mantineo, Elizabeth M Brydon, Zexian Zeng, X Shirley Liu, and 1 others. 2023. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624.

Kyle J Travaglini, Ahmad N Nabhan, Lolita Penland, Rahul Sinha, Astrid Gillich, Rene V Sit, Stephen Chang, Stephanie D Conley, Yasuo Mori, Jun Seita, and 1 others. 2020. A molecular cell atlas of the human lung from single-cell rna sequencing. *Nature*, 587(7835):619–625.

Ching-Huei Tsou, Michal Ozery-Flato, Ella Barkan, Diwakar Mahajan, and Ben Shapira. 2025. Bioverse: Representation alignment of biomedical modalities to llms for multi-modal reasoning. *arXiv preprint arXiv:2510.01428*.

Yuequn Wang, Jun Wang, Yanyu Xu, Ning Liu, Bin Liu, Yuliang Li, and Guoxian Yu. 2025. Fmh2st: foundation model-based spatial transcriptomics generation from histological images. *Nucleic Acids Research*, 53(17):gkaf865.

Hongzhi Wen, Wenzhuo Tang, Xinnan Dai, Jiayuan Ding, Wei Jin, Yuying Xie, and Jiliang Tang. 2023. Cellplm: Pre-training of cell language model beyond single cells. *BioRxiv*, pages 2023–10.

Juncheng Wu, Changxin Wan, Zhicheng Ji, Yuyin Zhou, and Wenpin Hou. 2025. Epifoundation: A foundation model for single-cell atac-seq via peak-to-gene alignment. *bioRxiv*.

Yan Wu, Esther Wershof, Sebastian M Schmon, Marcel Nassar, Błażej Osiński, Ridvan Eksi, Zichao Yan, Rory Stark, Kun Zhang, and Thore Graepel. 2024. Perturbench: Benchmarking machine learning models for cellular perturbation analysis. *arXiv preprint arXiv:2408.10609*.

Yuxuan Wu and Fuchou Tang. 2025. scextract: leveraging large language models for fully automated single-cell rna-seq data annotation and prior-informed multi-dataset integration. *Genome Biology*, 26(1):174.

Yihang Xiao, Jinyi Liu, Yan Zheng, Xiaohan Xie, Jianye Hao, Mingzhi Li, Ruitao Wang, Fei Ni, Yuxiao Li, Jintian Luo, and 1 others. 2024. Cellagent: An llm-driven multi-agent framework for automated single-cell data analysis. *arXiv preprint arXiv:2407.09811*.

Chenling Xu, Romain Lopez, Edouard Mehlman, Jeffrey Regier, Michael I Jordan, and Nir Yosef. 2021. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Molecular systems biology*, 17(1):e9620.

Jing Xu, De-Shuang Huang, and Xiujun Zhang. 2024. scmformer integrates large-scale single-cell proteomics and transcriptomics data by multi-task transformer. *Advanced Science*, 11(19):2307835.

Fan Yang, Wenchuan Wang, Fang Wang, Yuan Fang, Duyu Tang, Junzhou Huang, Hui Lu, and Jianhua Yao. 2022. scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. *Nature Machine Intelligence*, 4(10):852–866.

Wenyi Yang, Pingping Wang, Shouping Xu, Tao Wang, Meng Luo, Yideng Cai, Chang Xu, Guangfu Xue, Jinhao Que, Qian Ding, and 1 others. 2024a. Deciphering

cell–cell communication at single-cell resolution for spatial transcriptomics with subgraph-based graph attention network. *Nature Communications*, 15(1):7101.

Xiaodong Yang, Guole Liu, Guihai Feng, Dechao Bu, Pengfei Wang, Jie Jiang, Shubai Chen, Qinmeng Yang, Hefan Miao, Yiyang Zhang, and 1 others. 2024b. Genecompass: deciphering universal gene regulatory mechanisms with a knowledge-informed cross-species foundation model. *Cell Research*, 34(12):830–845.

Wenjin Ye, Yuanchen Ma, Junkai Xiang, Hongjie Liang, Tao Wang, Qiuling Xiang, Andy Peng Xiang, Wu Song, Weiqiang Li, and Weijun Huang. 2024. Objectively evaluating the reliability of cell type annotation using llm-based strategies. *arXiv preprint arXiv:2409.15678*.

Zhen Yuan, Shaoqing Jiao, Yihang Xiao, and Jiajie Peng. 2025. scmamba: A scalable foundation model for single-cell multi-omics integration beyond highly variable feature selection. *arXiv preprint arXiv:2506.20697*.

Yuansong Zeng, Jiancong Xie, Ningyuan Shangguan, Zhuoyi Wei, Wenbing Li, Yun Su, Shuangyu Yang, Chengyang Zhang, Jinbo Zhang, Nan Fang, and 1 others. 2025. Cellfm: a large-scale foundation model pretrained on transcriptomics of 100 million human cells. *Nature Communications*, 16(1):4679.

Fan Zhang, Hao Chen, Zhihong Zhu, Ziheng Zhang, Zhenxi Lin, Ziyue Qiao, Yefeng Zheng, and Xian Wu. 2025a. A survey on foundation language models for single-cell biology. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 528–549.

Fan Zhang, Tianyu Liu, Zhihong Zhu, Hao Wu, Haixin Wang, Donghao Zhou, Yefeng Zheng, Kun Wang, Xian Wu, and Pheng-Ann Heng. 2025b. Cellverse: Do large language models really understand cell biology? *arXiv preprint arXiv:2505.07865*.

Suyuan Zhao, Jiahuan Zhang, and Zaiqing Nie. 2023. Large-scale cell representation learning via divide-and-conquer contrastive learning. *arXiv preprint arXiv:2306.04371*.

A Background and Related Studies

The intersection of large language models (LLMs) and single-cell biology has rapidly expanded in recent years, motivating several surveys and benchmark studies. The ACL 2025 survey on foundation models for single-cell biology (Zhang et al., 2025a) reviews pretrained and fine-tuned architectures across protein, gene, and cell representations. Broader reviews such as Lan et al. (2025) and Bian et al. (2024b) discuss large cellular models (LCMs), emphasizing trends in scaling, tokenization, and transfer learning across biological modalities.

Complementary benchmark efforts, including Single-Cell Omics Arena (Liu et al., 2024) and CellVerse (Zhang et al., 2025b), evaluate LLMs on tasks such as cell-type annotation and question answering, while applied systems like scInterpreter and scExtract (Guo et al., 2023; Wu and Tang, 2025) demonstrate integration into annotation and curation workflows. Widely used non-LLM foundation models, such as the scVI family (Lopez et al., 2018; Ergen et al., 2025; Gayoso et al., 2022; Xu et al., 2021), remain strong baselines for integration, batch correction, and generative modeling, and serve as important calibration points. However, they do not incorporate natural language grounding, ontology alignment, or agentic reasoning, and are therefore treated as contextual baselines rather than survey targets in this work.

Several recent frameworks and benchmarks further contextualize the trends identified in this survey. BMFM-RNA (Dandala et al., 2025; Tsou et al., 2025) introduces a modular Transformer-based framework with controlled comparisons across training objectives (e.g., masked language modeling versus Wasserstein-based losses) under standardized benchmarks, reinforcing our observation that objective choice and evaluation protocol can dominate architectural differences.

Large-scale spatial benchmarks such as HEST-1k (Jaume et al., 2024) and HESCAPE (Gindra et al., 2025) provide paired histology–transcriptomics and cross-condition evaluation settings, revealing trade-offs where improved cross-modal alignment may degrade direct gene-expression prediction. These findings refine our discussion of spatial and multimodal evaluation by highlighting the need to balance alignment metrics with biological fidelity.

Recent multimodal architectures such as SPATIA (Kong et al., 2025) emphasize hierarchical modeling across cellular, niche, and tissue scales using cross-attention and conditional generation, aligning with the Spatial/Multimodal family and underscoring the importance of multi-scale context in spatial reasoning.

Efficiency-oriented designs, exemplified by xTrimoGene (Gong et al., 2023), demonstrate that sparsity-aware architectures can substantially reduce compute and memory costs while preserving performance, adding nuance to scalability considerations beyond model size alone.

Finally, perturbation-focused evaluations have shown that classical methods such as PCA or scVI can outperform Transformer-based models on perturbation fidelity and causal structure preservation, supporting our argument that standard technical metrics (e.g., batch correction) may misrepresent biological validity and that perturbation-aware objectives remain an open challenge for LLM-based models.

Despite their contributions, existing surveys largely frame progress in terms of model architectures or prompting strategies, with limited analysis of how dataset availability, evaluation practice, and modality coverage constrain reported performance. Spatial, perturbation, and multimodal datasets are often discussed only peripherally, standardized benchmarks are rare, and plant single-cell resources are largely absent. Moreover, agentic and tool-augmented systems are typically presented as isolated examples rather than analyzed as a coherent methodological class. LLM4Cell addresses these gaps through a unified, data-centric synthesis that links models, datasets, tasks, and trust-related evaluation dimensions. Unlike transformer-centric or oncology-focused surveys, LLM4Cell centers on how language grounding and agentic design interact with single-cell datasets and evaluation practice, rather than architectural optimization or clinical deployment alone. Throughout the paper, claims about LLM advantages are interpreted relative to these classical foundations, particularly for integration and generation, rather than as absolute performance improvements.

B Data Collection

A.1 Search Strategy

To construct a comprehensive registry of large language models (LLMs) and agentic frameworks in single-cell biology, we systematically screened both peer-reviewed and preprint literature from 2020–2025 across multiple sources. Searches were

Aspect	Transformer-centric Single-Cell Reviews	Multi-omic Surveys	LLM4Cell
Primary scope	Model architectures for single-cell analysis (e.g., Transformers, VAEs)	Cross-modal AI for cancer and omics (genomics, imaging, EHRs)	LLMs and agentic systems for single-cell biology
Organizing axis	Architectural design and performance	Clinical applications and modality fusion	Datasets, tasks, and evaluation practice
Modalities covered	Primarily RNA; limited spatial or perturbation	Genomics, imaging, clinical data	RNA, ATAC, spatial, perturbation, plant
Language grounding	Not considered	Occasionally used for reports or labels	Explicit alignment with text and ontologies
Agentic reasoning	Absent	Rare or conceptual	Analyzed as a first-class category
Evaluation focus	Task accuracy and benchmarks	Clinical utility and outcomes	Grounding, robustness, fairness, privacy
Non-LLM baselines	Central focus	Mixed with deep learning methods	Used as calibration, not survey targets
Key limitation	Limited cross-domain synthesis	Limited cellular resolution	Lack of standardized reasoning benchmarks

Table 1: Qualitative comparison of LLM4Cell with transformer-centric single-cell reviews and broader multimodal oncology or multi-omics surveys.

conducted on PubMed, Google Scholar, arXiv, and Semantic Scholar using combinations of domain- and method-specific keywords, including:

- **General:** "large language model", "foundation model", "LLM", "multimodal model", "generative AI"
- **Domain-specific:** "single-cell", "scRNA-seq", "scATAC", "multiome", "spatial transcriptomics", "perturb-seq", "cell atlas", "biomedical foundation model"
- **Integration / reasoning:** "ontology alignment", "text-bridge", "agentic framework", "biological reasoning", "cross-modal integration", "annotation model"

After removing duplicates, the combined queries returned approximately **8,020 papers** from 2020–2025. Of these, 5,510 remained after filtering for English-language articles with available abstracts. We manually reviewed titles, abstracts, and model descriptions to identify works directly involving large-scale pretrained models, multimodal transformers, or language-based reasoning systems applied to single-cell or cellular-level omics data.

A.2 Inclusion and Screening Criteria

Studies were included if they met at least one of the following criteria:

1. Introduced or evaluated a large pretrained model ($\geq$10M parameters) for single-cell or multi-omic analysis.
2. Employed textual grounding, ontology prompts, or natural-language interfaces to interpret single-cell data.
3. Proposed agentic or reasoning-based frameworks for biological tasks (annotation, trajectory, perturbation, or integration).
4. Released code, preprints, or benchmarks relevant to cellular foundation models.

Methodological, biological, and dataset-based duplicates were merged under representative entries. This curation yielded **58 distinct models** across five methodological families: foundation, text-bridge, spatial/multimodal, epigenomic, and agentic frameworks. Each was annotated with task domain, modality, training data, and evaluation scope (Appendix Table 8).

A.3 Inclusion of Preprints

We included arXiv and bioRxiv preprints to ensure coverage of recent and influential models not yet published in peer-reviewed venues. In single-cell research, the majority of foundational architectures (e.g., *scGPT*, *scFoundation*, *CellLM*) initially appeared as preprints months before journal acceptance. Given the field's rapid evolution and small number of LLM-scale models, preprints represent essential early contributions to reproducibility and transparency. Each preprint was manually cross-verified for active GitHub or Zenodo links to confirm technical validity and data availability.

A.4 Resulting Corpus

The final corpus spans 58 methods across 37 unique institutions, 40+ publicly accessible datasets, and

8 major analytical tasks. This dataset–method pairing forms the empirical foundation for the taxonomic and domain analyses presented in the main paper.

C Evaluation Rubric Protocol

To analyze trends across models without imposing a unified benchmark, we assess each method using a ten-dimension evaluation rubric covering biological grounding, robustness, generalization, interpretability, scalability, fairness, privacy, and agentic behavior. Each dimension is evaluated using a three-level categorical scale Present, Partial, or Absent based solely on explicit evidence reported in the original publication, without extrapolation beyond the authors' claims. Rubric annotations for all models are recorded explicitly in the `S2_Method_Domains` sheet of the supplementary spreadsheet and Figure 3, enabling direct inspection of Present, Partial, and Absent assignments for each dimension. The rubric captures whether a capability is reported and evaluated, not whether it is correct, optimal, or competitive.

A dimension is marked as Present when the capability is directly implemented and empirically evaluated. At minimum, a Present assignment requires an explicit implementation tied to a reported experiment, ablation, or quantitative analysis targeting that dimension, rather than a descriptive claim or post hoc interpretation. Partial indicates that the dimension is meaningfully addressed for example through model design, stated objectives, or limited analysis but without comprehensive or task-specific evaluation. In the absence of clear evidence, the dimension is conservatively marked as Absent. No inference is made beyond what is documented by the authors.

Rationale for partial coverage. The *Partial* category is used to distinguish between explicit, task-level evaluation and weaker forms of evidence such as architectural intent, stated objectives, or limited demonstrations. Collapsing the rubric to a binary Present/Absent scheme would overstate coverage by treating design claims and comprehensive evaluations equivalently. By separating Partial from Present, the rubric preserves this distinction while remaining conservative: dimensions are marked as Present only when directly implemented and empirically evaluated, Partial when meaningfully addressed without full validation, and Absent otherwise.

For summary purposes, the *Domains Covered* count reported in Appendix Table 8 includes both Present and Partial dimensions. This aggregation reflects the reported scope of each method rather than evaluation completeness or performance strength. As a result, coverage counts should be interpreted as a descriptive indicator of which domains a model engages with, not as a measure of quality or superiority. The rubric is not intended to rank models or provide absolute performance comparisons. Instead, it is designed to surface uneven coverage, reporting gaps, and evaluation blind spots across model families, tasks, and modalities.

Annotation consistency. Rubric annotations were produced through a multi-stage internal review process. An initial set of authors independently annotated methods using the shared rubric protocol. A second group of authors then independently reviewed these assignments. Discrepancies or borderline cases were discussed jointly and resolved through consensus, guided by explicit decision rules documented in the rubric (e.g., criteria distinguishing text-bridge from multimodal models).

We did not compute formal inter-annotator agreement statistics, as the rubric is intended as a descriptive synthesis of reported capabilities rather than a subjective labeling benchmark. Instead, consistency was enforced through independent review, conservative decision rules, and consensus-based resolution. All final assignments are released in the supplementary registry to support auditability and reuse. This process reflects standard practice for survey curation rather than a controlled annotation study.

Evidence sources and annotation process. Rubric assignments were based on explicit evidence in the original papers, including (i) method descriptions, (ii) reported experiments and metrics, (iii) ablation or evaluation sections, and (iv) stated design objectives when empirical validation was absent. Each model was independently annotated by two authors using the same written criteria. Disagreements were flagged and resolved through discussion with a third author, following a conservative rule favoring lower coverage (e.g., Partial over Present) in ambiguous cases. This process aimed to ensure consistent application of rubric thresholds across model families and tasks. We do not report formal inter-annotator agreement statistics, as

the rubric is descriptive rather than evaluative, but calibration examples were used to align decisions.

Counting and aggregation rules. To ensure consistent aggregation across heterogeneous methods, we apply the following counting rules. Models that support multiple analytical tasks (e.g., annotation and integration) are counted once for each task they explicitly address. Models that span multiple methodological families are assigned a single *primary family*, determined by their core training objective and dominant architectural role (e.g., foundation pretraining, language grounding, or agentic control). Secondary capabilities are discussed qualitatively but are not double-counted in summary statistics. These rules prevent inflation of counts while preserving visibility of multi-task and hybrid models.

Scope of cross-study performance observations. Observations relating reported performance gains to data scale or modality pairing are intended as qualitative trends derived from cross-study synthesis, rather than quantified statistical relationships. A formal meta-analysis (e.g., aggregating annotation accuracy, ARI, or perturbation correlation across studies) is currently infeasible due to heterogeneous datasets, evaluation metrics, preprocessing pipelines, and non-comparable experimental settings across publications. We therefore avoid effect-size estimation and interpret these patterns as descriptive signals of where progress is reported, not as causal or statistical claims. The released supplementary registry is designed to enable future quantitative analyses once standardized benchmarks and shared evaluation protocols become available.

Scope and reporting bias. Rubric annotations are based on capabilities and evaluations explicitly reported by the original authors and do not involve independent re-implementation or re-analysis. As a result, the rubric may reflect reporting biases present in the literature. This design choice is intentional and aligns with the goal of surveying reported practices rather than validating model performance. By releasing the full annotation registry, we enable readers to audit, reinterpret, or revise these assignments as the field evolves.

Proposed reference panel benchmark. To facilitate reproducible comparison without requiring full-scale re-evaluation, we outline a compact reference panel consisting of one to two widely used datasets per task (e.g., PBMC for annotation, Visium DLPFC for spatial mapping, sci-Plex for perturbation). A small representative model subset—covering foundation, text-bridge, multimodal, and agentic families—could be evaluated on this panel using standardized task-specific metrics. This design is intended as a lightweight calibration benchmark to validate high-level survey conclusions rather than to establish new state-of-the-art results.

Meta-observations and quantification. Cross-study observations relating reported gains to data scale or modality pairing are intended as qualitative trends derived from synthesis of published results, not as quantified correlations or effect-size estimates. Formal meta-analysis is currently infeasible due to heterogeneous datasets, metrics, and evaluation protocols. We therefore avoid statistical claims and interpret these patterns as descriptive signals rather than causal relationships.

Boundary cases: text-bridge vs multimodal. Models that use textual metadata (e.g., cell-type labels, captions, or gene descriptions) solely as weak supervision or auxiliary input—without explicitly aligning molecular embeddings with language representations—are classified as *multimodal* rather than text-bridge in our taxonomy. In contrast, text-bridge models explicitly optimize for semantic alignment between molecular and textual spaces through contrastive objectives, shared embedding spaces, or encoder–decoder translation.

For example, a spatial model that conditions on histology captions or uses text labels during training but does not learn a joint language–molecule embedding is treated as multimodal. Conversely, models such as Cell2Sentence or GenePT, which directly align gene or cell embeddings with natural-language representations, are categorized as text-bridge. This distinction reflects whether language serves as weak supervision or as a first-class semantic representation.

D Extended Dataset Summaries

D.1 RNA Atlases

Foundational scRNA-seq atlases such as *Tabula Sapiens* (>1 M cells, 24 tissues) and *Tabula Muris Senis* (mouse, multi-organ across aging) provide references for annotation and batch-effect evaluation. Organ-specific datasets like the *Human Lung Cell Atlas (HLCA)* and the *Allen Brain Map* enable

domain-specific benchmarking and ontology-aware evaluation.

D.2 Chromatin and ATAC Data

The *Cusanovich mouse sci-ATAC* atlas, human adult/fetal scATAC atlases, and joint RNA–ATAC modalities (*SHARE-seq*, *SNARE-seq2*) characterize regulatory landscapes supporting trajectory and GRN inference, though sparsity limits large-scale pretraining.

D.3 Multiome and Tri-Modal Data

Datasets such as *TEA-seq*, *DOGMA-seq*, *ASAP-seq*, and *CITE-seq* pair transcriptomic, epigenomic, and proteomic modalities, enabling cross-modal translation and representation learning. The *Multiome Benchmark Pack* aggregates >40 datasets for standardization but remains smaller than large RNA atlases.

D.4 Spatial Transcriptomics and Imaging

Platforms including *10x Visium/HD*, *Slide-seqV2*, and *DBiT-seq* provide spatial gene-expression maps, while imaging technologies (*MERFISH*, *STARmap*, *Stereo-seq*, *CosMx*, *Xenium*) capture near single-molecule resolution for tissue-level reasoning.

D.5 Perturbation and Drug-Response Screens

Large CRISPR-based datasets (*Replogle 2022 Perturb-seq*, *Norman 2019*, *sci-Plex*) and community benchmarks (*Virtual Cell Challenge*) quantify transcriptional responses for causal and policy evaluation in agentic frameworks.

D.6 Plant Single-Cell Datasets

scPlantDB (67 datasets, 17 species, 2.5 M cells) and *PlantscRNAdb* provide unified plant atlases; *Arabidopsis E-CURD-4* and transgenic tobacco scRNA-seq datasets enable cross-kingdom and stress-response modeling.

D.7 Critical Observations

(1) RNA atlases dominate in scale; **(2)** Chromatin and spatial data are fragmented; **(3)** Privacy and licensing restrict clinical data; **(4)** Cross-species coverage is limited; **(5)** Paired modalities offer promising benchmarks for next-generation LLMs.

Representative datasets and links are provided in Tables A1–A3.

E Extended Details on Model Taxonomy

E.1 Foundation Models (Extended Details)

Foundation-scale models constitute the base layer for language-driven single-cell analysis. They are trained on molecular profiles—typically scRNA-seq or integrated multi-organ datasets—to learn generalizable representations of cellular states, gene programs, and perturbations.

Training corpora and scale. Most use atlases containing 10^6–10^8 cells. *scGPT* and *scFoundation* were pretrained on human–mouse atlases from Tabula Sapiens and HCA; *Geneformer* aggregates thousands of GEO datasets; *iSEEEK* and *CellFM* span 11–17 organs and species, enabling cross-tissue generalization.

Architectural variations. Most adopt Transformer backbones with gene tokens as contextual units. *tGPT* and *scBERT* encode ranked or discrete gene sequences, while *Geneformer* and *scFoundation* use masked-token modeling. Hybrid designs such as *scGraphformer* (Fan et al., 2024) and *scRDiT* (Dong et al., 2025) incorporate graph or diffusion-based attention, improving spatial and regulatory context.

Learning objectives. Common objectives include masked-gene prediction, expression denoising, rank-based reconstruction, and cross-species contrastive learning (*UCE*, *GeneCompass*). Optimization typically minimizes cosine or KL divergence across gene embeddings, similar to masked-language modeling in NLP.

Applications and limitations. Foundation embeddings transfer effectively to annotation, integration, and trajectory inference but lack explicit biological grounding. Ontology alignment and reasoning remain external, and interpretability is largely post-hoc (attention heatmaps, gene-ranking). These limitations prompted the development of Text-Bridge and Agentic frameworks discussed in later sections.

E.2 Text-Bridge LLMs (Extended Details)

Text-bridge models are defined as methods that explicitly learn joint representations between molecular features and natural language or ontological concepts, rather than using text solely as auxiliary supervision or metadata. Concretely, a model is categorized as text-bridge only if its training objective enforces alignment between molecular embeddings and textual representations (e.g., via contrastive learning, shared embedding spaces, or sequence-to-text generation).

Representative examples include *scELMo*, which fuses ELMo-derived gene metadata with cell

latent vectors; *CellLM*, which uses ontology-aware prompts for natural-language annotation; *Cell2Sentence* and *Cell2Text*, which map gene-expression vectors to descriptive sentences through cross-modal contrastive objectives; and *GenePT*, which jointly pre-trains gene and disease embeddings using biomedical literature. In contrast, models that incorporate textual labels, captions, or descriptions only as weak supervision without explicit alignment objectives are categorized as multimodal rather than text-bridge.

Architecturally, text-bridge systems typically combine Transformer-based molecular encoders with domain-tuned language models (e.g., BERT variants), optimized using cosine-similarity or contrastive losses. These models score highly in *Explainability* and *Emerging Paradigms* dimensions (Appendix Table 8), but remain limited by vocabulary coverage, ontology completeness, and the absence of full multi-omic grounding.

E.3 Spatial and Multimodal Models (Extended Details)

Spatial models combine molecular and positional information to learn tissue-aware embeddings. *TransformerST* performs cross-scale attention between spatial spots and gene tokens; *spaLLM* aligns histology-derived captions with expression vectors using LLM guidance; and *OmiCLIP* links image and molecular embeddings via contrastive pretraining. Multi-omic extensions such as *scMMGPT* and *FmH2ST* integrate RNA, ATAC, and protein modalities within a shared latent space. Grounding typically relies on marker-based or atlas alignment rather than textual ontologies. Evaluation metrics include spot-level reconstruction accuracy and correlation with histological segmentation (DLPFC, Visium HD). These models show high *biological grounding* but moderate *explainability* due to visual-feature opacity.

E.4 Alignment with PerturBench metrics. (Extended Details)

For perturbation and trajectory modeling, recent benchmarks such as PerturBench (Wu et al., 2024) emphasize rank-based and distributional metrics (e.g., Spearman correlation, Wasserstein distance) that better capture biological fidelity than reconstruction loss alone. Based on these insights, we recommend that future studies report at minimum (i) rank-based agreement of differentially expressed genes, (ii) distributional similarity between predicted and observed perturbation responses, and (iii) robustness across unseen perturbations. Adopting a common metric set would improve comparability across models and tasks.

E.5 Epigenomic Models (Extended Details)

Inputs include chromatin-accessibility matrices, motif sequences, and enhancer–promoter links. *EpiFoundation* pre-trains on the ENCODE scATAC compendium; *EpiBERT* encodes open-chromatin sequences using masked-region objectives; and *EpiAttend* models enhancer–promoter coupling through cross-region attention. State-space architectures (*scMamba*, *GeneMamba*) enable efficient context propagation over tens of thousands of genomic peaks. *ChromFound* and *GET* jointly model RNA and ATAC modalities for regulatory inference. Grounding leverages atlas-derived cCRE annotations and transcription-factor motifs. Performance is strong for *biological grounding* and *multi-omics alignment*, but interpretability is limited to motif attention visualization.

E.6 Agentic Frameworks (Extended Details)

Agentic models couple (i) a pretrained molecular encoder, (ii) an LLM-based controller, and (iii) external tool interfaces. *scAgent* executes multi-step annotation workflows via ontology queries; *CellVerse* coordinates multiple domain agents for transcriptomic, spatial, and literature reasoning; *EpiAgent* extends agentic control to enhancer–gene analysis; and *Teddy*/*scPilot* (Joodaki et al., 2024) provide lightweight orchestrators for benchmarking pipelines. Grounding sources include the Human Cell Atlas ontology, UBERON, and PubMed. These systems score highest in *Explainability* and *Emerging paradigms*, marking a transition from static embeddings to interactive, goal-directed modeling. Open challenges include evaluation of reasoning accuracy, privacy, and reproducibility.

Toward operational evaluation of agentic systems. To move beyond descriptive analysis, we outline concrete evaluation dimensions for agentic single-cell systems: (i) task completion success under fixed workflows, (ii) tool-use efficiency (number and correctness of tool calls), and (iii) reasoning stability under stress tests. Relevant stress scenarios include prompt perturbation, simulated tool failure, ontology inconsistencies, and counterfactual queries. Metrics adapted from agent evaluation (e.g., task success rate, tool-use error rate) could be instantiated for single-cell workflows such as annotation, dataset retrieval, and pathway analysis.

F Supplementary Tables and Reproducibility

To support transparency and reproducibility, we provide supplementary tables that systematically catalog the datasets and methods surveyed in this work. These materials are hosted via an anonymous Zenodo record and are intended to complement the taxonomy and qualitative analysis presented in the main text.

Supplementary Materials. The supplementary file consists of a single spreadsheet containing two sheets:

- **S1_Datasets**: A curated overview of datasets used in LLM-enabled cellular and single-cell studies. For each dataset, we report the data modality, primary analytical tasks, a brief description, dataset scale, and a public access link when available.
- **S2_Method_Domains**: A structured registry of representative methods and systems, including model name, publication venue and year, method category, supported modalities, grounding strategy, agentic design (if applicable), primary task, and per-method annotations for all ten evaluation rubric dimensions (Present / Partial / Absent), with citation links to the original sources.

Usage and Interpretation. These tables are designed as a structured reference that enables readers to (i) identify commonly used datasets across tasks and modalities, (ii) compare grounding strategies and agentic designs across methods, and (iii) trace how different biological domains are addressed by existing LLM-based approaches. The tables align directly with the taxonomy and discussion in Sections 4 of the main paper.

Auditability of rubric annotations. All rubric assignments in `S2_Method_Domains` are grounded in explicit evidence reported in the cited publications, such as described training objectives, architectural components, or evaluation experiments. Rather than reproducing textual excerpts, we provide direct citation links for each method, allowing readers to verify, reinterpret, or refine the annotations based on the original sources.

Reproducibility. All quantitative summaries, categorical breakdowns, and survey figures in the paper (e.g., distributions over modalities, tasks, grounding types, and domains) are derived directly from the released supplementary registry. The registry is provided in a machine-readable CSV-compatible spreadsheet and explicitly records, for each method, which of the ten evaluation domains are marked as *Present*, *Partial*, or *Absent*.

The *Domains Covered* counts reported in tables and figures are computed by aggregating these per-method annotations according to the documented counting rules (i.e., counting both Present and Partial as covered). The exact aggregation logic and decision criteria are described in the Evaluation Rubric Protocol, enabling readers to reproduce all reported counts and percentages by filtering or summing the corresponding columns (e.g., `Domain_X_Status`) using standard spreadsheet tools or simple scripts in Python or R. No additional preprocessing or proprietary resources are required. The supplementary registry is intentionally structured to support auditability, reuse, alternative aggregations, and extension to new models or domains. Example aggregation scripts are trivial (e.g., summing non-Absent entries per domain) and are described in the appendix.

Access. The full supplementary registry, including dataset metadata, model attributes, task mappings, and per-domain rubric annotations, is available via an anonymous Zenodo record: `https://zenodo.org/records/18141566`.

Upon acceptance, the registry will be versioned and linked to a public repository to facilitate stable citation and incremental updates.

Maintaining currency and evolving resources. LLM4Cell reflects the state of the literature at the time of writing, including peer-reviewed papers and widely cited preprints. To address forward-dated manuscripts and rapidly evolving repositories, we release a machine-readable registry of datasets and methods (Supplementary Tables S1–S2) that can be updated independently of the paper. While we do not maintain a live leaderboard, the registry is designed to support community-driven updates and extensions, enabling future work to track new models, datasets, and benchmarks as they emerge.

Handling preprints and evolving resources. The survey includes widely cited preprints where peer-reviewed alternatives are unavailable. To mitigate drift as results evolve, each entry in the supplementary registry records publication status and year, enabling updates independent of the paper text. Rather than maintaining a live leaderboard,

we release a versioned, machine-readable registry designed for periodic community updates, ensuring that model status, datasets, and evaluation claims can be revised as the literature matures.

G Datasets

Table 2: **RNAseq** single-cell datasets used in LLM-based single-cell research.

Dataset	Tasks	Description	Scale	Link / Citation
Tabula Sapiens v2	Annotation, Integration	Multi-organ *human* atlas (∼1.1 M cells, 28 organs across 24 donors) capturing cell-type heterogeneity and cross-tissue transcriptional variation; droplet and plate-seq modalities.	1.1 M cells / 28 organs / 24 donors	CZ Biohub Portal
Tabula Muris (mouse, multi-organ)	Annotation, Integration	Mouse multi-organ atlas of gene expression combining droplet and FACS; enables cross-tissue comparison and batch integration.	∼100,000 cells / 20 organs	HCA Project Page
Mouse Cell Atlas (scMCA)	Annotation, Integration	Comprehensive mouse single-cell atlas constructed with Microwell-seq; supports cell-type matching via scMCA tool.	>400,000 cells / >40 tissues	MCA Portal (ZJU)
Human Lung Cell Atlas (HLCA v1.0)	Annotation, Integration, Disease Modeling	Integrated human lung reference built from 49 scRNA-seq datasets (∼2.4 M cells) across 16 studies; unified epithelial, immune, and stromal annotations.	∼2.4 M / 49 datasets / 16 studies	HLCA v1.0 Portal
HCA lung project (example project page)	Annotation, Integration, Disease Modeling	HCA lung cohort integrating ∼2.4 M single cells from 49 datasets; harmonized annotations for airway, immune, and endothelial populations.	2.4 M cells / 49 datasets	HLCA Project Page
Allen Brain Map – Cell Types RNA-seq (human & mouse)	Annotation, Cross-species, Trajectory	Single-cell and single-nucleus transcriptomes from human and mouse brain regions; supports cross-species comparison and cell-type taxonomy benchmarking.	>1.8 M cells / multiple cortical & subcortical regions	Allen Brain Cell Types Database
Yale Lung Disease Cell Atlases (e.g., COPD)	Annotation, Disease Modeling	Single-cell RNA atlases of diseased human lungs (COPD, IPF) from Yale's Kaminski lab, capturing altered epithelial, endothelial, and immune populations.	∼300k cells (IPF + COPD + controls)	HCA/Yale Atlas

Table 3: **ATACseq** and chromatin-accessibility datasets used in LLM-based single-cell research.

Dataset	Tasks	Description	Scale	Link / Citation
Mouse sci-ATAC-seq atlas (Cusanovich et al., 2018)	Annotation, Trajectory, GRN inference	Landmark single-cell ATAC-seq atlas profiling ∼100,000 nuclei across 13 adult mouse tissues using combinatorial indexing (sci-ATAC-seq); enables cross-tissue regulatory and lineage analysis.	∼100k nuclei / 13 tissues	GSE111586, Science 2018
Human adult scATAC atlas (Zhang et al., 2021)	Annotation, GRN, Cross-tissue integration	Comprehensive single-cell chromatin accessibility atlas of adult human tissues, defining candidate cis-regulatory elements (cCREs) across 25 tissues and 222 cell types; foundation of the ENCODE human cCRE registry.	∼472k nuclei / 25 tissues	Nature 2021, ENCODE Portal
Human fetal scATAC atlas (Domcke et al., 2020; GSE149683)	Annotation, Developmental trajectory, GRN	Single-cell ATAC-seq atlas of human fetal tissues profiling 15 organs across mid-gestation; reveals developmental enhancer activity and regulatory lineage trajectories.	∼720k nuclei / 15 organs	Science 2020, GSE149683
Massively parallel scATAC (Satpathy et al., 2019)	Annotation, Trajectory, Regulatory modeling	Pioneering high-throughput scATAC-seq dataset of ∼200k immune and cancer nuclei enabling scalable mapping of open-chromatin landscapes; forms benchmark for lineage and immune-cell trajectory studies.	∼200k nuclei / blood and tumor tissues	Nat. Biotechnol. 2019, GSE123581
ENCODE portal (single-cell experiments)	Annotation, Integration, Regulatory modeling	Centralized repository from the ENCODE Consortium aggregating thousands of single-cell RNA and ATAC assays from human and mouse tissues; provides uniformly processed metadata, peak calls, and cCRE annotations for benchmarking.	>2,000 single-cell assays / multiple tissues	ENCODE Portal
T cell epigenetic atlas (Giles et al., 2022)	Trajectory, GRN, Disease modeling	Single-cell ATAC-seq atlas profiling chromatin accessibility across human T-cell activation, exhaustion, and differentiation states; supports trajectory reconstruction and immune-epigenetic modeling.	∼150k nuclei / T-cell subsets	Nat. Immunol. 2022

Table 4: **Multiome** datasets (paired/tri-omic: RNA + ADT, RNA + ATAC, ATAC + ADT, and tri-modal) used in LLM-based single-cell research.

Dataset	Tasks	Description	Scale	Link / Citation
UCSC Cell Browser Hub	Annotation, Integration, Visualization	Aggregated repository of hundreds of public single-cell datasets across species and modalities with metadata and embeddings; useful for exploratory analysis and reference selection.	>1,200 datasets / multi-species	UCSC Cell Browser Hub
Human Cell Atlas (HCA) data browser (multi-project)	Annotation, Integration, Cross-project mapping	A global human cell atlas aggregating ∼63.2 M cells across 515 projects and >11,000 donors, spanning diverse tissues and modalities; enables cross-study reference and meta-integration.	∼63.2 M cells / 515 projects / 11,000+ donors	HCA Data Portal
Azimuth reference collections (PBMC, lung, kidney, fetal) (RNA + ADT)	Annotation, Integration	Curated single-cell reference atlases by the Satija Lab for automated cell-type mapping in Seurat/Azimuth; harmonized labels and multimodal ADT features.	100k–1 M cells across multiple organs	Azimuth Data Portal
TEA-seq (tri-modal RNA + ATAC + ADT; GSE158013)	Integration, Perturbation, Multi-modal reasoning	Tri-modal PBMC dataset measuring transcriptome, chromatin accessibility, and surface proteins; supports alignment/fusion of omics layers and pretraining.	∼100k cells	GSE158013
DOGMA-seq (RNA + ATAC + Protein)	Integration, Perturbation, Cross-modal reasoning	Tri-modal profiling of transcriptome, chromatin, and proteins in human immune cells; benchmark for joint embeddings and cross-modal translation.	∼50k cells / PBMCs	Nature Cell Biol. 2022 / GSE184715
ASAP-seq (ATAC + Protein)	Integration, GRN, Epigenetic modeling	Paired chromatin accessibility and surface proteins via antibody-derived tags; links cis-regulatory variation with immune phenotypes.	∼100k nuclei / PBMCs	Nat. Biotechnol. 2021 / GSE162690
CITE-seq compendia (RNA + ADT)	Annotation, Integration, Transfer learning	Large compendium of paired RNA and surface-protein profiles across blood and tissue; enables multimodal foundation pretraining and zero-shot annotation.	>500k cells across multiple tissues	Nat. Methods 2019 / GSE100866
Multiome Benchmark Pack (QuKun Lab)	Integration, Scalability, Batch-effect analysis	Public benchmark suite consolidating 25 RNA + Protein, 12 RNA + ATAC, and 4 tri-omic datasets (CITE/TEA/DOGMA), standardized for cross-modal and large-scale integration testing.	41 datasets total	Benchmark Portal

Table 5: **Spatial transcriptomics and imaging** datasets used in LLM-based single-cell research.

Dataset	Tasks	Description	Scale	Link / Citation
Slide-seq / Slide-seqV2	Spatial mapping, Trajectory, Integration	High-resolution spatial transcriptomics of mouse brain (hippocampus, cerebellum, cortex) at ∼10 µm bead resolution; supports neighborhood and spatial-domain reconstruction benchmarks.	∼50k–100k spots per tissue	Nat. Biotechnol. 2020 / SCP815
MERFISH (imaging)	Spatial mapping, Pathway, Annotation	Multiplexed error-robust fluorescence in situ hybridization (MERFISH) datasets profiling millions of mouse-brain cells with 3D coordinates and 483–1,000-gene panels; benchmark for high-resolution spatial mapping.	∼4 M cells / 483–1,000 genes	Vizgen MERFISH Portal / Science 2018
Stereo-seq (STOmics)	Spatial mapping, Developmental trajectory	Genome-scale Stereo-seq datasets with submicron resolution; includes MOSTA (mouse embryo) and 3D *Drosophila* atlases for developmental and cross-species modeling.	∼100 µm × 100 µm tiles / millions of spots	Cell 2022 / STOmics Data Hub
DBiT-seq & spatial multi-omics	Spatial mapping, Integration, Multi-omic reasoning	Deterministic barcoding in tissue (DBiT-seq) capturing spatial RNA and protein expression in mouse embryo and human lymph node; supports multi-omic spatial benchmarks.	∼20k spots / 100 µm grids	Nat. Biotechnol. 2020 / GSE152506
10x Visium / Visium HD	Spatial mapping, Annotation, Integration	Widely used capture-array platform for spatial transcriptomics; Visium uses 55 µm spots, Visium HD extends to 2 µm grids with improved gene recovery and FFPE compatibility.	∼5k–55k spots per tissue	10x Genomics Visium Portal
Xenium / CosMx (in situ)	Spatial mapping, Clinical translation, Privacy	In situ spatial transcriptomics platforms (10x Xenium, NanoString CosMx) profiling thousands of transcripts in human FFPE and fresh-frozen tissues (e.g., breast, colon, NSCLC); bridge omics and histology for clinical translation.	10k–100k cells per section	Xenium Explorer / CosMx Portal
HEST-1k	Spatial alignment, Annotation, Robustness	Large-scale paired histology–spatial transcriptomics benchmark comprising over 1,000 tissue sections with aligned whole-slide images and spatial gene expression; designed to evaluate cross-slide generalization, and model robustness.	∼1,000 sections / paired WSI–ST	NeurIPS 2024 (HEST-1k)

Continued on next page

Dataset	Tasks	Description	Scale	Link / Citation
HESCAPE	Spatial perturbation, Generalization, Robustness	Perturbation-aware spatial transcriptomics benchmark capturing transcriptional and spatial responses across experimental conditions; supports evaluation of spatial robustness, perturbation generalization, and condition shift.	Multiple tissues / perturbation settings	ICCV 2025 (HESCAPE)
STimage-1K4M	Super-resolution, Spatial mapping, Imputation	Large-scale image–spatial transcriptomics benchmark with over 1.4 million paired histology–ST samples, enabling evaluation of super-resolution image-to-expression mapping.	∼1.4M image–ST pairs	NeurIPS 2024 (STimage-1K4M)

Table 6: **Perturbation and Drug-Response** single-cell datasets used in LLM-based frameworks.

Dataset	Tasks	Description	Scale	Link / Citation
Genome-scale Perturb-seq (Replogle 2022) processed data	Perturbation, GRN, Causal inference	Largest CRISPR-based Perturb-seq dataset profiling ∼2.5 M single cells across >2 000 genetic perturbations; benchmark for causal network inference and representation learning.	∼2.5 M cells / >2 000 perturbations	Cell 2022 / Figshare Dataset
Norman et al. 2019 Perturb-seq	Perturbation, Combinatorial GRN	Foundational combinatorial CRISPR Perturb-seq dataset (immune + cancer models); establishes feasibility of pooled functional genomics at single-cell resolution.	∼200 k cells / >250 combinations	Science 2019 / GSE133344
Dixit et al. 2016 (GSE90063)	Perturbation, Combinatorial GRN	Early Perturb-seq study targeting 24 genes in macrophages; establishes pipeline for pooled CRISPR screens with single-cell RNA-seq.	∼100 k cells / 24 target genes	Cell 2016 / GSE90063
Adamson et al. 2016	Perturbation, Stress-response, GRN	Single-cell CRISPR Perturb-seq of ER-stress pathways in K562 cells; benchmark for pathway-level perturbation responses.	∼30 k cells / 10 perturbations	Cell 2016 / GSE90060
sci-Plex collection (drug screens) + cellxgene portal	Perturbation, Stress-response, GRN	Multiplexed chemical-perturbation scRNA-seq of >650 k cells treated with 188 compounds via sci-Plex barcoding; pharmacotranscriptomic profiles for drug response modeling.	∼650 k cells / 188 drugs	Cell 2020 / cellxgene Collection

Continued on next page

Dataset	Tasks	Description	Scale	Link / Citation
Compressed Perturb-seq (immune LPS, 598 genes)	Drug response, Perturbation, Benchmarking	Low-multiplicity Perturb-seq library targeting 598 immune genes in macrophages under LPS stimulation; enables compressed experimental design for causal modeling.	∼300 k cells / 598 targets	bioRxiv 2021 / GSE179924
In vivo Perturb-seq (brain / ASD genes)	Perturbation, GRN, Neuroscience	*In vivo* CRISPR Perturb-seq targeting ∼30 ASD-linked genes in mouse cortex; maps neuronal gene-regulatory networks in native contexts.	∼200 k cells / 30 genes	Science 2023 / GSE216113
Virtual Cell Challenge PBMC cytokine perturbations	Perturbation, GRN, Benchmarking	Community benchmark dataset (ARC Institute 2024) of ∼300 k PBMC cells under cytokine stimulation and CRISPRi perturbations; standardized splits for LLM and state-transition evaluation.	∼300 k cells / 150 perturbed genes	ARC Challenge Repository / Kaggle

Table 7: **Plant RNA** single-cell transcriptomics datasets used in LLM research.

Dataset	Tasks	Description	Scale	Link / Citation
scPlantDB (meta-collection)	Annotation, Integration, Meta-analysis	Comprehensive plant single-cell transcriptome database integrating 67 datasets from 17 plant species (∼2.5 M cells); unified preprocessing and annotations; supports cross-species modeling and plant-specific LLMs.	∼2.5 M cells / 67 datasets / 17 species	scPlantDB Portal
PlantscRNAdb	Annotation, Marker discovery	Curated database of plant cell-type marker genes across four species (Arabidopsis, rice, maize, tomato); supports ontology-based annotation and cell identity benchmarking.	4 species / multiple tissues	PlantscRNAdb Portal
Arabidopsis scRNA-seq (E-CURD-4)	Annotation, Trajectory, Development	Baseline scRNA-seq dataset of Arabidopsis thaliana root and leaf tissues (∼10,779 cells); used for developmental lineage and differentiation analysis.	10,779 cells / 2 tissues	EBI Array-Express E-CURD-4
Tobacco leaf scRNA-seq (transgenic antibody line)	Annotation, Perturbation, Stress response	Single-cell transcriptomic profiling of transgenic *Nicotiana tabacum* leaves expressing llama antibody; identifies immune-like responses and cell-type heterogeneity under genetic perturbation.	∼25k cells / leaf tissue	Sci. Data 2023
Plant scRNA Browser (PscB)	Annotation, Visualization, Cross-tissue integration	Online visualization hub aggregating plant scRNA-seq datasets (Arabidopsis, rice, *Wolffia*) with harmonized metadata and interactive UMAP-based cell-type search.	15+ datasets / 3+ species	PscB Data Portal

H Methods Comparison

Table 8: Comparison of single-cell LLM and agentic methods. Coverage indicates the number of domain dimensions (out of 10) for which explicit evidence is reported in the original publication; it does not reflect performance or ranking.

Model (Year)	Published where	Category	Modality	Grounding Type	Agentic (Y/N)	Primary Task	Domains Covered
scGPT (Cui et al., 2024)	Nature Methods	Foundation	Multiomics (scRNA + optionally multi-omic modes)	Atlas (trained on large atlas of single-cell datasets)	No	Annotation, Integration, Perturbation (annotation as primary)	6
Geneformer (Theodoris et al., 2023)	Nature	Foundation	scRNA	None / Rank-based (implicit)	No	Cell classification, in-silico perturbation, network prediction	6
scFoundation (Hao et al., 2024)	Nature Methods	Foundation	scRNA / multi-omics (transcriptomics focus)	Atlas (large pretrained corpus)	No	Cell annotation, perturbation prediction, drug response, gene module inference	7
CellFM (Zeng et al., 2025)	Nature Communications	Foundation	scRNA	Atlas / value-projection	No	Learns embeddings from 100M cells and supports annotation, perturbation, gene function	7
iSEEEK (Shen et al., 2022)	Briefings in Bioinformatics	Foundation	scRNA	Atlas (11.9M cells, human + mouse)	No	Integrates massive single-cell datasets via gene-ranking similarity to enable scalable cross-dataset embedding and clustering	5
tGPT (Shen et al., 2023)	iScience	Foundation Model (Autoregressive)	scRNA	Atlas (22.3 M cells)	No	Treats gene-expression ranks as token sequences and learns generative embeddings for cell clustering, trajectory inference, and bulk-tissue analysis	5
scBERT (Yang et al., 2022)	Nature Machine Intelligence	Foundation Model (Encoder-only)	scRNA	Atlas (public scRNA-seq corpora)	No	Learns bidirectional gene–cell embeddings using BERT-style masked modeling for robust cell-type annotation and novel cell discovery	4

Continued on next page

Model (Year)	Published where	Category	Modality	Grounding Type	Agentic (Y/N)	Primary Task	Domains Covered
UCE (Universal Cell Embeddings) (Rosen et al., 2023)	BioRxiv	Foundation	scRNA (cross-species)	Atlas / binary masked self-supervision	No	Embeds any cell zero-shot across species into a shared latent space for clustering, lineage inference, annotation	6
GeneCompass (Yang et al., 2024b)	Nature-/Cell Research	Foundation Model	scRNA / cross-species	Ontology (GRN + co-expression + gene family knowledge)	No	Learns cross-species gene–cell embeddings and supports annotation, perturbation, dose-response, and GRN inference	6
scELMo (Liu et al., 2023)	BioRxiv	Text-Bridge LLM	scRNA + Metadata Text	LLM-generated embeddings from gene/metadata descriptions + raw data	No	Embeds cells via text-derived gene metadata embeddings combined with expression, then supports clustering, batch correction, annotation, perturbation	5
CellPLM (Wen et al., 2023)	ICLR	Spatial (Multi-modal Foundation Model)	scRNA + Spatial (SRT)	Atlas + Spatial relations (uses SRT in pretraining)	No	Treats cells as tokens and tissues as sentences, leveraging spatially-resolved transcriptomics and a Gaussian-mixture prior to encode inter-cell relations for denoising, spatial imputation, and perturbation prediction.	4
scMoFormer (Tang et al., 2023)	ArXiv	Spatial (Multi-modal Foundation Model)	scRNA + Protein / Multi-omics	Atlas + domain knowledge in cross-modality aggregation	No	Uses modality-specific transformers and cross-attention to impute missing modalities, classify cells, and fuse multimodal representations	4

Continued on next page

Model (Year)	Published where	Category	Modality	Grounding Type	Agentic (Y/N)	Primary Task	Domains Covered
scFormer (Xu et al., 2024)	ArXiv	Spatial (Multi-modal Foundation Model) (or Foundation + multi-omics)	Transcript-omics + Proteomics / multi-modal	Atlas + transformer-based fusion	No	Aligns and integrates multi-omics single-cell data, recovers missing modalities, and transfers labels across modalities	4
scMulan (Bian et al., 2024a)	BioRxiv	Foundation Model	scRNA + Metadata	Atlas + prompt-conditioned generative modeling	No	Encodes each cell as a "c-sentence" integrating expression + metadata; supports zero-shot annotation, batch integration, and conditional generation	3
scPRINT (Kalfon et al., 2025)	Nature Commu-nications	Foundation Model	scRNA	Atlas (50M+ cell pretraining)	No	Learns cell embeddings and infers cell-specific gene regulatory networks; supports zero-shot denoising, batch correction, label prediction, expression reconstruction	5
scGraphformer (Fan et al., 2024)	Nature Commu-nications Biology	Foundation Model	scRNA	Atlas + relational prior (kNN bias, refined)	No	Learns a cell–cell graph via a transformer-GNN hybrid for better classification and interaction inference	4
scRDiT (Dong et al., 2025)	ArXiv	Foundation Model	scRNA (transcrip-tome)	Atlas-like generative prior / diffusion modeling	No	Generates synthetic scRNA-seq samples via diffusion transformer + DDIM for accelerated sampling	2
scGFT (Nouri, 2025)	Nature Commu-nications Biology	Foundation Model	scRNA	Fourier-based perturbation / reconstruc-tion (train-free)	No	Synthesizes new single-cell expression profiles by perturbing Fourier components in frequency space.	4

Continued on next page

Model (Year)	Published where	Category	Modality	Grounding Type	Agentic (Y/N)	Primary Task	Domains Covered
scTrans (Lu et al., 2024)	IJCAI	Foundation Model	scRNA	Sub-vector masked completion over gene modules	No	Learns multi-scale sub-vector tokens to perform gene-selective cell-type annotation via masked completion and contrastive regularization	3
scGT (Graph Transformer) (Qi et al., 2025b)	Bioinfo-rmatics Advance	Graph / Multi-omics Integration	scRNA + scATAC	Observed / Hybrid Graph	No	Multi-omics integration + label transfer	4
TransformerST (Lu et al., 2024)	Briefings in Bioinfor-matics	Spatial / Multi-modal FM	Spatial (histology + gene ex-pression)	Spatial (histology + gene expression)	No	Super-resolution gene expression & tissue clustering	3
scGPT-spatial (Cui et al., 2024)	BioRxiv	Spatial (Multi-modal Foundation Model)	Spatial Transcrip-tomics + scRNA prior	Atlas + spatial-aware decoding	No	Extends scGPT via continual pretraining to spatial data, supports multi-slide integration, cell-type deconvolution, and spatial gene imputation	3
HEIST (Madhu et al., 2025)	BioRxiv	Spatial (Multi-modal Foundation Model)	Spatial transcrip-tomics + proteomics	Hierarchical graph modeling (spatial + GRNs)	No	Learns joint embeddings of cells and genes in spatial context to perform cell annotation, gene imputation, spatial clustering, clinical outcome prediction	4
stFormer (Cao et al., 2024)	BioRxiv	Spatial (Multi-modal Foundation Model)	Spatial transcrip-tomics + ligand context	Atlas + biased cross-attention to ligand niche genes	No	Learns gene embeddings contextualized by ligand signals in spatial microenvironments; aids clustering, ligand–receptor inference, and perturbation simulation	4
FmH2ST (Wang et al., 2025)	Nucleic Acids Research	Spatial (Multi-modal Foundation Model)	Histology image + spatial transcrip-tomics	Image foundation + dual graphs + spot branch	No	Predict spatial gene expression from histology using fused image and spot features, supporting denoising, heterogeneity detection, and regulatory inference	3

Continued on next page

Model (Year)	Published where	Category	Modality	Grounding Type	Agentic (Y/N)	Primary Task	Domains Covered
OmiCLIP (Cui et al., 2025)	Nature Methods	Spatial (Multi-modal Foundation Model)	Histology + Spatial Transcrip-tomics	Image–gene contrastive alignment (rank-based)	No	Learns unified visual-omics embeddings linking histopathology and spatial gene expression; enables cross-modal prediction, annotation, and tissue retrieval	7
QuST-LLM (Huang, 2024)	ArXiv	Text-Bridge LLM	Spatial transcrip-tomics + histology metadata	GO-term + gene enrichment + LLM narrative overlay	Yes	Converts ST data and ROIs into human-readable narratives and matches natural language queries to spatial regions via GO/LLM interpretation	4
SpaCCC (Yang et al., 2024a)	IEEE Xplore	Text-Bridge LLM	Spatial transcrip-tomics / transcrip-tomic genes (LRs)	LLM embeddings of ligand + receptor genes	No	Infers spatially resolved cell–cell communication by embedding LR pairs in LLM latent space + diffusion / permutation test filtering	3
spaLLM (Ji et al., 2024)	Briefings in Bioinfor-matics	Spatial (Multi-modal Foundation Model)	Spatial multi-omics (RNA, ATAC, protein)	scGPT embeddings + GNN + attention fusion	No	Enhances spatial domain identification by fusing LLM-derived embeddings and spatial-omics signals via multi-view attention	3
scMMGPT (2025) (Shi et al., 2025)	ArXiv	Text-Bridge LLM / Multi-modal_FM	RNA + Text (metadata)	Annotation, description	No	Generation - links single-cell and text PLMs to describe cells, generate pseudo-cells from text, and enhance annotation through text-conditioned reasoning	6
Cell2Sentence (C2S) (Levine et al., 2024)	PMLR	Text-Bridge LLM	RNA (scRNA-seq converted to "cell sen-tences")	Marker / Implicit (gene rank-order)	No	Generation, annotation & reconstruction - encodes cells as gene-ranked sentences and fine-tunes LLMs to classify or generate biologically meaningful cell text	6

Continued on next page

Model (Year)	Published where	Category	Modality	Grounding Type	Agentic (Y/N)	Primary Task	Domains Covered
C2S-Scale (Rizvi et al., 2025)	BioRxiv	Text-Bridge LLM+ Multi-modal Foundation Model	RNA + Text / Metadata	Atlas + Text / Implicit rank grounding	No	Chat, generation & annotation - trains large LLMs on cell sentences and biological text to enable perturbation prediction and multicellular summarization	7
Cell2Text (Kharouiche et al., 2025)	ArXiv	Text-Bridge LLM + Multi-modal Foundation Model	RNA → Natural Language	Gene-level embeddings + ontology / metadata grounding	No	Expression prediction & regulatory inference - learns regulatory syntax from chromatin accessibility and DNA motifs to predict expression and interpret TF–cis interactions	7
GenePT (Chen and Zou, 2024)	BioRxiv	Foundation Model / Text-Augmented	scRNA	Literature / Text embedding grounding	No	Embedding & downstream prediction - uses GPT-3.5 gene text embeddings to derive cell embeddings via weighted or ranked gene aggregation.	5-7
CellLM (Zhao et al., 2023)	ArXiv	Foundation Model / Represen-tation model	RNA (single cell expres-sion)	Implicit - contrastive embedding from expression data	No	Represent cells via a contrastive-learning transformer, optimize embedding space for tasks like cell-type annotation, drug sensitivity prediction	6
scExtract (Wu and Tang, 2025)	Genome Biology	Agentic Framework (Text-Bridge LLM)	RNA (scRNA-seq) + Text	Article-based parameter extraction	Yes	Annotation & Integration - uses LLMs to extract pipeline parameters from publications and apply them for dataset harmonization.	7

Continued on next page

Model (Year)	Published where	Category	Modality	Grounding Type	Agentic (Y/N)	Primary Task	Domains Covered
scAgent (Mao et al., 2025)	ArXiv	Agentic Framework (Text-Bridge hybrid)	RNA (scRNA-seq)	Reference atlas + marker gene reasoning + memory grounding	Yes	Universal cell annotation & novel cell discovery - scAgent uses an LLM planning module, memory, and tool modules to annotate cells across tissues, detect unknown types, and incrementally learn new annotations	7
EpiFoundation (Wu et al., 2025)	BioRxiv	Epigenomic Foundation Model / Multi-modal alignment	scATAC (chromatin accessibil-ity)	Peak-to-gene supervision (alignment to expression)	No	Cell embedding, annotation, batch correction, gene expression prediction - trains on sparse peak sets, aligns to gene expression supervision, and transfers learned embeddings for downstream ATAC tasks	7
EpiAgent (Chen et al., 2025b)	BioRxiv	Epigenomic Foundation Model / Multi-modal text-bridge hybrid	scATAC / chromatin accessibil-ity	Peak tokenization + external embeddings + regulatory supervision grounding	No	Embedding, annotation, imputation, and perturbation prediction - encodes chromatin accessibility as ranked cCRE tokens, enabling zero-shot cell annotation, peak imputation, and response prediction.	8
ChromFound (Jiao et al., 2025)	Nature	Epigenomic $\rightarrow$ Expression Trans-former / Regulatory FM	scATAC (chromatin accessibil-ity) + DNA sequence	Motif $\times$ peak matrix / masked regulatory grammar grounding	No	Expression prediction & regulatory inference - learns regulatory syntax from chromatin accessibility and DNA motifs to predict gene expression in seen and unseen cell types; also interprets transcription factor interactions and cis-regulatory elements	8

Continued on next page

Model (Year)	Published where	Category	Modality	Grounding Type	Agentic (Y/N)	Primary Task	Domains Covered
GET (General Expression Transformer) (Fu et al., 2025)	Nature	Text-Bridge LLM (Hybrid Fusion Model)	RNA (scRNA) / Text embeddings	Fusion of scGPT embeddings + text-encoded "cell sentences" grounding	No	Annotation (fusion embedding model) - combines scGPT-derived embeddings and LLM (text encoder) embeddings via a small fusion MLP to improve cell-type classification robustness across datasets	6
scMPT (Palayew et al., 2025)	ArXiv	Text-Bridge LLM (Hybrid Fusion Model)	RNA (scRNA) / Text embeddings	Fusion of scGPT embeddings + text-encoded "cell sentences" grounding	No	Annotation (fusion embedding model) - combines scGPT-derived embeddings and LLM (text encoder) embeddings via a small fusion MLP to improve cell-type classification robustness across datasets	6
EpiBERT (Javed et al., 2025)	Cell Genomics	Epigenomic Foundation Model / Multimodal Transformer	DNA sequence + chromatin accessibility	Masked-accessibility pretraining + motif & sequence fusion	No	Accessibility imputation, gene expression prediction & regulatory inference - predicts masked ATAC signals, then fine-tunes to predict expression and enhancer-gene links, generalizing to unseen cell types	8
scMamba (Yuan et al., 2025)	ArXiv	Multimodal / Foundation_FM	Multi-omics (RNA + others)	Implicit via integrated features; no explicit external grounding	No	Multi-omic embedding, integration & annotation across modalities without prior feature selection	7
GeneMamba (Qi et al., 2025a)	ArXiv	Foundation Model / State-Space Model	RNA (scRNA)	Implicit (via gene sequence context + pathway loss)	No	Multi-batch integration, cell-type annotation, gene correlation - uses a BiMamba state-space architecture with pathway-aware losses for scalable, context-rich modeling of scRNA	7

Continued on next page

Model (Year)	Published where	Category	Modality	Grounding Type	Agentic (Y/N)	Primary Task	Domains Covered
Nicheformer (Schaar et al., 2024)	BioRxiv	Foundation_FM/ Spatial / Multi-modal_FM	scRNA + spatial transcrip-tomics	Contextual tokenization + metadata + spatial neighbor-hood embedding	No	Spatial context prediction, spatial label / niche prediction, mapping spatial info to dissociated cells	7
scFormer Cell+ (Cui et al., 2022)	Bioarxiv	Foundation Model / Multi-modal Trans-former	RNA (+ metadata)	Joint gene-cell embedding with metadata tokens	No	Integration & Annotation - context-aware joint embedding for cross-species/tissue generalization	7
scPlantLLM (Cao et al., 2025)	Genomics, Pro-teomics and Bioinfor-matics	Text-Bridge LLM (Founda-tion Model)	RNA (plant scRNA-seq)	Gene token + binned expression embedding (Plant Atlas grounding)	No	Annotation, Integration, GRN inference - pretrains on plant single-cell data with masked LM and cell-type supervision, enabling cross-species annotation, clustering, and regulatory discovery in plant systems	7
LICT (Ye et al., 2024)	ArXiv	Text-Bridge LLM / Annotation LLM Hybrid	RNA (scRNA-seq)	Marker gene–based DE lists + LLM prompting	No	Annotation & Reliability - iterative LLM prompting with DE markers for label refinement and confidence scoring	5
Teddy (family of models) (Chevalier et al., 2025)	ArXiv	Foundation Model / Disease-aware Trans-former	scRNA (single-cell RNA-seq)	Self-supervised + supervised annotation supervision	No	Disease state classification / healthy vs diseased detection - trained to identify disease conditions of held-out donors and distinguish diseased vs healthy cells in new disease contexts	7
Pilot (Joodaki et al., 2024)	Github	Benchmark / Evaluation Frame-work	RNA	Model-agnostic pilot foundation testing	No	Benchmarking & Evaluation - lightweight pilot framework for early single-cell FM testing	4

Continued on next page

Model (Year)	Published where	Category	Modality	Grounding Type	Agentic (Y/N)	Primary Task	Domains Covered
CellVerse (Zhang et al., 2025b)	ArXiv	Benchmark / Evaluation	Multi-omics (RNA, CITE, ASAP, etc.)	Implicit via prompt encoding	No	QA benchmark for annotation, drug response, perturbation tasks	5
xTrimoGene (Gong et al., 2023)	NeurIPS / ArXiv	Foundation Model (Scalable Trans-former)	scRNA-seq	Sparse masking + auto-discretization	No	Representation learning + annotation, perturbation, drug synergy prediction	7
EpiAttend (Li et al., 2022)	NeurIPS 2022 Work-shop	Regulatory / Sequence-Epigenome Trans-former	DNA sequence + single-cell epige-nomic data	Sequence + cell-specific epigenome grounding	No	Predict cell type–specific gene expression by integrating DNA sequence and single-cell epigenomic tracks, linking enhancers and promoters	6
Spatial2-Sentence (Chen et al., 2025a)	ArXiv	Spatial / Text-Bridge hybrid	Spatial + expression (Imaging Mass Cy-tometry)	Spatial adjacency + expression similarity tokenization	No	Encode spatial & expression context into multi-sentence prompts for LLMs to perform cell-type classification and clinical status prediction	6
ChatCell (Fang et al., 2024)	ArXiv	Text-Bridge LLM (In-structional)	scRNA → "cell sentence"	Vocabulary adaptation + unified sequence generation of cell sentences	No	Natural language interface for single-cell tasks - allows users to query, annotate, generate, and explore scRNA data via text prompts. Hugging Face	6
CellAtria (Nouri et al., 2025)	BioRxiv	Agentic Frame-work	scRNA-seq + metadata	Ontology (graph + metadata)	Yes	Annotation & Ontology Mapping	7
CellAgent (Xiao et al., 2024)	ArXiv	Agentic Frame-work	scRNA-seq	None	Yes	Annotation & Ontology Mapping	7

I Appendix Figure

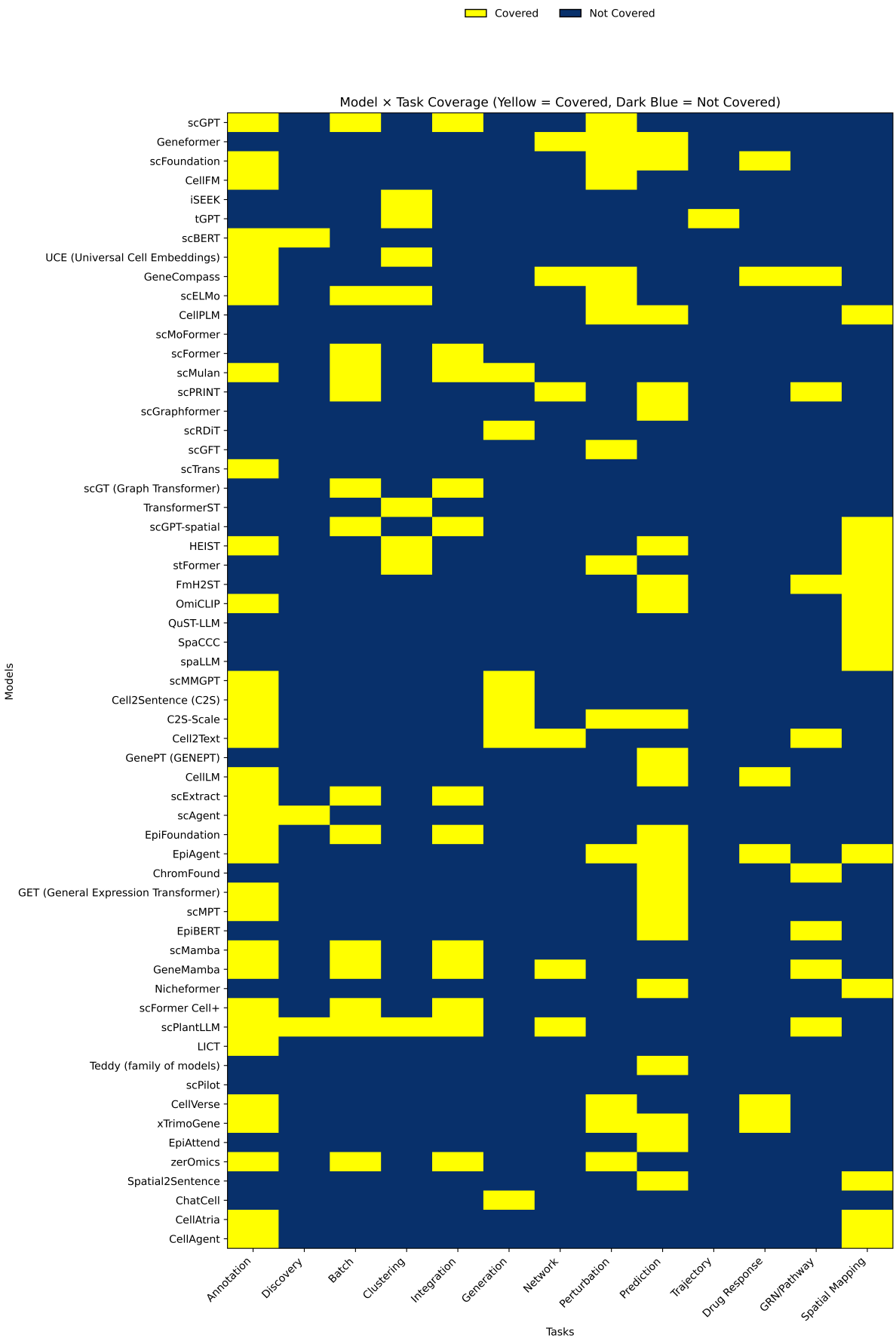

Figure 3: Task vs Model Heatmap. Underlying annotations are provided in Supplementary Tables S2.

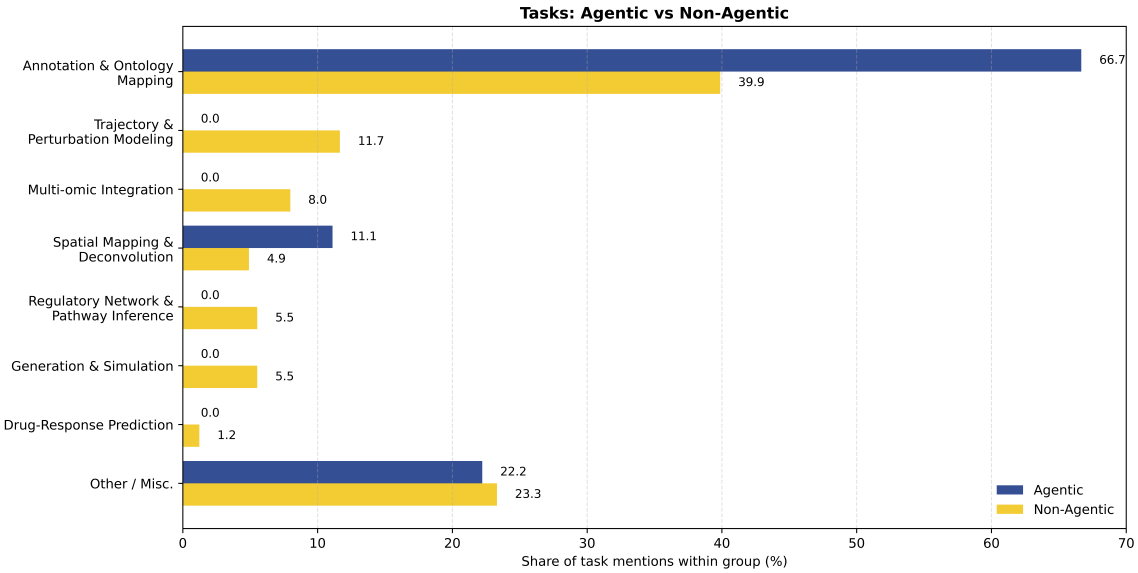

Figure 4: Comparison of task coverage between **agentic** (dark blue) and **non-agentic** (yellow) models. Agentic frameworks emphasize annotation, ontology mapping, spatial mapping while non-agentic models concentrate on trajectory, perturbation modeling, regulatory and pathway inference as well.

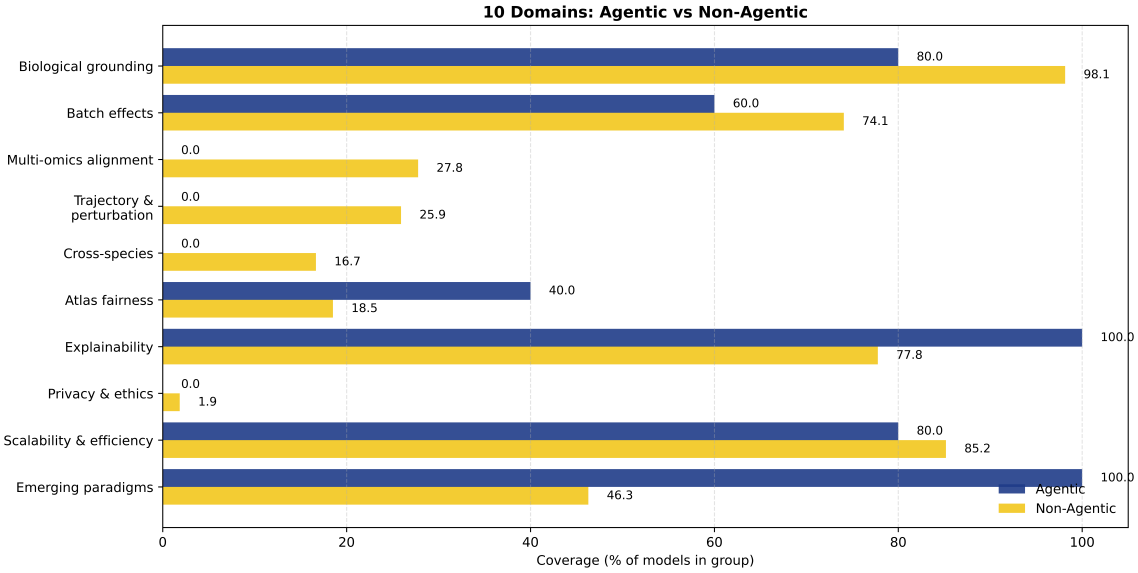

Figure 5: Comparison of domain coverage between **agentic** (dark blue) and **non-agentic** (yellow) models. Agentic frameworks emphasize explainability, fairness, and emerging paradigms, while non-agentic models concentrate on biological grounding and batch effects.