# BEYOND STATIC BENCHMARKS: LIVING EVALUATION FOR SINGLE-CELL FOUNDATION MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Single-cell datasets and sequencing technologies are evolving rapidly, yet the evaluation of single-cell foundation models (scFMs) still relies largely on fixed benchmarks that may overlap with expanding pretraining corpora. We introduce SCLIVEEVAL, a living, contamination-aware framework that continuously incorporates newly released datasets and updates scFM evaluation over time while screening for potential data exposure using available temporal and provenance evidence. Study-defined novel cell populations are used to evaluate generalization to emerging biology; comparisons across platform-specific cohorts assess whether relative model rankings changes across measurement contexts; and, as newly annotated datasets enter the benchmark, human–model annotation disagreement analysis examines divergence between model predictions and supplied labels, highlighting populations that require further review and informing model applicability. We observe clear context dependence in model behavior: novel-population detectability varies across datasets, while biological preservation, batch integration, and model rankings differ across platform-specific cohorts. Cross-model consensus also reveals recurrent annotation disagreements supported by donor-specific transcriptional evidence. These findings motivate a shift from static benchmarking toward living evaluation as single-cell data and technologies continue to evolve.

## 1 INTRODUCTION

Single-cell RNA sequencing (scRNA-seq) has expanded from studies of individual tissues and cell populations into large-scale atlas-building across tissues, species, disease states, and perturbation conditions. This expansion has driven the development of a growing number of single-cell foundation models (scFMs). Pretrained on millions of cellular expression profiles, these models learn transferable representations for downstream biological analysis (Baek et al., 2025). Systematic benchmarks evaluate scFM generalization and downstream utility across zero-shot representation quality, cell-type annotation, batch integration, gene-level tasks, and multi-omics integration (Kedzierska et al., 2025; Wu et al., 2025; Wang et al., 2026b; Liu et al., 2026). While these benchmarks standardize model comparison using fixed datasets and annotations, recent studies indicate that pretrained scFMs exhibit varying generalization across different tasks and zero-shot settings (Kedzierska et al., 2025; Boiarsky et al., 2024). Ongoing assessment is therefore important for establishing model applicability and utility as new datasets and biological contexts become available.

However, such ongoing assessment is complicated by the increasing difficulty of determining whether evaluation data were included in scFM pretraining, as pretraining corpora continue to expand. This concern has been documented in scFM benchmarking: Kedzierska et al. (2025) identified partial overlap between their Pancreas evaluation dataset and Geneformer pretraining data, and reported that their Tabula Sapiens and Immune evaluation datasets were included in scGPT pretraining. Such overlap complicates the interpretation of generalization. Cross-repository redundancy further complicates exposure assessment. For example, the Pan-human Azimuth study combined data from CELLxGENE, HuBMAP, GTEx, Azimuth, and DISCO and used an explicit duplicate-removal procedure because published datasets may be represented across multiple repositories (Sarkar et al., 2026). These findings motivate contamination-aware evaluation that considers both dataset release dates and source information.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

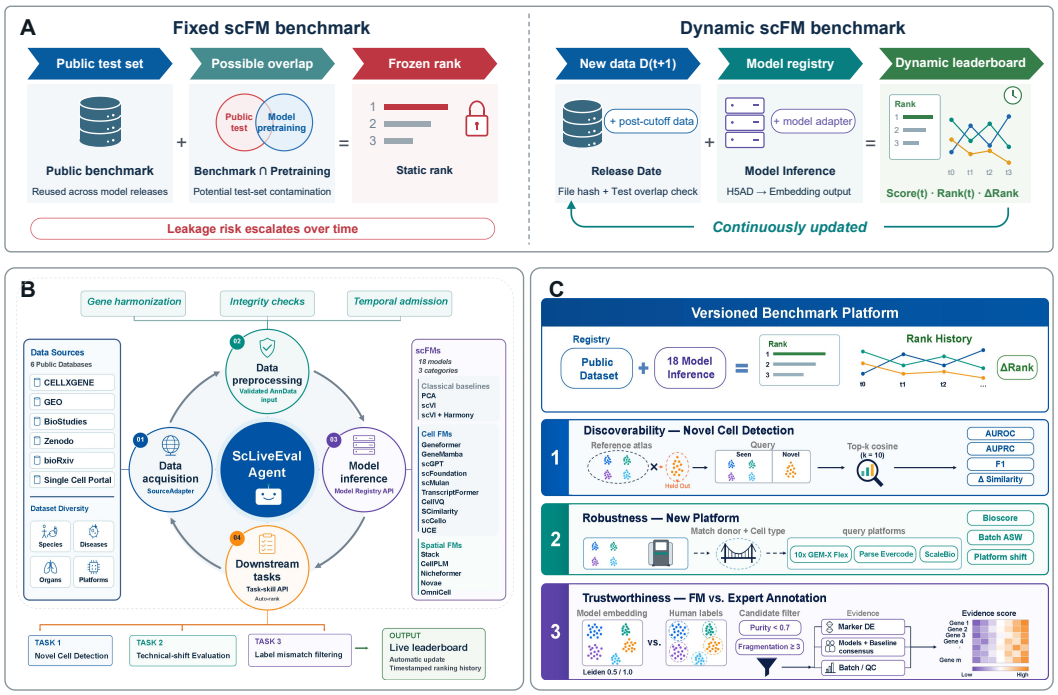


Figure 1: **Overview of SCLIVEEVAL.** (A) Static benchmarks repeatedly reuse public test sets and report fixed rankings, whereas SCLIVEEVAL incorporates post-cutoff data and tracks changes in model performance and rankings across successive benchmark versions. (B) The modular, agent-driven architecture of SCLIVEEVAL standardizes and orchestrates data acquisition, temporal admission checks, model inference, and downstream evaluation. (C) SCLIVEEVAL provides a dynamic, extensible platform for downstream benchmarking, supporting the incorporation of new datasets, models, and tasks over time.

Temporal controls already offer a practical response to this problem. In single-cell benchmarking, Wu et al. (2025) used AIDA v2, released after the evaluated models, to mitigate potential data leakage. In broader machine learning, LiveBench (White et al., 2025) constructs evaluation questions from recent information, while LiveCodeBench (Jain et al., 2025) continuously incorporates newly released programming problems. These precedents motivate extending temporal controls from individual dataset selections to continuously updated, versioned scFM evaluation. For scFMs, newly released datasets add more than test samples: they introduce biological contexts, sequencing technologies, and analytical needs that existing benchmarks may not cover. Living evaluation must therefore address both potential prior exposure and the relevance of evaluation settings to changing applications.

Newly generated datasets across diverse tissues, disease contexts, developmental stages, and perturbation conditions may reveal cell populations or states that are absent from existing reference collections (Varesi et al., 2026; Karakasheva et al., 2026; Wang et al., 2026a). Advances in sequencing technologies also introduce new platforms and protocols with differences in capture efficiency, library complexity, sparsity, and other measurement properties that can substantially affect downstream single-cell analyses (Mereu et al., 2020). These biological and measurement shifts require evaluation across newly released datasets and platforms. New datasets also raise practical questions about when to use scFMs for cell-type annotation and how to combine model outputs with expert judgment. For each dataset, researchers need to identify cell populations consistently supported by model outputs and annotations requiring further review.

To address these needs, we introduce SCLIVEEVAL, a living, contamination-aware evaluation framework for scFMs (Figure 1). Its agent-orchestrated workflow continuously incorporates newly released datasets, supports updates to evaluation tasks, and preserves reproducible benchmark versions. Dataset release dates are compared with model pretraining cutoffs to screen for potential

prior exposure to evaluation data. The framework performs new population detection on newly incorporated datasets and reassesses model performance as new sequencing platforms and generations become available. For annotation on new datasets, human–model disagreement analysis helps assess model applicability and identify populations requiring manual review.

Our key contributions are summarized as follows:

- We introduce SCLIVEEVAL, a living evaluation framework that combines continuous data and model updates with contamination-risk screening and reproducible benchmarking.
- We demonstrate that novel-cell detection capabilities depend strongly on the target population, supporting population-specific reassessment as new datasets enter the benchmark.
- We characterize how model rankings and metric profiles differ across platform-specific cohorts and matched platform-generation comparisons. Evaluating biological structure preservation and batch mixing in these contexts provides evidence for context-dependent model selection.
- We demonstrate the utility of human–model annotation disagreement analysis for newly released datasets. Cross-model consensus reveals recurrent annotation differences, providing evidence for model-assisted annotation review and biological interpretation.

## 2 RELATED WORK

**Single-cell foundation models.** We organize the evaluated scFMs into two broad families according to their modeling scope: Cell FMs and Spatial FMs. Cell FMs learn transferable cell- and gene-level representations from transcriptomic profiles. Models such as Geneformer (Theodoris et al., 2023), scGPT (Cui et al., 2024), and scFoundation (Hao et al., 2024) learn general representations from large-scale expression data, while SCimilarity (Heimberg et al., 2025), UCE (Rosen et al., 2026), and TranscriptFormer (Pearce et al., 2026) incorporate biological structure to support transfer across cell types, tissues, or species. Spatial FMs, including CellPLM (Wen et al., 2024), Nicheformer (Tejada-Lapuerta et al., 2025), and Novae (Blampey et al., 2025), further incorporate cellular neighborhoods and spatial tissue organization.

**Existing scFM benchmarks.** Kedzierska et al. (2025) compared zero-shot representations of Geneformer and scGPT with classical baselines across five datasets. Subsequent benchmarks broadened evaluation: scEval assessed multiple representation models across eight downstream tasks and 29 datasets (Liu et al., 2026), while Wu et al. (2025) evaluated six scFMs on biologically relevant gene- and cell-level tasks. SCMBench further extended benchmarking to single-cell multi-omics integration (Wang et al., 2026b). These studies establish standardized and reproducible comparisons, but their conclusions are based on fixed checkpoint–dataset combinations. As new datasets, biological populations, and sequencing technologies continue to emerge, this static design motivates evaluation frameworks that can be updated over time.

**Contamination-aware evaluation.** Training–evaluation overlap has been documented in scFM benchmarking (Kedzierska et al., 2025), and datasets released after the evaluated models have been used to mitigate leakage risk (Wu et al., 2025). In broader machine learning, LiveBench (White et al., 2025) and LiveCodeBench (Jain et al., 2025) extend temporal controls through repeated evaluation-data updates. In single-cell research, cross-repository duplication and incomplete pretraining disclosures complicate direct overlap verification. SCLIVEEVAL builds on these precedents by combining temporal screening, source tracking, available pretraining-cutoff information, and reproducible evaluation snapshots.

## 3 METHOD

Figure 1B illustrates SCLIVEEVAL, an agent-orchestrated framework for evaluating scFMs under evolving biological and technical conditions. As new datasets are incorporated, scLiveEval addresses three evaluation settings: biological shifts introduced by newly characterized cell populations, technical shifts arising from new sequencing platforms and platform generations, and annotation uncertainty highlighted by disagreement between supplied labels and model predictions.

### 3.1 AGENT-ORCHESTRATED EVALUATION WORKFLOW

**Data acquisition and standardization.** Source-specific connectors retrieve datasets and metadata from public resources, including CELLxGENE (CZI Cell Science Program et al., 2025), GEO (Barrett et al., 2013), BioStudies (Sarkans et al., 2018), Zenodo (European Organization for Nuclear Research & OpenAIRE, 2013), and the Single Cell Portal (Tarhan et al., 2023). Format adapters convert heterogeneous inputs to a common AnnData representation, harmonize gene identifiers, resolve duplicated features, and retain biological and technical annotations required for downstream evaluation.

**Contamination screening.** For each dataset $d$, SCLIVEEVAL records its accession, first available date $\tau_d$, source identifiers $\mathcal{I}_d$, processed-file version, and checksum $h_d$. For model $m$ with pretraining cutoff $c_m$, potential exposure is summarized as $r(d,m) = \max\{\mathbf{1}[\tau_d \le c_m], \mathbf{1}[\mathcal{I}_d \cap \mathcal{I}_m^{\text{pre}} \neq \varnothing], \mathbf{1}[h_d \in \mathcal{H}_m]\}$. Pairs are classified as *Low risk* when the dataset postdates the cutoff with no recorded identifier or hash overlap, *Flagged* when temporal, identifier, or hash evidence indicates possible exposure, and *Unknown* when the available records are insufficient for classification. The primary audit covered 11 dataset versions and 15 pretrained models (165 pairs). These classifications are retained alongside the evaluation results and can be updated as additional provenance and training records become available.

**Standardized model inference.** Each scFM is implemented through a model adapter that preserves its native normalization, gene-vocabulary mapping, tokenization, input requirements, and checkpoint configuration. For dataset $D_d$ and model $m$, the resulting representation is $Z_{m,d} = f_m(T_m(S(D_d))) \in \mathbb{R}^{n_d \times p_m}$, where $S$ denotes dataset standardization, $T_m$ model-specific preprocessing, and $f_m$ the pretrained encoder. PCA, scVI, and scVI with Harmony are exposed through the same evaluation interface and fitted according to the corresponding data split. Each run records the dataset version, model checkpoint, adapter version, software environment, split definition, random seed, and metric configuration in a common result schema.

### 3.2 EVALUATION TASKS

#### 3.2.1 BIOLOGICAL SHIFT: NOVEL-CELL DETECTION

For a labeled cohort $D = \{(x_i, y_i)\}$, let $U$ denote the held-out populations. Study-defined targets follow the original study annotations. Frequency-defined targets satisfy $a_D(c) = n_D(c)/|D| < 0.05$ and are used only as an operational low-abundance criterion. We denote the three frequency-defined analysis cohorts as LA-1, LA-2, and LA-3. We construct disjoint sets with $\mathcal{R}_U = \{i : y_i \notin U,\ i \notin Q_{\text{seen}}\}$ and $Q = Q_{\text{seen}} \cup Q_{\text{novel}}$, where $Q_{\text{novel}} = \{i : y_i \in U\}$ and $Q_{\text{seen}}$ is sampled from the remaining populations. Every sampled seen-query cell is removed from $\mathcal{R}_U$, ensuring $\mathcal{R}_U \cap Q = \varnothing$. Final split counts and cells omitted by legacy balancing manifests are reported in Supplementary Table S12. For LA-1, LA-2, and LA-3, respectively, $(|\mathcal{R}_U|, |Q_{\text{seen}}|, |Q_{\text{novel}}|) = (4{,}806, 356, 356)$, $(5{,}109, 79, 79)$, and $(2{,}604, 39, 39)$, with 844, 1,198, and 612 unused cells.

For query cell $i$, the novelty score is $s_i^{(m)} = 1 - \frac{1}{k}\sum_{j \in \mathcal{N}_k^{(m)}(i;\mathcal{R}_U)} \cos(z_i^{(m)}, z_j^{(m)})$, with $k = 10$. Higher scores indicate weaker support from the seen reference population. We report threshold-free AUROC, novel-positive AUPRC, and the mean seen–novel score difference. F1 is the maximum over a prespecified percentile grid and is treated as descriptive. Because the query is balanced at 50% novel cells, AUPRC and F1 are conditional on this evaluation prevalence.

#### 3.2.2 TECHNICAL SHIFT: SEQUENCING-PLATFORM EVALUATION

We evaluate whether relative model rankings are consistent across platform-specific cohorts using six broad cell types $\mathcal{C}$ shared across the evaluated datasets. For each platform $p$ and random seed $q$, we sample 150 cells from each cell type, yielding $D_p^{(q)} = \bigcup_{c \in \mathcal{C}} \text{Sample}(\mathcal{I}_{p,c}, 150; q)$, for $q = 1, \ldots, 20$, where $\mathcal{I}_{p,c}$ denotes cells of type $c$ on platform $p$. The same sampled cells are used across all models within each platform–seed pair.

Platform performance is evaluated using the scIB metric suite (Luecken et al., 2022). Biological preservation is summarized as $\text{BioScore}_{m,p} = (\text{NMI}_{m,p} + \text{ARI}_{m,p} + \text{ASW}_{m,p}^{\text{cell}} + \text{GC}_{m,p})/4$, with

higher values indicating stronger preservation of cell-type structure. Batch ASW and iLISI are reported separately for batch integration. Metrics are computed for each replicate, and model rankings are summarized across the 20 random seeds. For successive platform generations, Spearman rank correlation is used to quantify the consistency of model rankings across generations.

#### 3.2.3 HUMAN–MODEL ANNOTATION DISAGREEMENT ANALYSIS

We assess whether scFM representations can prioritize supplied annotations for review by comparing human and model annotation. For model $m$, embeddings are normalized as $\mathbf{z}_i^{(m)} = \mathbf{e}_i^{(m)} / \max(\|\mathbf{e}_i^{(m)}\|_2, 10^{-12})$, with cosine distance $d_m(i,j) = 1 - (\mathbf{z}_i^{(m)})^\top \mathbf{z}_j^{(m)}$. Let $\delta_i$ denote the donor of query $i$, let $R_{-\delta_i}$ contain eligible cells from the other donor, and let $\mathcal{N}_{15}^{(m)}(i; R_{-\delta_i})$ be the 15 nearest reference neighbors. Each model predicts when the candidate receives at least eight same votes; cells below this threshold are recorded as abstentions. Cells whose consensus differs from the supplied label are grouped by supplied-label $\rightarrow$ FM-label direction; directions with at least 20 cells in total and representation in both donors enter evidence review.

For each annotation disagreement direction, raw expression counts are aggregated by donor for disagreement candidate cells and the reference cells with the same supplied label. To identify signature genes, paired $t$-tests across donors are applied to $\log 1p(\mathrm{CPM})$ profiles. Candidate-upregulated genes supported in both donors are ranked by mean $\log_2$ fold-change, and the top 20 genes are compared with reference marker gene panel and literature: if they overlap more heavily with the FM-predicted cell type's markers, the case is resolved as *FM supported*; if they align better with the original human label's markers, it is resolved as *human supported*; otherwise, equal or inconclusive evidence leaves it *unresolved*. Additionally, potential technical artifacts—such as low-quality cells or batch effects—are flagged. Finally, visual tools including UMAP embeddings and,gene-programme module scores are incorporated for manual validation.

## 4 EXPERIMENTAL SETUP

**Models.** We evaluated 18 representations, including 15 pretrained scFMs and three data-fitted baselines. The pretrained models span ten Cell FMs and five Spatial FMs, covering diverse model architectures and pretraining strategies. PCA, scVI, and scVI+Harmony served as data-fitted baselines. All pretrained models were evaluated zero-shot using their released checkpoints and model-specific input pipelines. Model specifications, checkpoints, and available pretraining-cutoff information are provided in the Supplementary Material.

**Novel-cell datasets.** We collected human single-cell transcriptomic datasets from multiple public repositories, including recent studies published in 2026. Three studies spanning hematopoietic, intestinal, and placental tissues reported newly characterized cell populations (Varesi et al., 2026; Karakasheva et al., 2026; Wang et al., 2026a). To complement these study-defined novel populations, we introduced a low-abundance setting for cell populations that are sparsely represented in newly incorporated datasets. We therefore included three additional datasets from salivary-gland, meningeal, and craniofacial region tissues (Madissoon et al., 2023; Vinsland et al., 2025; Khouri-Farah et al., 2026). In these datasets, populations accounting for less than 5% of all cells were selected as targets, providing a consistent criterion for evaluating whether scFM representations can distinguish sparsely represented populations from the reference cells. The evaluated populations and settings are summarized in Supplementary Table S1.

**Retrieval and platform evaluation.** Novel-cell detection used cosine-based retrieval from seen-only reference pools with $k = 10$. Performance was measured by AUROC, novel-positive AUPRC, threshold-optimized F1, and the seen–novel similarity gap. Platform evaluation assessed biological structure preservation using BioScore on 10x GEM-X Flex v2, Parse Evercode WT v4, and ScaleBio QuantumScale, with repeated matched subsampling to control cell-type composition. We further examined ranking stability across matched Flex and ScaleBio generations. Sampling configurations and cohort sizes are provided in Supplementary Table S2.

**Human–model disagreement analysis.** We evaluated 15 foundation models on 14,015 eligible cells from the two-donor, 15,802-cell Parse Evercode V4 balanced PBMC dataset. Leave-one-

donor-out cosine 15-nearest-neighbor transfer required eight neighbor votes per model and 8 of 15 model votes for consensus. Directions containing at least 20 cells in total and represented in both donors were evaluated using donor-matched pseudobulk expression, technical covariates, UMAP, and gene-programme scores. Complete criteria and evidence sources are provided in the Supplementary Table S5, S6, and S8.

# 5 RESULTS

## 5.1 RAPID GROWTH OF SINGLE-CELL DATA MOTIVATES LIVING EVALUATION

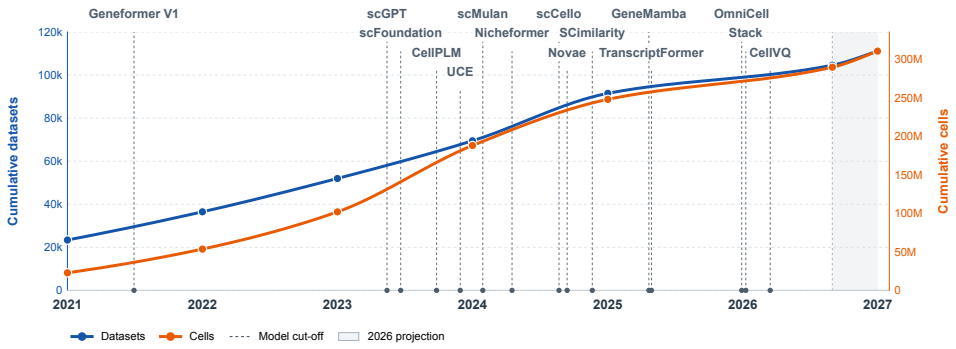


Figure 2: **Dataset and cell growth over time.** Estimated cumulative dataset and cell counts from five surveyed public resources (CELLxGENE, GEO, BioStudies, Zenodo, and the Single Cell Portal), shown alongside reported model pretraining cutoff dates. Counts were extracted through 31 August 2026; cross-source records were deduplicated using explicit accession/DOI matches and obvious duplicate records. The 2026 full-year endpoint is a linear 243/365-day annualization and is marked as projected.

The surveyed resources show rapid growth in publicly available single-cell data (Fig. 2). By 31 August 2026, the cross-source discovery series contained 104,610 deduplicated dataset records, while the date-resolved CELLxGENE series contained 289,373,106 cells. Linear annualization of the observed January–August rate projects 111,174 records and 310,283,834 cells for the full 2026 year. Together, these estimates reveal an expanding pool of biological and technical contexts for evaluating model checkpoints, including datasets that become available after documented training cutoffs.

## 5.2 BIOLOGICAL SHIFTS REVEAL POPULATION-SPECIFIC GENERALIZATION

For study-defined novel populations, representation performance varied substantially across Bone marrow, Colon, and Placenta (Fig. 3A). GeneMamba achieved the highest AUROC in Bone marrow (0.64), with PCA reaching 0.60, whereas scGPT performed strongest in Colon (0.87). In Placenta, scVI and scVI+Harmony achieved the highest AUROC values. No single representation therefore showed the same relative advantage across all newly characterized populations. These results show that detection of reference-excluded populations is population dependent, and that performance observed for one newly introduced population does not fully characterize a model's ability to represent another.

For low-abundance populations ($< 5\%$ abundance), detectability varied substantially across biological populations (Fig. 3B). In Meninges, the cross-model median AUROC was 0.206 for mechanoreceptor cells but 0.814 for radial glial cells. Their novelty-score distributions also differed, with mechanoreceptor cells remaining more similar to seen cells and radial glial cells showing clearer distinction from the reference population. These results show that low abundance alone does not determine whether a population can be identified as novel in representation space; detectability varies across populations even under the same abundance criterion.

Cross-model agreement on novelty was compared with annotated cell identities in Colon (Fig. 3C). High agreement was concentrated in the annotated novel population, while additional regions an-

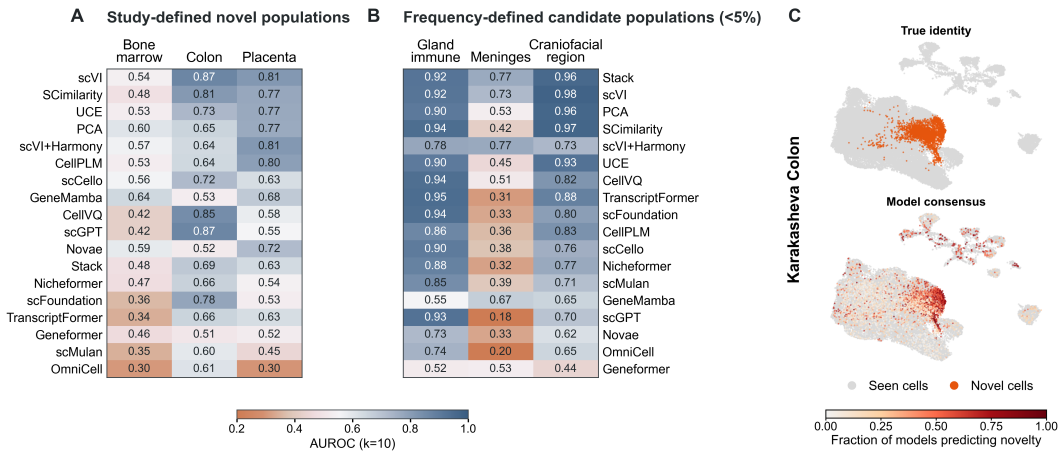


Figure 3: **Novel-cell detection across six datasets.** (A) AUROC for study-defined novel populations in Bone marrow, Colon, and Placenta. (B) AUROC for low-abundance populations ($< 5\%$) in Gland immune, Meninges, and Craniofacial region. Both heatmaps compare 18 representations using $k = 10$ reference neighbors. (C) Colon UMAPs showing annotated seen and novel cells (top) and the fraction of models predicting each cell as novel (bottom). Darker red indicates greater cross-model agreement.

notated as seen cells also showed elevated agreement, particularly in a distinct upper-right cluster. Cross-model consensus therefore captured the major novel population while also highlighting additional cell regions with novel-like representation profiles. These results indicate that novelty signals can extend beyond the originally annotated target population and may help prioritize cells for further biological review.

### 5.3 Model rankings differ across platform-specific cohorts and generations

Evaluation on 10x GEM-X Flex v2, Parse Evercode V4, and ScaleBio QuantumScale reveals distinct model performance profiles in biological preservation and batch integration (Fig. 4A). GeneMamba performs strongly on Flex v2 but occupies a weaker relative position on Parse and QuantumScale, whereas scGPT maintains more consistent biological preservation across the three platforms. Stack shows a comparatively balanced profile, combining strong Batch ASW with competitive biological preservation, while Geneformer achieves strong batch mixing but lower BioScore across platforms. The contrasting profiles show that model rankings and metric profiles differ across platform-specific cohorts, and that biological preservation and batch integration capture different aspects of representation quality. These cohort-specific differences show that model rankings vary across measurement contexts, supporting context-specific evaluation as new cohorts become available.

BioScore rank agreement was lower for Flex v1$\rightarrow$v2 ($\rho = 0.505$) than for ScaleBio v1.1$\rightarrow$QuantumScale ($\rho = 0.838$), indicating different degrees of ranking stability across technology families. On Flex, GeneMamba and OmniCell improved markedly in relative rank, while CellVQ declined; GeneMamba also improved in both BioScore and Batch ASW. On QuantumScale, scCello improved in biological-preservation rank while maintaining a similar Batch ASW position. Model-wise BioScore rank changes were weakly correlated across the two platform families ($\rho = 0.170$, $p = 0.515$), with Nicheformer, for example, improving on Flex but declining on ScaleBio. The distinct responses across platform families indicate that model performance can change differently as sequencing technologies are updated. These generation-specific changes support repeated evaluation as new platform versions become available, allowing model performance to be reassessed under the current measurement technology.

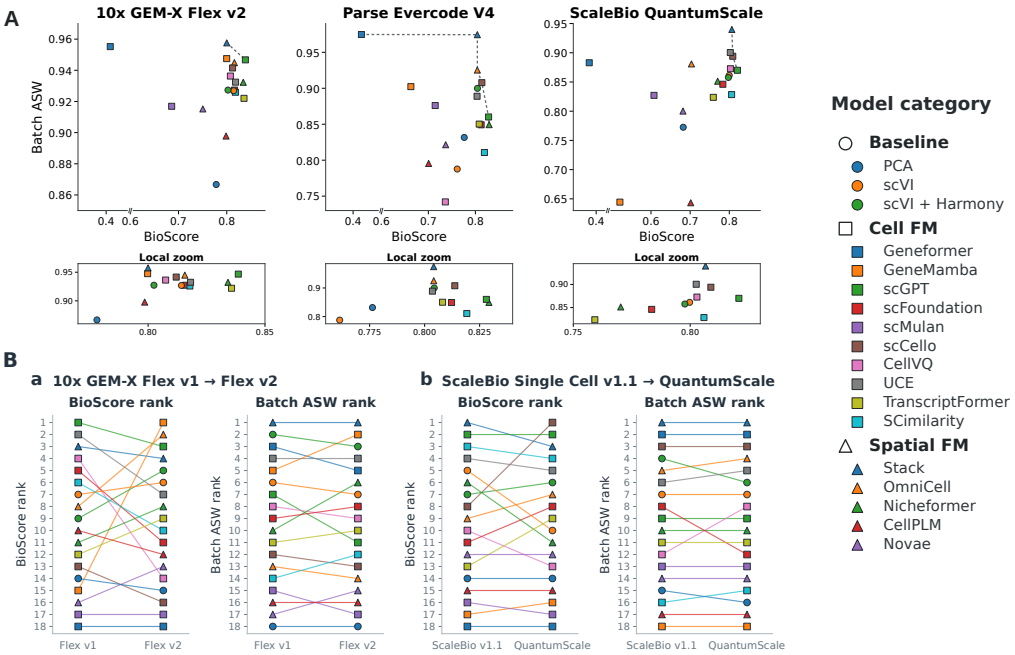


Figure 4: **Model performance across emerging sequencing platforms and generations.** (A) Biological preservation and batch mixing across three newly introduced sequencing platforms. (B) Changes in model rankings across successive platform generations.

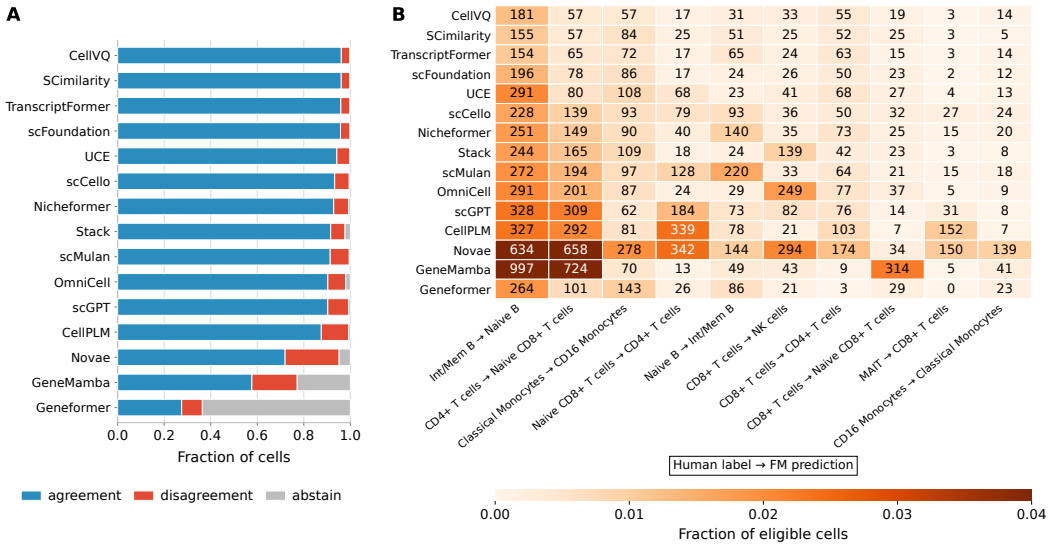


Figure 5: **Foundation-model prediction status and recurrent disagreement directions in Parse Evercode V4.** (A) Stacked bars show agreement, disagreement, and abstention fractions for 15 foundation models across 14,015 eligible cells. (B) Heatmap entries show cell counts for the ten most frequent human-label → FM-prediction directions for each model, with color indicating the fraction of eligible cells on total cells.

### 5.4 Human–model annotation disagreements

Across 15 FMs and 14,015 eligible Parse Evercode V4 cells, agreement with supplied labels ranged from 27.5% to 96.1%, whereas disagreement ranged from 3.8% to 23.2% (Fig. 5A). Geneformer

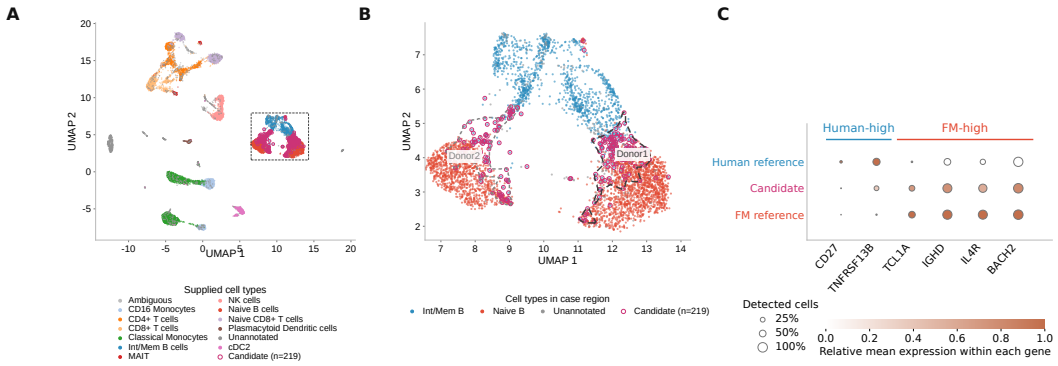


Figure 6: **Case study of the Int/Mem B → Naive B annotation disagreement.** (A) Global UMAP showing cell-type labels and the 219 candidate cells. (B) Donor-resolved UMAP highlighting candidate cells in Donor1 and Donor2. (C) Donor-balanced marker expression comparison across human-reference, candidate, and FM-supported groups.

had the highest abstention rate (63.6%), while Novae had the largest disagreement fraction. Using the selection criteria defined in the Methods, we nominated five directions for review. Donor-aware marker and technical arbitration classified two as FM-supported, one as a technical artifact, and two as unresolved. The FM-supported groups comprised 219 Int/Mem B → Naive B cells and 78 $CD4^+$ T → Naive $CD8^+$ T cells (Fig. 5B).

Detailed analysis of the larger group showed 219 candidates in the B-cell compartment in the full-atlas UMAP (Fig. 6A). The case-region UMAP resolved the same population across Donor1 and Donor2, with 127 and 92 candidates, respectively (Fig. 6B). The donor-balanced marker-expression dot plot supported a Naive B-like state (Fig. 6C): within each gene, relative mean expression was higher in candidates than in Int/Mem B references for the Naive B-high genes *TCL1A*, *IGHD*, *IL4R*, and *BACH2*, whereas the Int/Mem B-high genes *CD27* and *TNFRSF13B* showed lower relative mean expression in candidates, supporting review of these candidates as Naive B. Dot size indicates the fraction of cells in which each gene was detected. Together, donor-resolved localization and marker profiles demonstrate that cross-model consensus from foundation models effectively assists human experts in reviewing and annotating candidates from new datasets. Marker definitions and evidence are provided in Supplementary Table S5 and Supplementary Table S8, and expression summaries are provided in Supplementary Table S9.

## 6 Discussion

scLiveEval shows that evaluation of scFMs inherently dependent on the biological and technical context in which models are applied. Across newly introduced biological populations and sequencing platforms, model performance varied across evaluation settings, with different representations showing distinct strengths in population detection, biological structure preservation, and batch integration. These results highlight the value of maintaining evaluation as an evolving process that follows the expansion of single-cell data and technologies.

**Limitation.** The current evaluation covers publicly available transcriptomic datasets, three platform-specific cohorts, two matched-generation comparisons, and a two-donor annotation-review cohort. Platform cohorts retain differences in donor composition, gene content, batch structure, and sampling, so the platform analysis characterizes context-specific model rankings. Study-defined held-out populations follow source annotations, while contamination screening relies on release dates, provenance records, file hashes, and documented checkpoint cutoffs. These boundaries define the next expansion of scLiveEval through additional modalities, larger cohorts, independent annotation review, and repeated benchmark runs as the single-cell resource landscape expands.

## AI USE STATEMENT

Generative AI tools were used to assist with language editing and translation, including improving grammar, clarity, concision, and academic style across the manuscript. They were also used to support literature search and reference checking, to suggest improvements to manuscript organization and presentation, and to provide feedback on the wording and notation of selected methodological descriptions. All AI-assisted text was reviewed and revised by the authors. Literature and citations suggested through AI assistance were checked against the original sources, and methodological descriptions, equations, numerical results, and scientific interpretations were verified against the underlying analyses and data by the authors. The authors made all final decisions regarding the study design, analyses, interpretation of results, and conclusions, and take full responsibility for the content of the manuscript.

## REFERENCES

Seungbyn Baek, Kyungwoo Song, and Insuk Lee. Single-cell foundation models: bringing artificial intelligence into cell biology. *Experimental & Molecular Medicine*, 57:2169–2181, October 2025. ISSN 2092-6413. doi: 10.1038/s12276-025-01547-5. URL https://doi.org/10.1038/s12276-025-01547-5.

Tanya Barrett, Stephen E. Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F. Kim, Maxim Tomashevsky, Kimberly A. Marshall, Katherine H. Phillippy, Patti M. Sherman, Michelle Holko, Andrey Yefanov, Hyeseung Lee, Naigong Zhang, Cynthia L. Robertson, Nadezhda Serova, Sean Davis, and Alexandra Soboleva. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research*, 41(D1):D991–D995, 2013. doi: 10.1093/nar/gks1193.

Quentin Blampey, Hakim Benkirane, Nadège Bercovici, Kevin Mulder, Grégoire Gessain, Florent Ginhoux, Fabrice André, and Paul-Henry Cournède. Novae: a graph-based foundation model for spatial transcriptomics data. *Nature Methods*, 22(12):2539–2550, December 2025. ISSN 1548-7105. doi: 10.1038/s41592-025-02899-6. URL https://doi.org/10.1038/s41592-025-02899-6.

Rebecca Boiarsky, Nalini M. Singh, Alejandro Buendia, Ava P. Amini, Gad Getz, and David Sontag. Deeper evaluation of a single-cell foundation model. *Nature Machine Intelligence*, 6(12):1443–1446, December 2024. ISSN 2522-5839. doi: 10.1038/s42256-024-00949-w. URL https://doi.org/10.1038/s42256-024-00949-w.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21(8):1470–1480, August 2024. ISSN 1548-7105. doi: 10.1038/s41592-024-02201-0. URL https://doi.org/10.1038/s41592-024-02201-0.

CZI Cell Science Program, Shibla Abdulla, Brian Aevermann, Pedro Assis, et al. CZ CELLxGENE Discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *Nucleic Acids Research*, 53(D1):D886–D900, 2025. doi: 10.1093/nar/gkae1142.

European Organization for Nuclear Research and OpenAIRE. Zenodo, 2013. URL https://www.zenodo.org/.

Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, and Le Song. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, 21(8):1481–1491, August 2024. ISSN 1548-7105. doi: 10.1038/s41592-024-02305-7. URL https://doi.org/10.1038/s41592-024-02305-7.

Graham Heimberg, Tony Kuo, Daryle J. DePianto, Omar Salem, Tobias Heigl, Nathaniel Diamant, Gabriele Scalia, Tommaso Biancalani, Shannon J. Turley, Jason R. Rock, Héctor Corrada Bravo, Josh Kaminker, Jason A. Vander Heiden, and Aviv Regev. A cell atlas foundation model for scalable search of similar human cells. *Nature*, 638:1085–1094, February 2025. ISSN 1476-4687. doi: 10.1038/s41586-024-08411-y. URL https://doi.org/10.1038/s41586-024-08411-y.

Naman Jain, Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. LiveCodeBench: Holistic and Contamination Free Evaluation of Large Language Models for Code. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 58791–58831, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/94074dd5a072d28ff75a76dabed43767-Paper-Conference.pdf.

Tatiana A. Karakasheva, Clara Morral Martinez, Yusen Zhou, et al. An epigenetic basis for sustained inflammatory epithelial progenitor cell states in crohn's disease. *Cellular and Molecular Gastroenterology and Hepatology*, 20(3):101665, 2026. doi: 10.1016/j.jcmgh.2025.101665.

Kasia Z. Kedzierska, Lorin Crawford, Ava P. Amini, and Alex X. Lu. Zero-shot evaluation reveals limitations of single-cell foundation models. *Genome Biology*, 26(1):101, April 2025. ISSN 1474-760X. doi: 10.1186/s13059-025-03574-x. URL https://doi.org/10.1186/s13059-025-03574-x.

Nagham Khouri-Farah, Alexandra Manchel, Emma Wentworth Winchester, Brian M. Schilder, Kelsey Robinson, Sarah W. Curtis, Nathan G. Skene, Elizabeth J. Leslie-Clarkson, and Justin Cotney. Gene expression dynamics of human and mouse craniofacial development at the single-cell level. *Nature Communications*, 17(1):3714, March 2026. ISSN 2041-1723. doi: 10.1038/s41467-026-70232-6. URL https://doi.org/10.1038/s41467-026-70232-6.

Tianyu Liu, Kexing Li, Yuge Wang, Hongyu Li, and Hongyu Zhao. Evaluating the utilities of foundation models in single-cell data analysis. *Advanced Science*, 13(27):e14490, 2026. doi: https://doi.org/10.1002/advs.202514490. URL https://advanced.onlinelibrary.wiley.com/doi/abs/10.1002/advs.202514490.

Malte D. Luecken, M. Büttner, K. Chaichoompu, A. Danese, M. Interlandi, M. F. Mueller, D. C. Strobl, L. Zappia, M. Dugas, M. Colomé-Tatché, and Fabian J. Theis. Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods*, 19(1):41–50, January 2022. ISSN 1548-7105. doi: 10.1038/s41592-021-01336-8. URL https://doi.org/10.1038/s41592-021-01336-8.

Elo Madissoon, Amanda J. Oliver, Vitalii Kleshchevnikov, Anna Wilbrey-Clark, Krzysztof Polanski, Nathan Richoz, Ana Ribeiro Orsi, Lira Mamanova, Liam Bolt, Rasa Elmentaite, J. Patrick Pett, Ni Huang, Chuan Xu, Peng He, Monika Dabrowska, Sophie Pritchard, Liz Tuck, Elena Prigmore, Shani Perera, Andrew Knights, Agnes Oszlanczi, Adam Hunter, Sara F. Vieira, Minal Patel, Rik G. H. Lindeboom, Lia S. Campos, Kazuhiko Matsuo, Takashi Nakayama, Masahiro Yoshida, Kaylee B. Worlock, Marko Z. Nikolić, Nikitas Georgakopoulos, Krishnaa T. Mahbubani, Kourosh Saeb-Parsy, Omer Ali Bayraktar, Menna R. Clatworthy, Oliver Stegle, Natsuhiko Kumasaka, Sarah A. Teichmann, and Kerstin B. Meyer. A spatially resolved atlas of the human lung characterizes a gland-associated immune niche. *Nature Genetics*, 55(1):66–77, January 2023. ISSN 1546-1718. doi: 10.1038/s41588-022-01243-4. URL https://doi.org/10.1038/s41588-022-01243-4.

Elisabetta Mereu, Atefeh Lafzi, Catia Moutinho, Christoph Ziegenhain, Davis J. McCarthy, Adrián Álvarez Varela, Eduard Batlle, Sagar, Dominic Grün, Julia K. Lau, Stéphane C. Boutet, Chad Sanada, Aik Ooi, Robert C. Jones, Kelly Kaihara, Chris Brampton, Yasha Talaga, Yohei Sasagawa, Kaori Tanaka, Tetsutaro Hayashi, Caroline Braeuning, Cornelius Fischer, Sascha Sauer, Timo Trefzer, Christian Conrad, Xian Adiconis, Lan T. Nguyen, Aviv Regev, Joshua Z. Levin, Swati Parekh, Aleksandar Janjic, Lucas E. Wange, Johannes W. Bagnoli, Wolfgang Enard, Marta Gut, Rickard Sandberg, Itoshi Nikaido, Ivo Gut, Oliver Stegle, and Holger Heyn. Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. *Nature Biotechnology*, 38:747–755, June 2020. ISSN 1546-1696. doi: 10.1038/s41587-020-0469-4. URL https://doi.org/10.1038/s41587-020-0469-4.

James D. Pearce, Sara E. Simmonds, Gita Mahmoudabadi, Lakshmi Krishnan, Giovanni Palla, Ana-Maria Istrate, Alexander Tarashansky, Benjamin Nelson, Omar Valenzuela, Donghui Li, Stephen R. Quake, and Theofanis Karaletsos. Transcriptformer: A generative cell atlas across 1.5 billion years of evolution. *Science*, 393(6806):aec8514, 2026. doi: 10.1126/science.aec8514. URL https://www.science.org/doi/abs/10.1126/science.aec8514.

Yanay Rosen, Yusuf Roohani, Ayush Agrawal, Leon Samotorčan, Stephen R. Quake, Jure Leskovec, and Tabula Sapiens Consortium. Universal cell embedding provides a foundation model for cell biology. *Nature*, 656(8126):183–191, August 2026. ISSN 1476-4687. doi: 10.1038/s41586-026-10689-z. URL https://doi.org/10.1038/s41586-026-10689-z.

Ugis Sarkans, Mikhail Gostev, Awais Athar, Ehsan Behrangi, Olga Melnichuk, Ahmed Ali, Jasmine Minguet, Juan Camillo Rada, Catherine Snow, Andrew Tikhonov, Alvis Brazma, and Johanna McEntyre. The BioStudies database—one stop shop for all data supporting a life sciences study. *Nucleic Acids Research*, 46(D1):D1266–D1270, 2018. doi: 10.1093/nar/gkx965.

Sourav Sarkar, Zhuoyan Li, Gesmira Molla, Anagha Shenoy, Brian Zhang, David Collins, Nicole Vasilevsky, Joseph P. Gaut, Aleix Puig-Barbe, Andreas Bueckle, David Osumi-Sutherland, Katy Börner, Sanjay Jain, and Rahul Satija. Organism-scale annotation with Pan-human Azimuth. *bioRxiv*, 2026. doi: 10.64898/2026.07.16.738997. URL https://www.biorxiv.org/content/early/2026/07/21/2026.07.16.738997.

Leyla Tarhan, Jon Bistline, Jean Chang, Bryan Galloway, Emily Hanna, and Eric Weitz. Single cell portal: an interactive home for single-cell genomics data. *bioRxiv*, 2023. doi: 10.1101/2023.07.13.548886.

Alejandro Tejada-Lapuerta, Anna C. Schaar, Robert Gutgesell, Giovanni Palla, Lennard Halle, Mariia Minaeva, Larsen Vornholz, Leander Dony, Francesca Drummer, Till Richter, Mojtaba Bahrami, and Fabian J. Theis. Nicheformer: a foundation model for single-cell and spatial omics. *Nature Methods*, 22(12):2525–2538, December 2025. ISSN 1548-7105. doi: 10.1038/s41592-025-02814-z. URL https://doi.org/10.1038/s41592-025-02814-z.

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 Patrick T. Ellinor. Transfer learning enables predictions in network biology. *Nature*, 618:616–624, June 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06139-9. URL https://doi.org/10.1038/s41586-023-06139-9.

Angelica Varesi, Murtaza S. Nagree, Isabella Di Biasio, Andy G. X. Zeng, Sayyam Shah, Michael Zhang, Hyerin Kim, Alex Murison, Rohail Badami, Olga Gan, Liqing Jin, Jessica McLeod, Sheela A. Abraham, Mark D. Minden, Andrea Arruda, Igor Novitzky-Basso, Jonas Mattsson, John E. Dick, and Stephanie Z. Xie. Atp2b1 expression identifies human hematopoietic stem cells with superior repopulation and self-renewal. *Proceedings of the National Academy of Sciences*, 123(19):e2604380123, 2026. doi: 10.1073/pnas.2604380123. URL https://www.pnas.org/doi/abs/10.1073/pnas.2604380123.

Elin Vinsland, Sergio Marco Salas, Ivana Kapustová, Lijuan Hu, Simone Webb, Xiaofei Li, Xiaoling He, Mats Nilsson, Muzlifah Haniffa, Roger Barker, Oscar Persson, David R. Raleigh, Erik Sundström, Peter Lönnerberg, and Sten Linnarsson. Cell atlas of the developing human meninges reveals a dura origin of meningioma., July 2025. ISSN: 2692-8205 Journal Abbreviation: bioRxiv Pages: 2025.07.08.663122 Publication Title: bioRxiv : the preprint server for biology.

Cheng Wang, Yan Zhou, Yuejun Wang, Tuhin Kumar Guha, Zhida Luo, Anxhela Mustafaraj, Tara I. McIntyre, Marisa E. Schwab, Brittany R. Davidson, Gabriella C. Reeder, Ronald J. Wong, Sarah K. England, Juan M. Gonzalez, Robert Blelloch, Alexis J. Combes, Linda C. Giudice, Adrian Erlebacher, Tippi C. MacKenzie, David K. Stevenson, Gary M. Shaw, Michael P. Snyder, Xiaofei Sun, Virginia D. Winn, Susan J. Fisher, and Jingjing Li. Single-cell spatiotemporal dissection of the human maternal–fetal interface. *Nature*, 653(8113):167–179, May 2026a. ISSN 1476-4687. doi: 10.1038/s41586-026-10316-x. URL https://doi.org/10.1038/s41586-026-10316-x.

Yixuan Wang, Yimin Fan, Xuesong Wang, Tingyang Yu, Yongshuo Zong, Xinyuan Liu, Gaoyang Zhong, Meitong Liu, Qing Li, Kin Hei Lee, Khachatur Dallakyan, Zhichao Hu, Yaqian Qi, Junjie Huang, Gengjie Jia, Jiao Yuan, Ting-Fung Chan, Xin Gao, Irwin King, and Yu Li. SCMBench: benchmarking domain-specific and foundation models for single-cell multi-omics data integration. *Nature Communications*, 17(1):5967, May 2026b. ISSN 2041-1723. doi: 10.1038/s41467-026-72570-x. URL https://doi.org/10.1038/s41467-026-72570-x.

Hongzhi Wen, Wenzhuo Tang, Xinnan Dai, Jiayuan Ding, Wei Jin, Yuying Xie, and Jiliang Tang. CellPLM: Pre-training of Cell Language Model Beyond Single Cells. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 5649–5673, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/173bcdbd820edbe43284e0941109ada9-Paper-Conference.pdf.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Sandha, Siddartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. LiveBench: A Challenging, Contamination-Limited LLM Benchmark. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 91595–91631, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/e4a46394ba5378b3f9a186a5b4c650d1-Paper-Conference.pdf.

Jialu Wu, Qing Ye, Yilin Wang, Renling Hu, Yiheng Zhu, Mingze Yin, Tianyue Wang, Jike Wang, Chang-Yu Hsieh, and Tingjun Hou. Biology-driven insights into the power of single-cell foundation models. *Genome Biology*, 26(1):334, October 2025. ISSN 1474-760X. doi: 10.1186/s13059-025-03781-6. URL https://doi.org/10.1186/s13059-025-03781-6.