SCALING LARGE LANGUAGE MODELS FOR NEXT-GENERATION SINGLE-CELL ANALYSIS

Syed Asad Rizvi[*,†]
Yale University, Google Research
syed.rizvi@yale.edu

Daniel Levine[*]
Yale University
daniel.levine@yale.edu

Aakash Patel[*]
Yale University
aakash.patel.ap2853@yale.edu

Shiyang Zhang[*]
Yale University
shiyang.zhang@yale.edu

Eric Wang[*]
Google DeepMind
ericzwang@google.com

Sizhuang He
Yale University

David Zhang
Yale University

Cerise Tang
Yale University

Zhuoyang Lyu
Brown University

Rayyan Darji
Yale University

Chang Li
Yale University

Emily Sun
Yale University

David Jeong
Yale University

Lawrence Zhao
Yale University

Jennifer Kwan
Yale University

David Braun
Yale University

Brian Hafler
Yale University

Jeffrey Ishizuka
Yale University

Rahul M. Dhodapkar
University of Southern California

Hattie Chung
Yale University

Shekoofeh Azizi
Google DeepMind
shekazizi@google.com

Bryan Perozzi
Google Research
hubris@google.com

David van Dijk[‡]
Yale University
david.vandijk@yale.edu

April 15, 2025

ABSTRACT

Single-cell RNA sequencing has transformed our understanding of cellular diversity, yet current single-cell foundation models (scFMs) remain limited in their scalability, flexibility across diverse tasks, and ability to natively integrate textual information. In this work, we build upon the Cell2Sentence (C2S) framework, which represents scRNA-seq profiles as textual "cell sentences," to train Large Language Models (LLMs) on a corpus comprising over one billion tokens of transcriptomic data, biological text, and metadata. By scaling model size to 27 billion parameters, we observe consistent improvements in predictive and generative capabilities, as well as the capacity for advanced downstream tasks requiring synthesis of information across multicellular contexts. Through targeted fine-tuning supported by modern reinforcement learning techniques, our approach excels in tasks such as perturbation response prediction, natural language interpretation, and complex biological reasoning. By unifying transcriptomic and textual data at unprecedented scales, this approach not only surpasses both specialized single-cell models and general-purpose LLMs, but also establishes a powerful platform for next-generation single-cell analysis, paving the way for the development of "virtual cells."

* = Equal contribution
† = Work partially done during internship at Google Research
‡ = Corresponding author

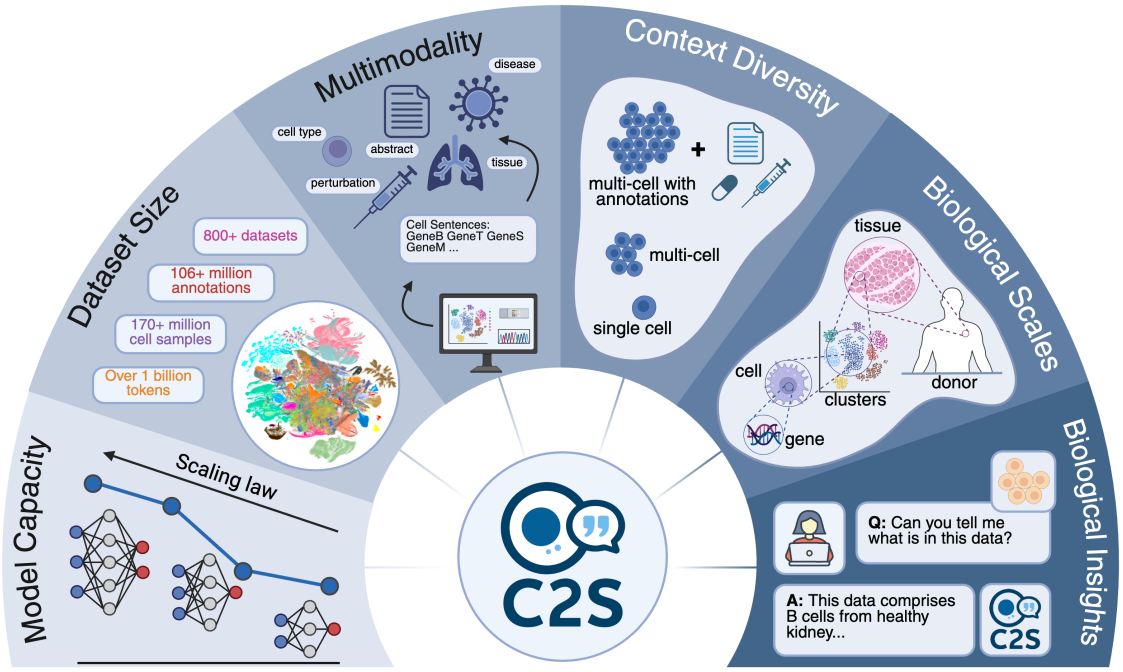

Figure 1: Scaling LLM-based single-cell analysis. A multidimensional expansion of the C2S [14] framework, demonstrating advances in model capacity, dataset size, multimodality, multi-cell support, and integration across biological scales, from single cells to organism-wide insights in natural language. This framework bridges computational innovation with biological discovery, accelerating next-generation single-cell analysis.

1 Introduction

Single-cell RNA sequencing (scRNA-seq) has revolutionized our understanding of cellular heterogeneity by enabling the profiling of gene expression at single-cell resolution [1]. This technology has generated massive data atlases such as CellxGene [2] and the Human Cell Atlas [3], offering unparalleled opportunities for computational methods to extract biological insights from this data. Recent transcriptomic foundation models (FMs), such as scGPT [4], Geneformer [5], scFoundation [6], and scGenePT [7] have shown promise in modeling single-cell transcriptomic data at scale. Despite these advances, current models are often constrained by bespoke architectures, hindering their scalability to larger model sizes, integration of different data modalities, and ability to perform diverse generative and predictive tasks. These limitations restrict the ability of expression-only foundation models to synthesize insights across datasets, modalities, and biological contexts, and highlight the need for an alternative approach capable of addressing these challenges while maintaining flexibility and scalability.

Large Language Models (LLMs) [8, 9, 10], offer a promising solution to these challenges. Widely used in Natural Language Processing (NLP), LLMs exhibit robust scaling behavior in performance over diverse downstream tasks [11, 12]. Their ability to process vast text corpora and generalize effectively to new applications makes them well-suited for addressing the limitations of current expression-only models. Cell2Sentence (C2S) [13, 14] leverages the capabilities of LLMs through *data engineering*, transforming high-dimensional single-cell data into a textual format compatible with these models. By converting scRNA-seq profiles into "cell sentences" — sequences of gene names ordered by expression level — C2S positions single-cell data within the LLM framework, providing better scalability and infrastructure advantages than specialized model architectures. This data transformation strategy simplifies model development and deployment, and enables easy integration of transcriptomic data with diverse modalities, including metadata, experimental conditions, and textual descriptions from biological publications.

Here, we introduce the next generation of C2S models, **C2S-Scale**, which significantly improves the C2S paradigm in terms of: (a) model capacity, (b) model performance, (c) dataset size and multimodality, (d) context length and diversity, and (e) downstream applications, highlighted in Figure 1. The C2S-Scale model family establishes scaling laws in single-cell analysis and represents a significant step towards next-generation, language-powered tools for biological discovery, paving the way for virtual cell platforms that integrate transcriptomic data, natural language, and contextual information.

Our contributions can be summarized as follows:

1. Scaling Single-Cell Analysis with LLMs: We introduce C2S-Scale, a new family of LLMs designed to robustly scale single-cell analysis across multiple axes:

(a) **Larger Model Capacity:** C2S-Scale comprises models ranging from 410 million to 27 billion parameters (410M, 1B, 2B, 9B, and 27B), based on the Gemma-2 [15] and Pythia [16] LLM architectures. This represents a substantial increase in model capacity compared to existing single-cell foundation models, enabling the capture of more complex relationships within the data.

(b) **Increased Performance at Scale:** We establish performance scaling laws for LLMs in single-cell analysis, demonstrating significant improvements in both predictive and generative tasks with increasing model size from 410 million to 27 billion parameters. Evaluation on held-out test sets demonstrates improved generalization across diverse single-cell tasks with larger models. These scaling trends are observed in both full fine-tuning and parameter-efficient regimes, highlighting the practical utility of scaling even with limited computational resources.

(c) **Data Size and Multimodality:** C2S-Scale models are trained on a massive, 1-billion token multimodal corpus encompassing more than 50 million human and mouse cells with associated metadata and annotations, collected from publicly available single-cell atlases such as the Human Cell Atlas [3] and CellxGene [2]. By training on both transcriptomic data along with corresponding biological text (e.g. paper abstracts), C2S aligns single-cell transcriptomic data with natural language and biological context. This corpus is formatted into a set of 150 million multi-task training samples (detailed in Table 1), allowing the LLM to learn diverse tasks while simultaneously integrating annotations and free-text information.

(d) **Long-Context, Multi-Cell Capabilities:** C2S-Scale models support extended context lengths up to 8192 tokens, enabling more comprehensive multimodal and multi-cell input. Importantly, C2S-Scale can process and generate data for multiple cells simultaneously, enabling analysis of cellular interactions and complex biological processes. The extended context also allows for the integration of diverse contextual information, including biological annotations, manuscript text, perturbation conditions, and more detailed task-specific instructions.

(e) **Diverse Downstream Applications:** C2S-Scale models are fine-tuned and evaluated on a significantly broader range of downstream tasks than previous models, encompassing challenging biological reasoning tasks such as perturbation prediction, nuanced natural language interpretation of single-cell data, and complex question answering, showcasing the versatility and applicability of our approach.

2. Reinforcement Learning for Enhanced Performance: Inspired by the use of reinforcement learning in NLP to align LLMs with user preferences, we leverage Group Relative Policy Optimization (GRPO) [17] to further refine C2S for targeted single-cell tasks. We quantify the performance improvements achieved with GRPO on challenging question answering benchmarks as well as perturbation response prediction.

3. A Novel Metric for Evaluating Single-Cell Generative Models: We introduce the **single-cell Fréchet Inception Distance (scFID)**, an adaptation of the widely used Fréchet Inception Distance (FID) for evaluating image generative models. Unlike expression-level metrics, which can be dominated by high-dimensional noise and outlier genes, scFID leverages a single-cell foundation model embedding space to assess the quality of generated cells in a biologically meaningful way.

4. Open-Source Models and Resources: We release our code and model weights to the community to facilitate broader adoption and further development of LLM-based single-cell analysis. This includes resources for constructing transcriptomic-language integrated datasets and prompts for LLM-based analysis.

2 Results

In this section, we demonstrate the broad capabilities of C2S-Scale on a diverse range of single-cell tasks, highlighting the benefits of scaling LLM-based single-cell analysis. First, we evaluate trained C2S-Scale models ranging from 410 million to 27 billion parameters, demonstrating scaling laws in performance across predictive and generative tasks. We then present key results for natural language interpretation of scRNA-seq data, spatial reasoning, question answering, and perturbation response prediction tasks.

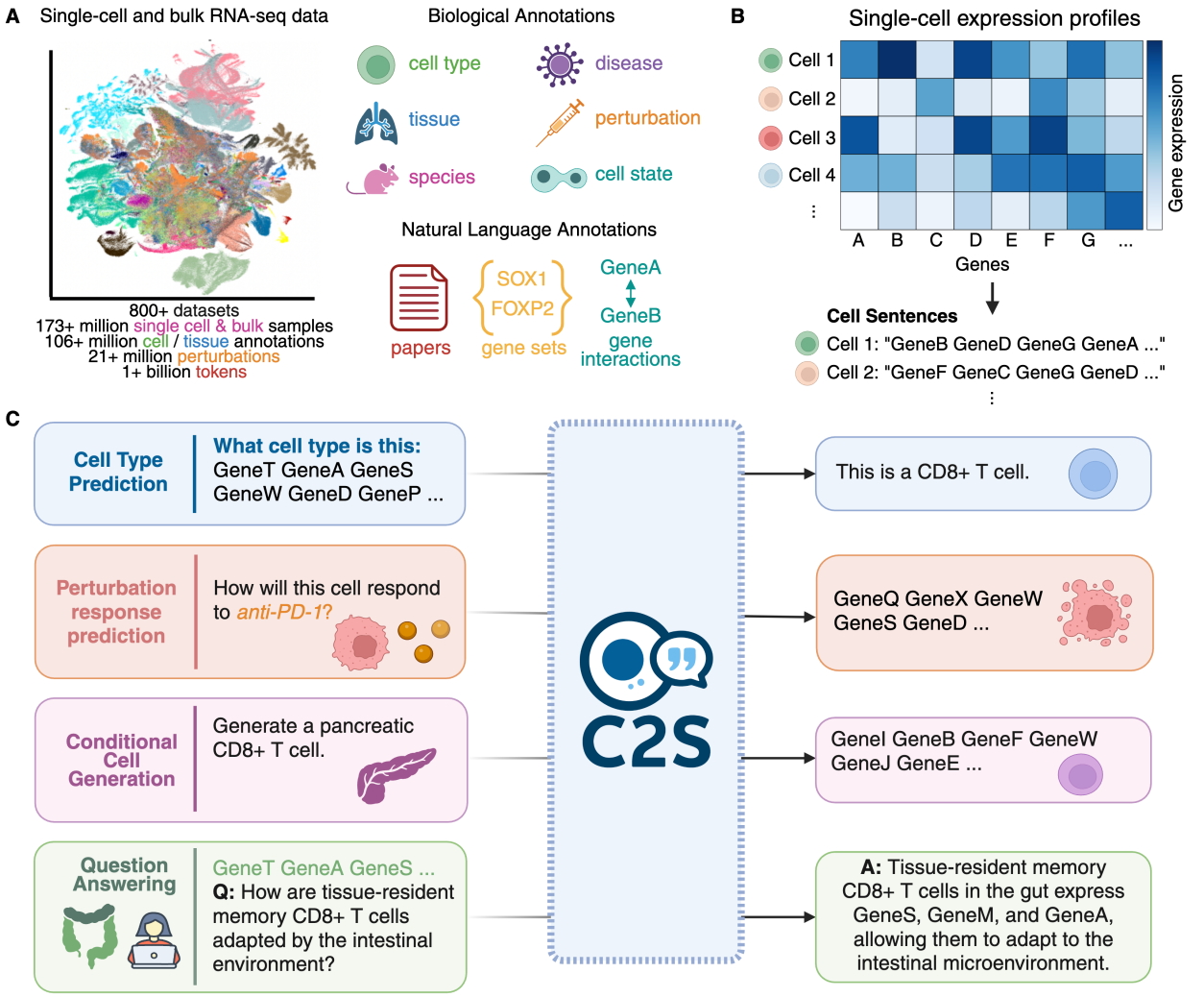

Figure 2: C2S-Scale bridges scRNA-seq data and natural language by training LLMs to perform single-cell analysis tasks on diverse, multimodal data. (A) A multimodal corpus of over 50 million human and mouse transcriptomes is gathered from public data atlases, encompassing cellular expression from a diverse range of tissues, textual annotations, papers, gene sets, and disease labels from scRNA-seq studies. (B) C2S rank-orders genes by expression and converts them to natural language "cell sentences", leveraging powerful LLM architectures without the need for custom modifications. (C) C2S supports diverse downstream use cases, including perturbation prediction, generative tasks, and advanced biological reasoning tasks such as question answering.

2.1 LLM Framework and Training

C2S-Scale uses an LLM-based framework for single-cell analysis, illustrated in Figure 2, building upon and scaling up the original Cell2Sentence framework [13, 14]. To represent cells in natural language, C2S-Scale ranks the expressed genes of each cell in descending order of expression and concatenates their gene names, separated by spaces, creating a "cell sentence" (Figure 2B). This representation preserves relative gene expression while enabling the LLM to leverage its pre-existing knowledge associated with gene names acquired during large-scale pre-training on natural language data. The transformation from expression to cell sentence representation is reversible with minimal information loss due to the strong relationship between relative position and original gene expression [13, 14] (examples provided in Figure 9).

Training C2S-Scale consists of two phases: a self-supervised general pre-training phase on our large-scale corpus, followed by additional tuning for specific tasks. To assemble the pre-training corpus, we collected over 50 million human and mouse transcriptomes from a diverse range of tissues gathered from the CellxGene [2] and Human Cell Atlas [3] data atlases, along with associated annotations, papers, and metadata. We pretrain C2S-Scale on a variety of

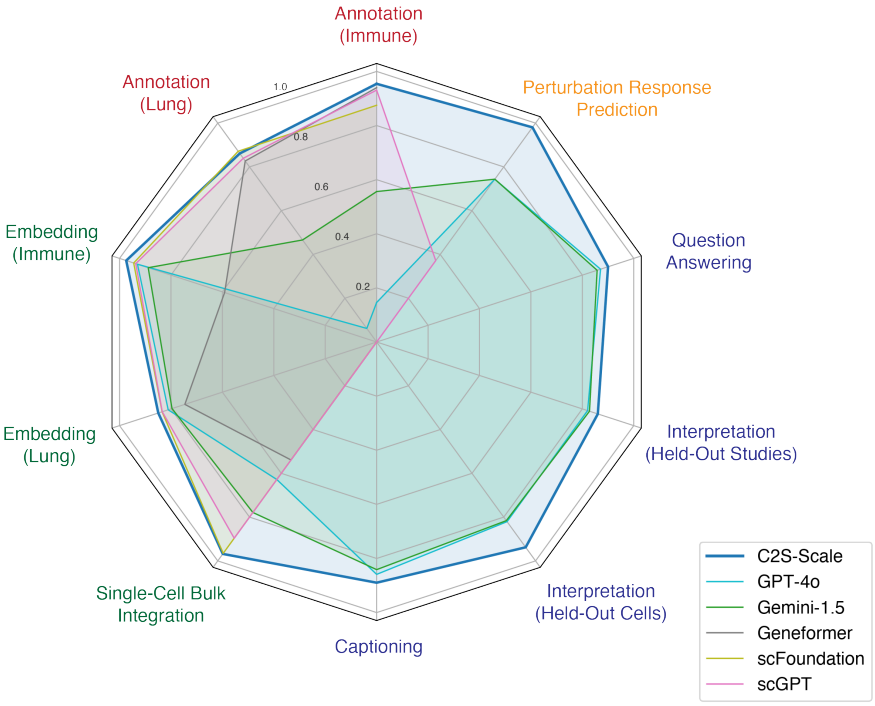

Figure 3: C2S-Scale outperforms both transcriptomic and natural language foundation models across diverse predictive and generative single-cell tasks. Tasks include standard single-cell analysis tasks such as cell type annotation (red) and cell embedding (green), a generative perturbation response prediction task (orange), and natural language interpretation tasks including cluster captioning, dataset interpretation, and question answering tasks (blue). Raw performance numbers are available in the Supplement. C2S-Scale is the only model capable of spanning the entire range of single-cell analysis tasks, and demonstrates competitive performance on all tasks.

tasks constructed using samples from the raw corpus, encompassing predictive and generative tasks on both single and multi-cell context (Table 1). This allows the LLM to learn to model cell sentences while simultaneously learning to follow prompt instructions for common scRNA-seq analysis tasks. During the fine-tuning phase, the pretrained model is specialized for a particular task on a new dataset.

2.2 State-of-the-art Predictive and Generative Capabilities

C2S-Scale demonstrates strong performance across a diverse spectrum of single-cell transcriptomic tasks, outperforming or matching existing state-of-the-art transcriptomic and natural language foundation models (Figure 3). For traditional single-cell analysis tasks like cell type annotation, C2S-Scale is provided with a cell sentence and prompted to predict the corresponding cell type label in natural language. On these tasks, C2S-Scale achieves results competitive with other specialized scFMs such as scGPT [4] and Geneformer [5] on immune tissue [18] and lung tissue [19] datasets. For cell embedding tasks, we leverage the pretrained C2S-Scale models to generate rich cell embeddings given a cell sentence as input. C2S-Scale produces informative cell embeddings that capture both transcriptional and contextual information from natural language. We also construct a multi-modal integration task where we assess the zero-shot similarity of embeddings from paired single-cell and bulk data. We find that C2S has the most consistent embeddings despite none of the models being pre-trained on bulk data, suggesting that C2S innately captures a more biologically meaningful representation of cellular states, likely due to the nature of the cell sentence transformation.

Additionally, C2S-Scale excels in generative tasks without requiring task-dependent architectural modifications, a feature absent in most other transcriptomic foundation models. For perturbation response prediction, C2S-Scale generates accurate predictions of cellular transcriptional responses to various perturbations, generalizing even to combinatorial and previously unseen conditions. This task is described further in Section 2.7. On natural language tasks involving reasoning about scRNA-seq data, C2S-Scale sets a new standard by outperforming state-of-the-art (SOTA) and open-source LLMs such as Llama [20, 21], GPT-4o [22] and Gemini [23] at cluster captioning, dataset interpretation, and question answering tasks. Remarkably, C2S-Scale generalizes effectively to completely unseen

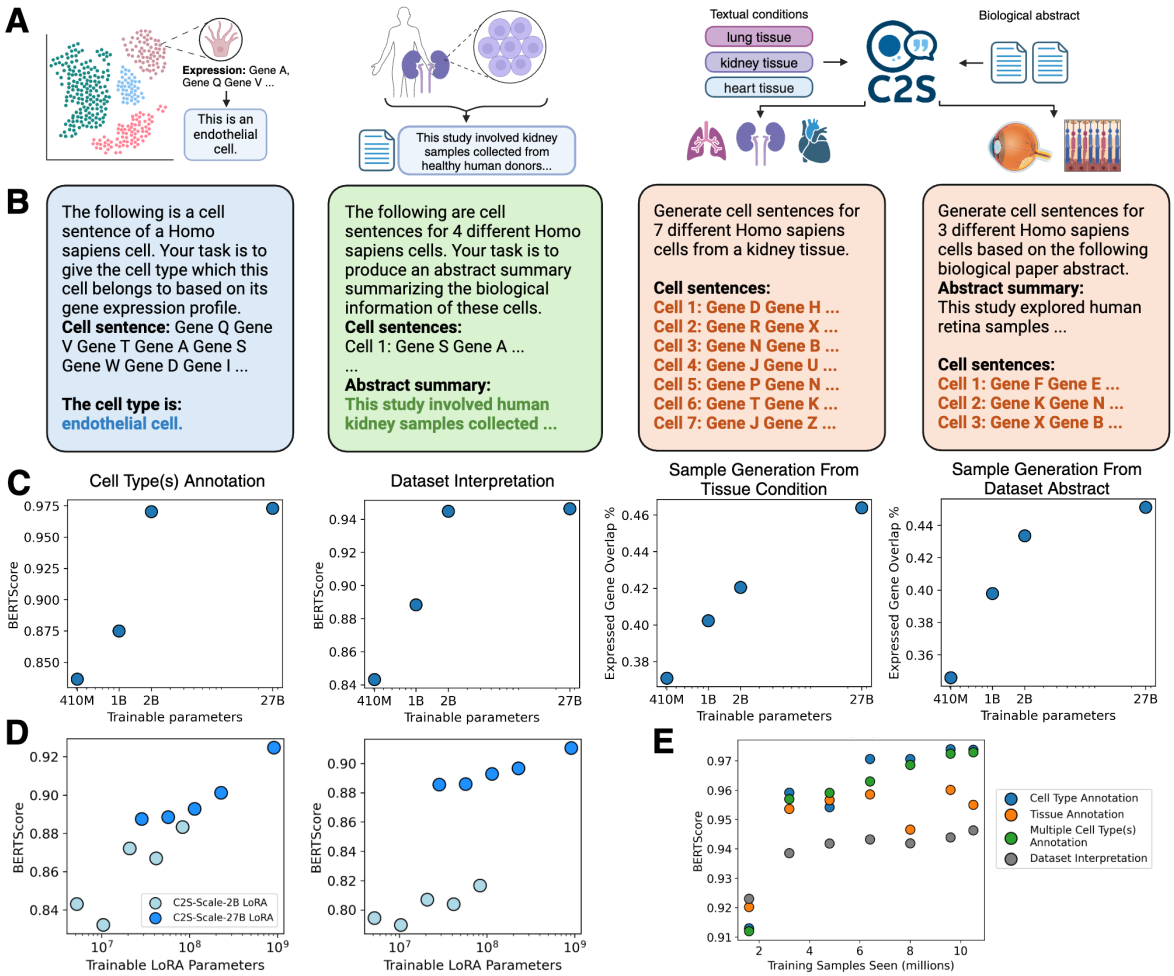

Figure 4: Cell2Sentence demonstrates consistent scaling in performance with increasing model capacity across diverse single-cell analysis tasks. (A) Examples of predictive and generative tasks on single-cell data. (B) Natural language prompts and responses for tasks in (A), colored by expression generation (red), predictive (blue), and language generation (green) tasks. (C) Performance scaling of full-finetuned C2S models on conditional sample generation, cell type annotation, tissue sample annotation, and dataset interpretation. (D) LoRA-finetuned C2S-Scale-2B and 27B models demonstrate performance scaling with increased model capacity in the parameter-efficient regime. (E) Performance scaling by number of training samples seen by C2S-Scale-27B.

scRNA-seq studies (Figure 3), demonstrating its robust interpretative capabilities for novel data. The ability to generate biologically meaningful insights in natural language makes C2S-Scale a uniquely powerful and accessible tool for interacting with and interpreting single-cell data. Detailed description of each task and evaluation methodology can be found in Section 5.

Importantly, C2S-Scale is the *only* model capable of spanning this entire range of single-cell analysis tasks, encompassing both predictive and generative tasks, as well as integrating single-cell data with natural language understanding and reasoning. This positions C2S-Scale as a universal and comprehensive tool for next-generation single-cell analysis.

2.3 Scaling Laws for LLMs in Single-Cell Analysis

Large language models (LLMs) are known to exhibit predictable scaling behavior in natural language tasks [11, 12]. We find that similar scaling laws emerge in the context of single-cell analysis when LLMs are trained on natural language representations of transcriptomic data. As model capacity increases, C2S-Scale models demonstrate consistent improvements across predictive and generative tasks, including cell type annotation, tissue inference, and conditional cell generation (Figure 4C).

These scaling trends are observed in both fully fine-tuned and parameter-efficient training regimes (Figure 4D). In addition to model scaling, we find that performance also improves consistently with increased training data for a fixed model size, as shown for the 27B model in Figure 4E. Together, these results suggest that scaling LLMs—both in capacity and dataset size—can significantly enhance biological reasoning capabilities, mirroring the benefits seen in general NLP.

2.4 Natural Language Interpretation at Multiple Scales of Biology

Natural language interpretation is a underexplored aspect of single-cell analysis, enabling researchers to bridge experimental scRNA-seq data with existing biological literature and providing a user-friendly tool for biologists to interact with and interpret their data. Existing LLM-based single-cell models such as GenePT [24] and scGenePT [7] offered limited integration of natural language and single-cell data, focusing primarily on using language embeddings in single-cell architectures and tasks. C2S-Scale bridges large-scale training on transcriptomic data with the natural language pre-training and generative capabilities of LLMs, enabling natural language interpretation of scRNA-seq data at multiple scales of biology, illustrated in Figure 5A.

We benchmark C2S-Scale on a series of natural language interpretation tasks at various scales of biology, evaluating its ability to reason about and generate meaningful descriptions about data. At the **individual cell level**, C2S-Scale is able to accurately annotate cell types in natural language given cell sentences as input. The model is first fine-tuned on a diverse immune tissue dataset [18] to predict cell type labels in natural language. C2S-scale is able to correctly classify almost all cell types on a held-out partition of the immune tissue data (Figure 5B), demonstrating C2S-Scale's effectiveness at standard single-cell analyses.

At the **cluster level**, we introduce a novel task called Cluster Captioning, where the goal is to generate biologically meaningful descriptions for groups of cells from the same tissue and batch within a scRNA-seq dataset. To create training data for this task, we use GPT-4o [22] to generate natural language captions for cell clusters derived from annotated datasets (Methods Section 5.6). C2S-Scale is fine-tuned to predict these captions given multiple input cell sentences from each cluster and is evaluated on held-out clusters not seen during training. Performance is measured using BioBERTScore [25], which quantifies semantic similarity between generated and ground-truth captions. As shown in Figure 5C, C2S-Scale outperforms all baseline LLMs on this task, demonstrating its ability to interpret and summarize expression patterns at the cluster level.

At the **dataset level**, we further evaluate interpretive ability through a Dataset Interpretation task, where the model receives multiple cell sentences from a scRNA-seq dataset and is tasked with generating a high-level summary in the style of a biological abstract. These summaries are expected to describe key features of the dataset, including dominant cell types, tissues, disease states, or perturbations (examples provided in Figure 10). Figure 5D shows that C2S-Scale achieves the highest BERTScore among all evaluated models—including LLaMA [20, 21, 26], Meditron [27], BioMistral [28], Gemini [23], and GPT-4o [22]. Notably, C2S-Scale generalizes well to entirely unseen datasets, producing summaries that remain relevant and informative (Figure 5E), highlighting its robust natural language understanding of scRNA-seq data.

Overall, C2S-Scale enables natural language interpretation at multiple scales, spanning single cells, clusters, and datasets. Its ability to integrate textual and biological data unlocks new opportunities for biologists to explore, annotate, and generate insights from scRNA-seq data in natural language.

2.5 C2S-Scale Learns Spatial Reasoning from Multi-cell Context and Interaction Data

Understanding spatial organization in tissues is fundamental to uncovering the mechanisms that govern cellular interactions, particularly in how they drive disease progression and tissue homeostasis [29, 30, 31]. Cellular niches, defined by their specific cell types, signaling molecules, and extracellular matrix components, play a crucial role in regulating these processes. Accurately predicting spatial relationships among cells from transcriptomic data alone is challenging, as traditional approaches often rely on explicitly structured spatial models or predefined interaction networks [32, 33, 34].

Although C2S-Scale was not explicitly designed for spatial reasoning, its ability to incorporate multi-cellular context provides a natural mechanism for modeling spatial organization. We hypothesize that by sampling and encoding cells from shared neighborhoods, C2S-Scale can infer spatial relationships without requiring architectural modifications. To test this, we evaluate the model's performance in predicting spatial neighborhoods using a human liver spatial RNA-seq dataset [35]. Additionally, we simultaneously train C2S-Scale on related tasks aimed at improving its spatial understanding: niche label prediction, neighbor cell generation, and determining whether multiple cells belong to the

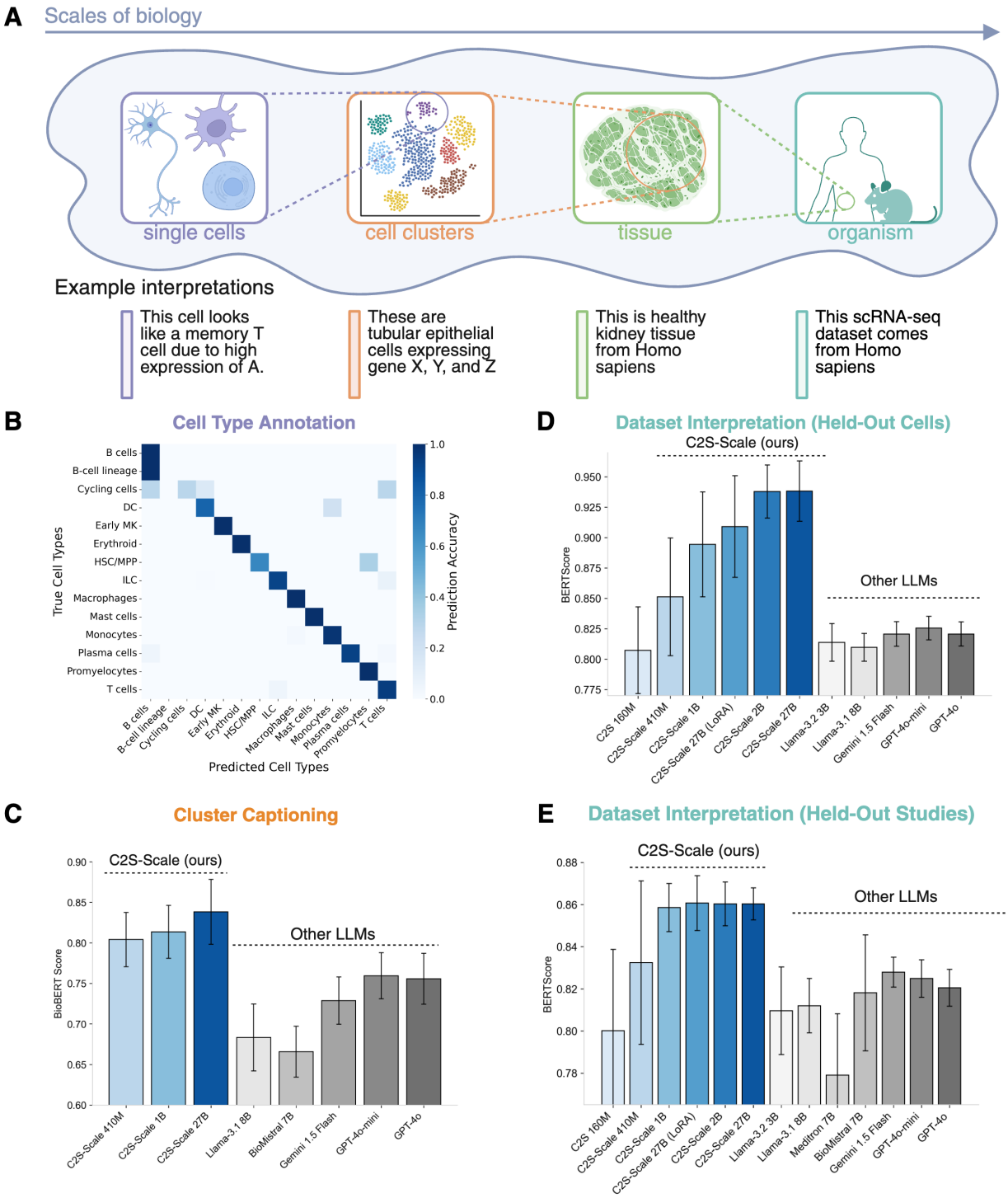

Figure 5: C2S-Scale enables natural language interpretation of scRNA-seq data at multiple scales, from single cells to entire datasets. (A) Different scales of biological data interpretation, from single cells to organism and dataset-level annotation. (B) Ground truth and predicted cell types for immune cells extracted from 16 different tissues of adult human donors [18], demonstrating the ability of C2S-Scale to annotate data at the single-cell level. (C) Performance and example prompts for C2S-Scale on predicting cell-cell interaction in a lymph tissue spatial dataset. (D) Cluster captioning performance on unseen scRNA-seq data clusters. Models are given multi-cell context from unseen data clusters and tasked with captioning the data, measured by BERTScore. (E) Performance of C2S-Scale models on natural language interpretation of entire scRNA-seq datasets on held-out cells and held-out studies. Error bars for (D) - (E) represent standard deviation across test set samples.

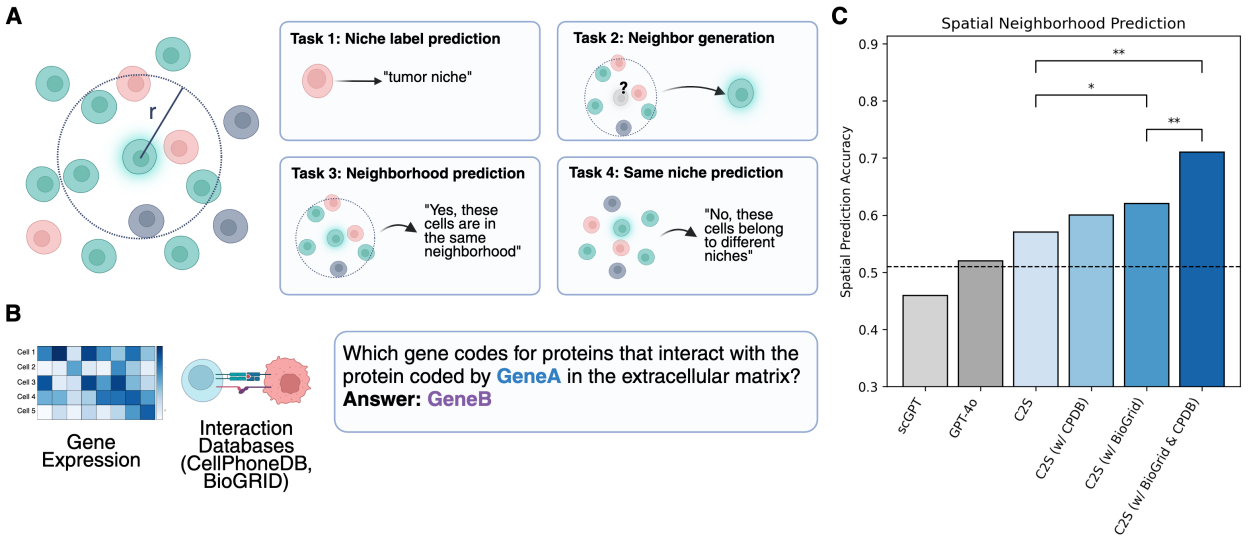

Figure 6: C2S-Scale can interpret multicellular spatial context and predict niche neighborhoods. (A) We fine-tune C2S-Scale on a variety of single and multi-cellular spatial tasks designed to enable C2S-Scale to perform spatial reasoning, including predicting niche labels, generating spatial neighbors, and identifying whether cells belong to the same neighborhood or niche. A "neighborhood" is defined to be cells within a fixed radius from a central cell. (B) We use publicly available gene interaction databases including BioGRID and CellPhoneDB to construct natural language interaction prompts about gene interactions. To maximize relevance, BioGRID is filtered to include only genes expressed in the CosMx dataset and restricted to extracellular proteins. (C) C2S outperforms scGPT and GPT-4o in spatial neighborhood identification accuracy. Additionally, integrating gene interactions from BioGRID and CellPhoneDB individually improves performance, and their combination provides the greatest improvement. These results highlight the multi-task transfer learning potential of C2S-Scale for spatially-aware biological modeling.

same niche (Figure 6A). By training on these complementary tasks, C2S-Scale learns robust representations of spatial organization, significantly outperforming both scGPT and GPT-4o in neighborhood prediction (Figure 6C).

We further hypothesize that incorporating external biological knowledge—specifically, gene interaction networks—can enhance spatial reasoning. Receptor-ligand and other protein-protein interactions are central to cell-cell communication, yet many scFMs are unable to integrate this information. Instead of imposing predefined rules, we simply expose C2S-Scale to receptor-ligand interactions from CellPhoneDB [36] and protein interaction data from BioGRID [37], formatted as natural language prompts (Figure 6B). This approach allows the model to implicitly integrate prior knowledge while maintaining flexibility in how it applies this information.

Fine-tuning with gene interaction data further improves C2S-Scale's ability to predict spatial relationships, reinforcing the hypothesis that external molecular context enhances spatial reasoning (Figure 6B). Notably, adding either CellPhoneDB or BioGRID data individually improves performance, demonstrating that both receptor-ligand and protein-protein interaction knowledge contribute to spatial reasoning (Figure 6C). Moreover, combining both datasets results in the greatest improvement, suggesting that integrating diverse biological interaction sources allows LLMs to develop a richer understanding of multicellular organization and interaction.

A key advantage of C2S-Scale is its ability to integrate diverse data sources without requiring explicitly structured incorporation of external knowledge. Unlike traditional methods that rely on predefined pathways or manually curated interaction models, C2S-Scale implicitly learns to incorporate relevant information during training. This highlights a fundamental strength of C2S: rather than designing bespoke architectures for specific tasks, we can provide relevant data, and the model autonomously determines how to utilize it. This capability extends beyond spatial reasoning and suggests broad applicability for integrating multimodal biological data.

2.6 Single-Cell Question Answering (QA) through Reinforcement Learning

QA tasks form a core part of NLP, providing a standard test to measuring a model's ability to understand information and apply reasoning [38, 39, 40, 41]. In biomedical research, QA tasks are particularly valuable for assessing advanced reasoning in domain-specific contexts, as evidenced by the development of numerous specialized QA datasets for

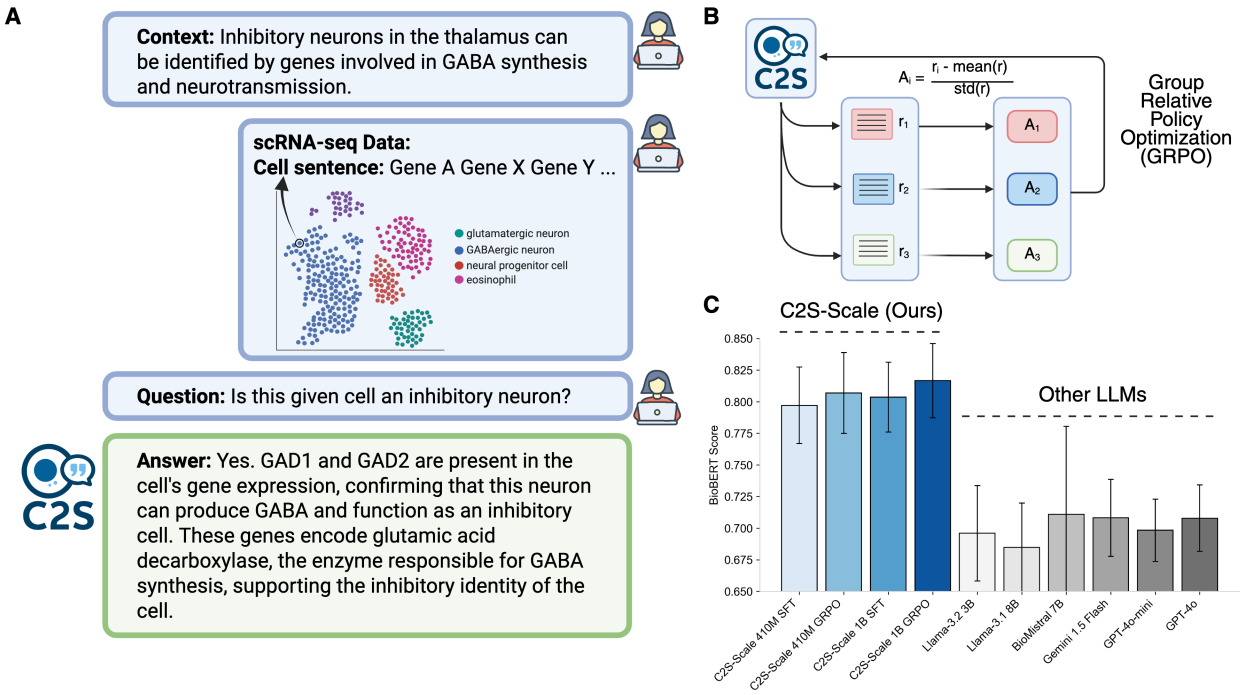

Figure 7: C2S-Scale demonstrates superior single-cell question answering performance compared to state-of-the-art (SOTA) LLMs. (A) Example QA scenario based on scRNA-seq data. (B) Overview of the GRPO framework [17], which further refines model performance by training on preference data. (C) Empirical comparison of C2S-Scale and SOTA LLMs on single-cell QA tasks, highlighting C2S-Scale's advantage in domain-specific reasoning. Error bars represent standard deviation across test set QA samples.

medical [42, 43] and biological [44] applications. Building on this foundation, we introduce a single-cell Question Answering (scQA) task to assess the ability of foundation models to reason about and interpret single-cell transcriptomic data.

The scQA dataset consists of two thousand question answer pairs, each containing: (i) an associated biological context, (ii) relevant scRNA-seq data sampled from clusters or cell type annotations, (iii) a main question, and (iv) a final answer. Additionally, each answer is annotated with keywords to help evaluate response quality. To construct the dataset, we sample cells from scRNA-seq datasets, provide the sampled data along with associated biological manuscripts to GPT-4.5 [22], and prompt it to generate meaningful questions (Figure 7A).

After supervised fine-tuning (SFT), C2S-Scale surpasses the performance of state-of-the-art LLMs on scQA (Figure 7C), demonstrating the advantages of specialized training on transcriptomic data paired with natural language. To further improve C2S-Scale's question answering capabilities, we employ Reinforcement Learning (RL) [45] through Group Relative Policy Optimization (GRPO) to further optimize the model to generated preferred responses to questions (Figure 7B). By using BioBERT Score as the reward function, we guide C2S-Scale toward producing higher-quality answers aligned with biological insights. Following GRPO training, C2S-Scale significantly outperforms the SFT baseline on the scQA dataset, highlighting the potential of RL techniques to optimize LLMs for specialized single-cell applications.

2.7 Perturbation Response Prediction

Single-cell foundation models offer remarkable opportunities for conducting large-scale virtual perturbation experiments that would otherwise be infeasible or prohibitively expensive in a laboratory setting. Here, we demonstrate C2S-Scale's flexibility and accuracy in predicting responses to previously unseen perturbations across diverse settings (Figure 8A).

The prompts used to train C2S-Scale are illustrated in Figure 8C. Training proceeds in two stages: supervised fine-tuning (SFT) followed by reinforcement learning (RL). During SFT, the model is trained to predict gene expression profiles from untreated cells under target perturbation conditions. In the second stage, we apply GRPO [17], an online reinforcement learning algorithm, to optimize perturbation responses with respect to biologically relevant objectives.

While C2S-Scale generates full expression profiles, screening experiments often focus on specific phenotypes rather than all genes. GRPO addresses this by targeting gene programs of interest—apoptosis for the L1000 dataset [46], reflecting the goal of inducing programmed cell death in cancer cells as a therapeutic mechanism, and interferon response for the Dong et al. dataset [47], to capture inflammatory responses to cytokine stimulation. The reward signal is computed over these subsets of genes (Figure 8F), enabling targeted optimization and improving generalization to out-of-distribution settings (Figure 8G).

We introduce a new metric, scFID (Figure 8B), an adaptation of the FID metric [48] widely used in computer vision to evaluate the realism of generated images. scFID replaces the Inception-v3 model [49] with a single-cell foundation model to embed transcriptomic data, enabling biologically meaningful comparisons between real and generated cells. Unlike expression-level metrics, which are sensitive to noise and outliers, scFID offers a robust evaluation in the learned feature space. See Methods (Section 5.7) for details and theoretical connections to Wasserstein distance.

C2S-Scale outperforms existing methods on the Dong et al. dataset, accurately predicting responses to unseen cytokine perturbations on entire gene expression profiles. It generalizes to novel combinations of cell type, cytokine, and exposure duration, producing responses that closely match ground truth (Figure 8E). Compared to scGen, CellOT, and scGPT, C2S-Scale performs best on fully unseen combinatorial perturbations, capturing nonlinear synergistic effects. Quantitative results (Figure 8F) show superior MMD, Wasserstein, and scFID scores. GRPO further reduces scFID on interferon-related genes, improving biological fidelity on immune pathways (Figure 8G).

The L1000 results further underscore C2S-Scale's versatility in modeling perturbation responses across both single-cell and bulk transcriptomic data. We evaluate performance on a subset of apoptosis-related genes, focusing on the model's ability to generalize to unseen compound treatments. Figure 8G shows a consistent performance gain when applying GRPO, with notable improvements in both Kendall's τ and Pearson's r for models of 410M and 1B parameters. These results demonstrate that reinforcement learning not only improves alignment with biologically meaningful responses but also enhances the model's generalization to perturbations outside the training distribution.

3 Discussion

Our work introduces C2S-Scale, a family of LLMs for single-cell analysis that leverages the benefits of state-of-the-art LLMs out of the box. By converting transcriptomic profiles into "cell sentences," C2S-Scale avoids the need for bespoke model architectures while readily integrating contextual information from annotations, metadata, and biological texts. This data engineering paradigm yields a flexible system capable of predictive and generative single-cell tasks, and our results demonstrate that scaling C2S-Scale up to 27 billion parameters systematically boosts performance, mirroring similar scaling phenomena observed in the broader field of NLP.

Moreover, we show that C2S-Scale bridges the gap between raw transcriptomic information and natural language-based interpretation by supporting tasks at multiple scales, ranging from cell type annotation to entire dataset-level summarization. We propose new evaluation datasets for these interpretation tasks and demonstrate that LLMs trained in the C2S-Scale framework provide meaningful captions and summarizations of single-cell data, even in cases where the dataset is completely new to the model. By aligning expression data with rich textual metadata and biological domain knowledge, our approach highlights the potential of language-based modeling to offer biologically informed explanations and generate insights unavailable to purely expression-only systems.

Higher-capacity models and more diverse training corpora can unlock advanced capabilities, such as the integration of epigenomic, proteomic, and clinical data into a single multimodal model. In parallel, increasing transparency and explainability in LLM decision making will be essential for building trust and accelerating adoption of these tools in single-cell research. Reinforcement Learning and other innovations in LLM alignment will provide a path forward for aligning LLMs to preferred responses in the context of biological tasks. By directly linking natural language and transcriptomic data, C2S sets the stage for transformative innovations in biological discovery and personalized medicine.

4 Limitations

4.1 Addressing Limitations of Causal Attention in Gene Expression Modeling

While our approach demonstrates strong empirical performance in modeling single-cell gene expression using autoregressive language models, we acknowledge that causal attention's inherent unidirectionality—favoring high-to-low gene expression dependencies—could theoretically limit the modeling of true causal biological interactions that flow from low- to high-expression genes. However, we contend that this constraint does not significantly impede our objectives and can be mitigated through several complementary strategies. First, our approach aligns with successful

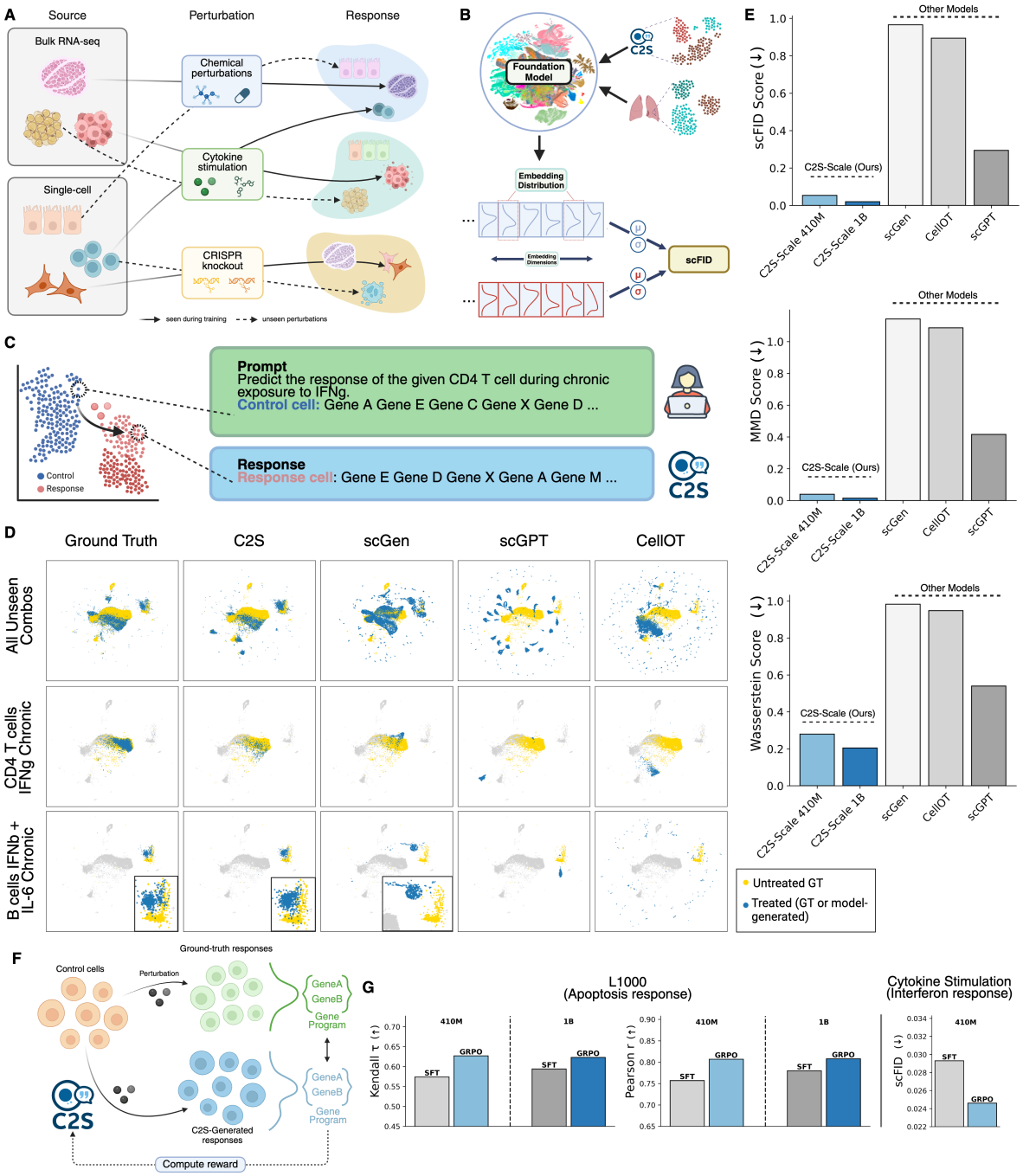

Figure 8: C2S-Scale models outperform existing methods in predicting cellular responses to unseen perturbations. (A) Overview of the C2S-Scale perturbation prediction framework, which supports diverse perturbation types including drugs, cytokines, and genetic knockouts. (B) Diagram of the scFID metric, computed in foundation model latent space, analogous to FID in computer vision. (C) Prompt and response example for perturbation prediction. (D) UMAPs comparing predicted vs. ground-truth responses for unseen perturbations across four models. Rows show: (1) all combinatorial perturbations, (2) CD4 T cells under IFN-γ, (3) B cells under the held-out IFN-β + IL-6 stimulation. C2S-Scale aligns closely with ground truth in all cases. (E) Benchmark metrics show C2S-Scale outperforms scGen, scGPT, and CellOT across all evaluation criteria. (F) GRPO framework for perturbation prediction: models generate perturbed responses and receive rewards based on gene program similarity. (G) GRPO improves over SFT on L1000 (apoptosis response) and cytokine stimulation (interferon response) tasks, with gains in Kendall's τ, Pearson's r, and scFID.

paradigms from vision-language models, where arbitrary tokenization orders paired with causal attention still achieve state-of-the-art performance [50]. Similar to hybrid vision architectures that combine causal and non-causal attention layers, our framework could incorporate indirect bidirectional context through auxiliary reasoning tokens or non-causal gene interactions.

Multi-cell context and reasoning as a corrective mechanism The model's reasoning capabilities provide additional corrective potential. Emerging evidence from language modeling demonstrates that explicit reasoning steps can compensate for causal attention limitations [51, 52, 53]. In our context, intermediate tokens representing biological pathways or gene interactions enable iterative prediction refinement, effectively circumventing strict unidirectionality. Furthermore, our multi-cell training framework enables implicit bidirectionality—low-expression genes in one cell can influence high-expression genes in the following cell, approximating bidirectional attention across a multi-cell context.

Correlation, not causation It is important to emphasize that our model is designed to capture predictive correlations over inferring causal gene relationships. This mirrors natural language processing, where autoregressive models successfully capture statistical correlations despite occasional misalignment between word order and true causal relationships (e.g. passive constructions) [54, 55]. Our results confirm that expression correlations provide sufficient predictive power for key biological analysis tasks.

Architectural enhancements Looking forward, we propose three architectural enhancements to further mitigate this limitation: (1) bidirectional attention by partitioning gene sequences, (2) variable gene ordering during training to induce order invariance, and (3) hybrid attention architectures blending causal and non-causal attention layers. While our current approach already demonstrates that sequential modeling of gene expression—despite lacking natural ordering—leverages pre-trained LLMs without requiring custom architectures, these enhancements aim to further improve biological fidelity and predictive power.

In summary, while causal attention restricts bidirectionality within individual cells, its ability to capture correlations aligns with our predictive objectives. The combined effects of multi-cell context, reasoning mechanisms, and prospective architectural improvements position this approach as a robust foundation for single-cell analysis, with multiple pathways available for extending its biological fidelity.

4.2 Hallucination and Interpretability

A known challenge with large language models is their tendency to generate plausible but incorrect outputs, often referred to as hallucinations. While our benchmarking focuses on structured biological tasks with ground-truth labels, more open-ended interpretation tasks—such as abstract generation or cluster captioning—may be susceptible to such errors. Developing domain-specific safeguards, such as biological fact-checking mechanisms or constrained decoding strategies, remains an important direction for improving interpretability and reliability in high-stakes settings.

5 Methods

The following section details the data collection, processing, and and formatting for multi-task samples, as well as the model architecture for Large Language Models.

5.1 Data Collection

To construct the C2S-Scale pre-training corpus, we assembled a large-scale dataset encompassing over 50 million single-cell transcriptomic profiles from human and mouse tissues. These scRNA-seq datasets were sourced from established public repositories, including the CellxGene [2] and Human Cell Atlas [3] data portals. The datasets span a broad range of biological contexts, and include associated annotations and textual data such as cell type and tissue annotations, disease states, experimental conditions, and associated biological papers and abstracts. We applied standard preprocessing pipelines for scRNA-seq data, including quality control, normalization, and log-transformation, following established conventions [56]. For each dataset, any available annotations, including cell type, tissue type, disease state, donor ID, development stage, species, and associated paper were kept for constructing natural language prompts after converting the raw transcriptomic data into cell sentences. This forms a multimodal training corpus with linked transcriptomic and natural language data.

5.2 Cell Sentence Transformation

To adapt high-dimensional single-cell gene expression data into a format compatible with natural language processing, we converted expression profiles into textual representations termed "cell sentences." For each cell, let $X \in \mathbb{R}^D$ be the

expression vector, where X_k denotes the normalized expression value of gene k in that cell. The cell sentence for X is constructed by rank-ordering the genes within a cell by their expression levels and taking the K most highly expressed genes. If S is a list of indices from 1 to D sorted in descending order based on expression level in X, then

$$\text{cell sentence}(X) := \text{“gene}(S[1])\ \text{gene}(S[2])\ \ldots\ \text{gene}(S[K])\text{”}. \tag{1}$$

The gene names are in natural language, forming a sentence interpretable by language models (exemplified in Figure 2). Under this framework, there is no need to extend or modify the vocabulary of the language model, and it allows any LLM architecture to tokenize gene names according to their existing vocabulary. This has two primary benefits: (i) by avoiding architectural modifications, the C2S framework is immediately applicable to any LLM architecture or innovation, and (ii) the LLM is able to recognize gene names and associate prior knowledge about that gene obtained during self-supervised pre-training on natural language data, which has been shown to be significant for large-scale pretrained LLMs [24].

The cell sentence transformation into textual sequences retains the underlying biological information by preserving the rank-order of gene expression. We find there is a strong linear relationship (in log space) between a gene's rank in the cell sentence and the (normalized) expression level, validating the fidelity of this transformation. This relationship is shown in Supplementary Figure 9 for two scRNA-seq datasets. A linear model fitted between rank and original expression can predict the original gene expression values given a gene's rank with an R^2 of 85% (Figure 9), demonstrating that minimal information is lost during conversion to cell sentences. This interchangeability allows us to utilize the strength of LLMs in natural language processing while retaining the ability to convert back to gene expression vectors for traditional single-cell analysis methods. The parameters of the linear model for each scRNA-seq dataset used during training are saved to enable reversible transformation from cell sentences back to expression values during inference.

Multi-Task Prompt Formatting. By operating in natural language, C2S-Scale enables diverse input and output context for predictive and generative single-cell analysis tasks, including cell type and tissue annotation, multi-cell generation tasks, and dataset interpretation tasks. The full list of pre-training tasks, inputs, and outputs of the model are detailed in Table 1. To construct prompts for specific tasks, each prompt combines the cell sentence representation of one or multiple cells with task-specific instructions to guide the model to perform the specific task. For predictive tasks, the cell sentence information is given as part of the input prompt, and the response contains the metadata label of interest. For example, for the cell type annotation task, the input would contain the cell sentence and a natural language prompt such as “Predict the cell type of this cell:”, and the output would be the cell type label. For generative tasks this is reversed; metadata conditions are given in the input prompt, and the output response contains the cell sentence(s). Metadata given in natural language prompts can include cell type, tissue annotations, perturbation conditions, biological abstracts and text, and disease states, to provide additional biological context or conditions. This approach ensures that C2S-Scale learns to interpret and perform complex biological tasks within the framework of natural language, and enables it to generalize to diverse applications.

5.3 C2S-Scale Architecture and Pre-training

Word Embedding in Transformers. The C2S-Scale framework uses LLMs, which are based on the Transformer architecture [8], to model cell sentences and perform single-cell analysis in natural language. Language models represent input sequences of text as sequences of high-dimensional vectors known as “word embeddings”, suitable for processing by neural networks. Each word in a cell sentence corresponds to a gene name, which is further split into tokens using the pretrained tokenizer associated with the model's backbone architecture. By reusing the existing tokenizer associated with the LLM, we avoid introducing new vocabulary and maintain compatibility with the model's pre-training knowledge.

The tokenized gene names are embedded into vector spaces by means of an embedding layer trained alongside the model. These embeddings capture the semantic information of genes, informed by both biological context and the language model's prior knowledge. This representation enables the Transformer to interpret and process complex gene expression patterns encoded in cell sentences.

Attention Mechanism. Central to modern language model architectures is the attention mechanism [57], which allows the model to identify and focus on key components of input sequences. Self-attention, the predominant method used in Transformer models [8], is employed to compute attention scores between tokens. This mechanism enables the model to dynamically weigh the importance of different genes within a cell sentence, depending on the task. For example, the model may emphasize lineage-defining marker genes for cell type classification tasks while focusing on perturbation-associated genes for prediction tasks.

The attention mechanism also facilitates the integration of additional contextual metadata, such as cell type or tissue labels, by attending to these features alongside the cell sentences. This ensures that the model considers both the textual representation of gene expression and the accompanying biological context during processing.

Table 1: Pre-training task inputs and outputs for C2S-Scale multi-task training. For multi-cell tasks, multiple cells are sampled from the same donor sample with the same tissue label.

Task name	Type	Input information	Target output	Metric
Single cell language modeling	Single-cell	–	Single cell sentence	Overlap %
Cell type annotation	Single-cell	Single cell sentence	Cell type	BertScore
Conditional cell generation	Single-cell	Cell type of one cell	Single cell sentence	Overlap %
Multiple cell language modeling	Multi-cell	–	Multiple cell sentences	Overlap %
Tissue sample annotation	Multi-cell	Multiple cell sentences	Tissue label	BertScore
Sample cell type(s) annotation	Multi-cell	Multiple cell sentences	Cell types of multiple cells	BertScore
Conditional sample generation (tissue)	Multi-cell	Tissue annotation	Multiple cell sentences	Overlap %
Conditional sample generation (cell type)	Multi-cell	Cell types of multiple cells	Multiple cell sentences	Overlap %
Conditional sample generation (abstract)	Multi-cell	Paper abstract	Multiple cell sentences	Overlap %
Natural language interpretation	Multi-cell	Multiple cell sentences	Paper abstract	BertScore
Gene set enumeration	Gene set	Gene set name	List of genes in gene set	Overlap %
Gene set naming	Gene set	List of genes in gene set	Gene set name	BertScore

Transformer Architecture. LLMs are decoder-only Transformer architectures, chosen for its proven capabilities in sequential data modeling and generative tasks [22]. The Transformer consists of stacked blocks, each comprising a self-attention layer followed by a feedforward network with residual connections and layer normalization [8]. This modular design enables scalable and efficient learning across a wide range of tasks.

Key architectural components include:

1. Self-Attention Layers: These layers compute relationships between all tokens in the input sequence, allowing the model to capture long-range dependencies in gene expression data.
2. Feedforward Networks: Each attention layer is followed by a feedforward network, which applies non-linear transformations to enhance feature extraction.
3. Residual Connections and Layer Normalization: These components stabilize training and facilitate gradient flow, enabling the model to scale effectively to large parameter sizes.

Pre-training Objectives. The pre-training objective of LLMs is next token prediction [58], a foundational task in generative language modeling introduced . In this setup, the model learns to predict the next token in a sequence given all preceding tokens, enabling it to capture complex dependencies and semantic relationships within the input data. For cell sentences, this objective involves predicting the next gene name in the rank-ordered sequence based on the expression levels of preceding genes, while incorporating contextual metadata, such as cell type or tissue annotations, when provided. While prior work such as Geneformer [5] also rank-orders genes and uses a masked modeling objective to predict genes in the sequence, their formulation is not in natural language and lacks the autoregressive framing central to generative LLMs. In contrast, our approach trains the model to understand gene expression patterns and their hierarchical organization through natural language modeling, conditioning it to integrate biological context naturally via autoregressive generation. The sequential nature of next token prediction aligns seamlessly with downstream generative tasks, such as cell sentence generation and annotation, ensuring that the model can generate coherent and biologically meaningful outputs when applied to single-cell analyses.

Training Setup. The pre-training was conducted on a corpus of over 50 million single-cell transcriptomes and associated metadata and textual annotations, as described in the previous section. We employed multi-task learning to jointly optimize the model across predictive and generative tasks described in Table 1, allowing it to develop a comprehensive understanding of single-cell data linked with natural language. The training utilized modern optimizers and techniques, such as AdamW and gradient checkpointing, to efficiently manage computational resources for models ranging from 1 billion to 27 billion parameters. We used Huggingface [59] and PyTorch [60] to train LLM models up to the 1B parameter scale, and afterwards used Jax and TPU-based compute resources to train models from 2B to 27B capacity.

5.4 Scaling Evaluation

To evaluate scaling behavior in C2S-Scale models, we benchmarked models ranging from 410 million to 27 billion parameters, based on the Gemma 2 [15] and Pythia [16] architectures. We assessed performance on a held-out set of 500 test samples spanning multiple single-cell tasks listed in Table 1, including cell type annotation, tissue classification, dataset interpretation, and conditional sample generation. Both fully fine-tuned and LoRA-finetuned variants [61] were evaluated to assess scaling behavior under different computational budgets.

Performance was measured using BERTScore [25] between generated and reference outputs for predictive tasks such as cell type annotation and dataset interpretation, providing a semantic measure of response quality. For generative tasks like conditional cell generation, we evaluated performance by measuring gene overlap between generated and target cell sentences.

5.5 Post-training Methods

Supervised fine-tuning. After pre-training, C2S-Scale is fine-tuned on task-specific datasets to adapt the model to downstream applications in single-cell analysis. During this stage, the model is trained using labeled data for tasks such as cell type annotation, tissue-level classification, and cell generation. Next-token-prediction [58] is again used for the supervised fine-tuning phase, with natural language prompts formatted for the downstream task.

To maintain efficiency and minimize overfitting, we employ parameter-efficient fine-tuning techniques, including LoRA (Low-Rank Adaptation) and lightweight adapter layers. These methods allow fine-tuning a subset of model parameters while keeping the majority of the pretrained weights frozen. This approach enables rapid adaptation to specific tasks without requiring extensive computational resources or large labeled datasets.

Reinforcement Learning. To further enhance performance on generative and interpretative tasks, we draw on RL techniques aimed at aligning LLM outputs to preferred standards using reward modeling 45. Specifically, we employ GRPO, a reward-based method that directly updates the model parameters based on gradient signals tied to task-specific criteria, thereby aligning C2S outputs with biological accuracy and interpretability.

The GRPO process starts with generating multiple candidate outputs for each training example using the SFT model. These candidates are then ranked by quality; in conventional NLP settings, human preference rankings are often used. However, in C2S-Scale, we rely on domain-specific criteria and automated metrics such as BERTScore [25] to assess semantic similarity to reference answers, as well as the biological plausibility of responses for tasks like question answering. By optimizing against these ranked outputs, GRPO fine-tunes the model to favor higher-scoring (i.e., higher-quality and more biologically aligned) answers.

Compared to other RL methods, such as Proximal Policy Optimization (PPO) [62], GRPO offers a more streamlined workflow: rather than requiring a separate reward model, it directly incorporates the reward signals—here, bioBERT-based or domain-specific metrics—into the gradient updates. This direct integration simplifies the alignment process, making it particularly efficient for large-scale models like C2S-Scale. By focusing the optimization on biologically relevant metrics, GRPO enables consistent improvements in specialized single-cell tasks, ensuring that C2S-Scale steadily refines its outputs in a manner consistent with expert expectations and high-quality biological insights.

5.6 Downstream Tasks

Cell type annotation. For the cell type annotation task, we fine-tune the model to predict cell type labels on an immune tissue dataset [63] and a lung dataset [19]. We use 80% of cells from each dataset for training and reserve 20% for evaluation. C2S-Scale is provided with a cell sentence and a natural language prompt, such as “Predict the cell type of this cell:”. C2S-Scale is fine-tuned for this task using the same next-token prediction objective [58] as the pre-training step, predicting cell type labels in natural language. Other scFMs are tuned using prediction heads on top of the pretrained transformer weights in accordance with the recommended strategies for each model.

Cell generation. For cell generation tasks, we finetune the model to unconditionally or conditionally generate cell expression on the immune tissue and lungdatasets. The model is given a natural language prompt containing relevant metadata for conditional generation, or no information in the case of unconditional generation, and is tasked with generating a cell sentence of K genes representing the expression of the cell under that condition. For instance, to conditionally generate a B cell, the model might be given a prompt such as: "Generate a list of 1000 genes in order of descending expression which represent a Homo sapiens cell of cell type B cell."

Cell embedding. For cell embedding, we use trained C2S-Scale foundation models (e.g. C2S-Scale 1B) trained on the C2S multimodal corpus to embed cells without any dataset-specific fine-tuning. To embed cells, we first format input prompts for C2S-Scale in the same manner as in cell type prediction tasks. However, instead of decoding token predictions, we take the last hidden state from the last layer of the C2S-Scale model, and average pool the latents in order to form our embedding of the input prompt. We note that this procedure can be done for multi-cell contexts as well as contexts that involve different metadata and condition components in natural language prompts, making C2S-Scale a diverse embedding model for transcriptomic and language inputs.

Single-cell bulk integration. Multimodal integration is essential for capturing the complexity of biological systems, as different data modalities provide complementary perspectives on cellular function. Each modality has its own strengths

and limitations — some offer high resolution at the cost of sparsity, while others provide broader coverage but lack single-cell detail. Therefore, models that can integrate modalities can provide a more complete and robust understanding of cellular behavior, improving both interpretability and predictive power in biological analysis.

To assess this, we design a simple single-cell and bulk RNA seq integration task. Using the single-cell lung tissue data from [19], we construct pseudo-bulk samples by aggregating over donor, cell type, and batch. For each pseudobulk sample, we randomly sample ten single-cell samples from the same conditions to construct pairs. We embed each single-cell and pseudobulk sample individually using each model and compute the cosine similarity between the paired single-cell and bulk samples. Following [64], we use the "fraction of samples closer than the true match" (FOSCTTM) to evaluate the performance of each model. A FOSCTTM of 0 corresponds to a perfect model (the cosine similarity of matched pairs is higher than any other pair), whereas a FOSCTTM close to 0.5 means the cosine similarity between the matched pairs is about as good as the cosine similarity between random pairs.

Cluster captioning. To generate the cluster captioning dataset, we select 30 scRNA-seq datasets and perform standard preprocessing, clustering, and differential expression analysis. We then prompt GPT-4o [22] to generate five captions for a cluster based on the cell type, tissue type, organism, disease, top three differentially expressed genes, and the full text of the associated paper. This resulted in a total dataset of 1,723 captions from 345 distinct clusters. To produce the final training data, we randomly sample two cells from a cluster to construct the training prompt, and a caption from that cluster as the target. The C2S-Scale models were fine-tuned using supervised fine-tuning with a next-token prediction learning objective with a learning rate of 1×10^{-5}, weight decay of 0.01, and a batch size of 64. All models were evaluated on the same holdout test set consisting of clusters unseen in the training data.

Dataset interpretation. For the dataset-level interpretation task, we create two test sets for dataset-level interpretation: (i) a training distribution dataset interpretation test set, where scRNA-seq data and paper abstracts come from 613 of the scRNA-seq datasets gathered from CellxGene [2] as a part of the C2S-Scale training corpus, and (ii) a out-of-distribution (OOD) evaluation set where the papers and data are completely unseen by the C2S-Scale model. By evaluating dataset-level interpretation on scRNA-seq studies from both the training corpus and out of distribution data, we create a challenging generalization benchmark for writing meaningful interpretations of scRNA-seq data.

Each dataset interpretation sample was created by sampling between 5 and 20 cells from the same tissue and donor in a given scRNA-seq dataset, and formatting a prompt with the multi-cell context that tasked the model with generating a biological abstract summary to describe the data. The ground truth for the abstract summary of the data was obtained by taking the abstract of the paper associated with the scRNA-seq study; to create more diversity in the biological abstracts seen across samples, we create 500 variations of each dataset abstract using GPT-3.5-Turbo-1106, to prevent the model from simply memorizing a few hundred dataset abstracts. For each multi-cell context, we choose one of the abstract summaries as the ground truth target summary. Example abstract summaries can be found in Figure 10.

To create the training corpus distribution dataset interpretation test set, we first gather held-out abstract generation samples from the training corpus. These are multi-cell contexts and samples which the model had not seen during training since they were a part of held-out validation and test sets of the C2S-Scale corpus, however since each dataset only contains 1 abstract, the held-out samples will still contain similar information to training set abstract generation samples that the model has seen. We sample 5 held-out abstract generation samples from 613 datasets gathered from CellxGene [2], yielding a total test set of 3065 dataset interpretation samples.

For the out-of-distribution dataset interpretation test set, we constructed new abstract generation samples by dowloading two new datasets from CellxGene that were either published recently (after the initial C2S-Scale corpus gathering period) or verified to not be a part of the C2S-Scale training corpus: (i) a pancreas tissue [65] and a human retina [66] dataset. We constructed 200 samples from each dataset, again creating 50 variations of the abstract of each dataset to again provide more diversity in summary language.

Spatial niche prediction. We utilized the CosMx Spatial Molecular Imager Human Liver dataset [35], which provides annotated spatially-resolved single-cell data from both normal and hepatocellular carcinoma liver tissues from two different donors. This dataset encompasses over 800,000 single cells across a total of approximately 180mm^2 of liver tissue, with expression measured on a set of 1,000 curated genes. The dataset was processed to filter out genes expressed in fewer than three cells and cells expressing fewer than 50 genes. It was then normalized to a total count of 1×10^4 and the base 10 logarithm was applied. Spatial coordinates were saved to define neighborhoods and faciliate spatial analyses. We define a neighborhood to be a radius of 0.02 pixels (approximately $20\mu m$), chosen to maximize the number of cells we can fit into the model's context. The dataset was split into training and test sets based on spatial coordinates to prevent spatial leakage between sets.

To train C2S-Scale on spatial and multi-cellular relationships, we designed the following tasks:

1. **Niche label prediction:** Given a cell sentence for a single cell, predict the niche label annotation for that cell.

2. **Conditional Neighbor Generation:** Given multiple cell sentences from a neighborhood, generate a novel cell sentence that would belong to the same neighborhood.
3. **Spatial neighborhood prediction:** Given multiple cell sentences, predict whether these cells come from the same neighborhood.
4. **Same niche prediction:** Give multiple cell sentences, predict whether all of these cells have the same niche label or different niches.

To construct prompts, cell sentences were randomly sampled from the appropriate data split. Multi-cell contexts were created by taking all cells in the sampled cell's neighborhood for positive samples, or an equivalent number of randomly sampled cells outside the neighborhood as negative samples.

Additionally, to enhance the model's understanding of cell communication, we included gene interaction metadata from CellPhoneDB [36] and BioGRID [37]. We restricted the data to only retain interactions involving the 1,000 genes in the CosMx data, and also only to genes coding for extra-cellular proteins (determined using MatrixDB [67]).

Question answering. We begin by using the GPT-4.5 model to generate question–answer pairs from three sections of each manuscript—abstracts, discussions, and results—as well as data sampled from that study. Each scRNA-seq study contributes 20 QA pairs, for a total of approximately 1600 QA pairs used for SFT. We conduct SFT with a learning rate of 1×10^{-5} and 100 warmup steps.

Following SFT, we apply GRPO to further refine answer quality. To create the GRPO training set, we collect an additional 600 samples from unseen studies, with each sample prompting the SFT model to generate 32 candidate answers. We then use BioBERT to compute a reward score for each candidate answer against the ground truth answer provided by GPT-4.5, capturing its biological plausibility. These BioBERT-derived scores serve as the primary reward signals, guiding the GRPO update step and optimizing model parameters to favor biologically accurate, contextually relevant responses. For GRPO training, we set $\beta = 0.03$ and use a learning rate of 5×10^{-7}. Finally, we evaluate the GRPO-refined model on a new test set derived from unseen studies, and compare its performance against a commonly used LLM, as illustrated in Figure 7.

Perturbation prediction. The Dong et al. dataset includes immune cells exposed to individual and combinatorial cytokines, with each cell annotated by type, stimulation, and exposure length—yielding 133 conditions. We retain the 5000 most highly variable genes and evaluate models in the scGPT embedding space [4] using maximum mean discrepancy (MMD), Wasserstein distance, and scFID (Section 5.7). This embedding-based evaluation provides more meaningful comparisons than expression-level metrics, which can be skewed by a small number of genes with extreme values.

The training of C2S models for the Dong et al. dataset followed a structured two-stage process to effectively predict responses to unseen cytokine stimulations. The test dataset featured three tiers of held-out perturbations with increasing difficulty: (1) a completely excluded combinatorial perturbation (interferon-β + IL-6), (2) one perturbation entirely held out for each cell type across both chronic and acute conditions (B: interferon-III, CD4 T: interferon-γ, CD8 T: interferon-α2, Dendritic: interferon-β (no chronic cells), NK: IL-6), and (3) one perturbation excluded in either chronic or acute conditions for each cell type while the other condition remained in training (B: acute interferon-β, CD4 T: acute interferon-β + interferon-γ, CD8 T: chronic TNF-α, NK: chronic interferon-III). In the first stage, the model was fine-tuned using supervised learning on both cell sentence generation and natural language label prediction, where it simultaneously predicted all three labels—cell type, perturbation, and exposure—ensuring it learned bidirectional relationships between conditions and gene expression. This fine-tuning stage was conducted for 3–4 epochs using the Hugging Face Trainer on a single H100 GPU.

The second stage employed GRPO to refine perturbation response generation. For the Dong et al. dataset, the reward was computed as the negative mean squared error between generated and ground truth cells, randomly paired under the same condition labels and embedded using scGPT. GRPO training used 32 generated responses and 32 real cells per prompt, and was conducted on 4 H100 GPUs for 3 epochs. The interferon subset used for GRPO was defined as the union of the MSigDB [68] interferon-α and interferon-γ hallmark gene sets, intersected with the highly variable genes (HVGs) from the dataset, resulting in 95 genes.

To benchmark against other perturbation response models, we included scGen, CellOT, and scGPT. For scGen, we used the pertpy library [69] to generate perturbation predictions. For CellOT, we followed the standard procedure but replaced the encoder with the pretrained encoder from scGen. For scGPT, we added linear encoders for cell type, perturbation, and exposure, projecting binary vectors into dense vectors, and then added these embeddings to each gene token embedding before forwarding them through the model.

For the L1000 dataset [46], we trained on the 978 landmark genes following quantile normalization. We paired untreated and treated samples by matching the cell line name. To evaluate generalization, we selected 50 perturbations with fewer

than 1,000 total samples and held out half the cell lines in each perturbation as test data. We used Kendall's Tau as the reward function during reinforcement learning, as it properly accounts for tied ranks. This is especially important for L1000 where non-expressed genes share the same lowest rank. SFT was conducted using a batch size of 2 and gradient accumulation of 32, with a learning rate of 1e-4. Training ran on a single H100 GPU for 3,500 steps (approximately one epoch, though not all data is seen due to dataset size). For GRPO, the model was trained with a batch size of 8 and gradient accumulation of 4. We generated 24 responses per prompt. The learning rate was set to 1e-6 with a beta value of 5e-3. Training was distributed across four H100 GPUs—three for model training and one for vLLM-based response generation. GRPO ran for approximately 3,000 steps over 3 epochs, although as with SFT, the model likely saw less than a full epoch due to data scale. Only the apoptosis genes from the MSigDB hallmark set that were present in the L1000 landmark gene list were used during GRPO, totaling 40 genes.

5.7 Single-Cell Fréchet Inception Distance

The scFID is an adaptation of the FID [48] tailored for evaluating generative models in single-cell transcriptomics. While the traditional FID employs the Inception v3 model [49] to extract features from images, scFID utilizes scGPT [4] as its foundation model to embed single-cell gene expression profiles. Notably, scFID is flexible and can incorporate any suitable foundation model for embedding. The scFID quantifies the similarity between the distributions of real and generated single-cell embeddings by assuming that these distributions are multivariate normal (Gaussian). Under this assumption, the scFID computes the Wasserstein distance between the two Gaussian distributions, providing a measure of how closely the generated data resembles the real data in the embedding space.

Mathematically, given two sets of single-cell embeddings—one from real cells and one from generated cells—scFID is defined as:

$$\text{scFID} = \|\mu_r - \mu_g\|_2^2 + \operatorname{tr}\left(\Sigma_r + \Sigma_g - 2\left(\Sigma_r \Sigma_g\right)^{\frac{1}{2}}\right)$$

where:

- μ_r and μ_g are the mean vectors of the real and generated cell embeddings, respectively,
- Σ_r and Σ_g are the covariance matrices of the real and generated cell embeddings, respectively,
- tr denotes the trace of a matrix.

To evaluate generative model performance across various conditions, we compute the scFID for each unique combination of test labels—such as specific cell types, perturbations, and exposure durations—and then average these individual scFID values.

6 Acknowledgements

The authors thank collaborators and contributors from across institutions for their invaluable support and insights throughout this project. This work was supported in part by the National Institutes of Health (NIH) grant R35GM143072–01 and the Yale Colton Center Award, both awarded to Dr. David van Dijk.

All figures were created in BioRender, https://BioRender.com.

References

[1] Philipp Angerer, Lukas Simon, Sophie Tritschler, F Alexander Wolf, David Fischer, and Fabian J Theis. Single cells make big data: new challenges and opportunities in transcriptomics. *Current opinion in systems biology*, 4:85–91, 2017.

[2] CZI Cell Science Program, Shibla Abdulla, Brian Aevermann, Pedro Assis, Seve Badajoz, Sidney M Bell, Emanuele Bezzi, Batuhan Cakir, Jim Chaffer, Signe Chambers, et al. Cz cellxgene discover: A single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *Nucleic Acids Research*, page gkae1142, 2024.

[3] Aviv Regev, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, et al. The human cell atlas. *elife*, 6:e27041, 2017.

[4] 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*, pages 1–11, 2024.

[5] 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, et al. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023.

[6] 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*, pages 1–11, 2024.

[7] Ana-Maria Istrate, Donghui Li, and Theofanis Karaletsos. scgenept: Is language all you need for modeling single-cell perturbations? *bioRxiv*, pages 2024–10, 2024.

[8] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[9] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[10] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[11] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[12] Biao Zhang, Zhongtao Liu, Colin Cherry, and Orhan Firat. When scaling meets llm finetuning: The effect of data, model and finetuning method. *arXiv preprint arXiv:2402.17193*, 2024.

[13] Rahul M Dhodapkar. Representing cells as sentences enables natural-language processing for single-cell transcriptomics. *bioRxiv*, pages 2022–09, 2022.

[14] Daniel Levine, Syed Asad Rizvi, Sacha Lévy, Nazreen Pallikkavaliyaveetil, David Zhang, Xingyu Chen, Sina Ghadermarzi, Ruiming Wu, Zihe Zheng, Ivan Vrkic, et al. Cell2sentence: Teaching large language models the language of biology. *bioRxiv*, pages 2023–09, 2023.

[15] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

[16] Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR, 2023.

[17] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[18] C Domínguez Conde, C Xu, LB Jarvis, DB Rainbow, SB Wells, T Gomes, SK Howlett, O Suchanek, K Polanski, HW King, et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science*, 376(6594):eabl5197, 2022.

[19] Jieun Kim, Eun-Young Eo, Bokyong Kim, Heetak Lee, Jihoon Kim, Bon-Kyoung Koo, Hyung-Jun Kim, Sukki Cho, Jinho Kim, and Young-Jae Cho. Transcriptomic analysis of air–liquid interface culture in human lung organoids reveals regulators of epithelial differentiation. *Cells*, 13(23):1991, 2024.

[20] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[21] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[22] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[23] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

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

[25] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

[26] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[27] Antoine Bosselut, Zeming Chen, Angelika Romanou, Antoine Bonnet, Alejandro Hernández-Cano, Badr Alkhamissi, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, et al. Meditron: Open medical foundation models adapted for clinical practice. 2024.

[28] Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*, 2024.

[29] Rohit Arora, Christian Cao, Mehul Kumar, Sarthak Sinha, Ayan Chanda, Reid McNeil, Divya Samuel, Rahul K Arora, T Wayne Matthews, Shamir Chandarana, et al. Spatial transcriptomics reveals distinct and conserved tumor core and edge architectures that predict survival and targeted therapy response. *Nature Communications*, 14(1):5029, 2023.

[30] Mikala Egeblad, Elizabeth S Nakasone, and Zena Werb. Tumors as organs: complex tissues that interface with the entire organism. *Developmental cell*, 18(6):884–901, 2010.

[31] Giuliana Mannino, Cristina Russo, Grazia Maugeri, Giuseppe Musumeci, Nunzio Vicario, Daniele Tibullo, Rosario Giuffrida, Rosalba Parenti, and Debora Lo Furno. Adult stem cell niches for tissue homeostasis. *Journal of Cellular Physiology*, 237(1):239–257, 2022.

[32] Zixuan Cang and Qing Nie. Inferring spatial and signaling relationships between cells from single cell transcriptomic data. *Nature communications*, 11(1):2084, 2020.

[33] Suoqin Jin, Christian F Guerrero-Juarez, Lihua Zhang, Ivan Chang, Raul Ramos, Chen-Hsiang Kuan, Peggy Myung, Maksim V Plikus, and Qing Nie. Inference and analysis of cell-cell communication using cellchat. *Nature communications*, 12(1):1088, 2021.

[34] Erick Armingol, Adam Officer, Olivier Harismendy, and Nathan E Lewis. Deciphering cell–cell interactions and communication from gene expression. *Nature Reviews Genetics*, 22(2):71–88, 2021.

[35] Shanshan He, Ruchir Bhatt, Brian Birditt, Carl Brown, Emily Brown, Kan Chantranuvatana, Patrick Danaher, Dwayne Dunaway, Brian Filanoski, Ryan G Garrison, et al. High-plex multiomic analysis in ffpe tissue at single-cellular and subcellular resolution by spatial molecular imaging. *BioRxiv*, pages 2021–11, 2021.

[36] Kevin Troulé, Robert Petryszak, Martin Prete, James Cranley, Alicia Harasty, Zewen Kelvin Tuong, Sarah A Teichmann, Luz Garcia-Alonso, and Roser Vento-Tormo. Cellphonedb v5: inferring cell-cell communication from single-cell multiomics data. *arXiv preprint arXiv:2311.04567*, 2023.

[37] Rose Oughtred, Jennifer Rust, Christie Chang, Bobby-Joe Breitkreutz, Chris Stark, Andrew Willems, Lorrie Boucher, Genie Leung, Nadine Kolas, Frederick Zhang, et al. The biogrid database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Science*, 30(1):187–200, 2021.

[38] P Rajpurkar. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

[39] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[40] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*, 2024.

[41] Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*, 2023.

[42] Asma Ben Abacha and Dina Demner-Fushman. A question-entailment approach to question answering. *BMC bioinformatics*, 20:1–23, 2019.

[43] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR, 2022.

[44] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*, 2019.

[45] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[46] Aravind Subramanian, Rajiv Narayan, Steven M Corsello, David D Peck, Ted E Natoli, Xiaodong Lu, Joshua Gould, John F Davis, Andrew A Tubelli, Jacob K Asiedu, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452, 2017.

[47] Mingze Dong, Bao Wang, Jessica Wei, Antonio H. de O. Fonseca, Curtis J. Perry, Alexander Frey, Feriel Ouerghi, Ellen F. Foxman, Jeffrey J. Ishizuka, Rahul M. Dhodapkar, and David van Dijk. Causal identification of single-cell experimental perturbation effects with cinema-ot. *Nature Methods*, 20(11):1769–1779, 2023.

[48] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[49] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[50] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[51] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[52] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

[53] Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. *arXiv preprint arXiv:2305.15852*, 2023.

[54] Cara Su-Yi Leong and Tal Linzen. Language models can learn exceptions to syntactic rules. *arXiv preprint arXiv:2306.05969*, 2023.

[55] Michael Wilson, Jackson Petty, and Robert Frank. How abstract is linguistic generalization in large language models? experiments with argument structure. *Transactions of the Association for Computational Linguistics*, 11:1377–1395, 2023.

[56] Ashraful Haque, Jessica Engel, Sarah A Teichmann, and Tapio Lönnberg. A practical guide to single-cell rna-sequencing for biomedical research and clinical applications. *Genome medicine*, 9:1–12, 2017.

[57] Dzmitry Bahdanau. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[58] Alec Radford. Improving language understanding by generative pre-training. 2018.

[59] T Wolf. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

[60] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[61] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[62] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[63] C Xu, L Jarvis, T Gomes, S Howlett, D Rainbow, O Suchanek, H King, L Mamanova, K Polanski, N Huang, et al. Cross-tissue immune cell analysis reveals tissue-specific adaptations and clonal architecture in humans. 2021.

[64] Zhi-Jie Cao and Ge Gao. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nature Biotechnology*, 40(10):1458–1466, 2022.

[65] Maayan Baron, Adrian Veres, Samuel L Wolock, Aubrey L Faust, Renaud Gaujoux, Amedeo Vetere, Jennifer Hyoje Ryu, Bridget K Wagner, Shai S Shen-Orr, Allon M Klein, et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. *Cell systems*, 3(4):346–360, 2016.

[66] Zhen Zuo, Xuesen Cheng, Salma Ferdous, Jianming Shao, Jin Li, Yourong Bao, Jean Li, Jiaxiong Lu, Antonio Jacobo Lopez, Juliette Wohlschlegel, et al. Single cell dual-omic atlas of the human developing retina. *Nature Communications*, 15(1):6792, 2024.

[67] Olivier Clerc, Madeline Deniaud, Sylvain D Vallet, Alexandra Naba, Alain Rivet, Serge Perez, Nicolas Thierry-Mieg, and Sylvie Ricard-Blum. Matrixdb: integration of new data with a focus on glycosaminoglycan interactions. *Nucleic acids research*, 47(D1):D376–D381, 2019.

[68] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, and Jill P Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.*, 102(43):15545–15550, October 2005.

[69] L Heumos, Yuge Ji, Lilly May, Tessa D Green, Xinyue Zhang, Xichen Wu, Johannes Ostner, Stefan Peidli, Antonia Schumacher, Karin Hrovatin, M F Mueller, F Chong, Gregor Sturm, Alejandro Tejada, Emma Dann, Mingze Dong, Mojtaba Bahrami, Ilan Gold, Sergei Rybakov, Altana Namsaraeva, A Moinfar, Zihe Zheng, Eljas Roellin, Isra Mekki, C Sander, M Lotfollahi, Herbert B Schiller, and Fabian J Theis. Pertpy: an end-to-end framework for perturbation analysis. *bioRxiv*, August 2024.

7 Supplementary

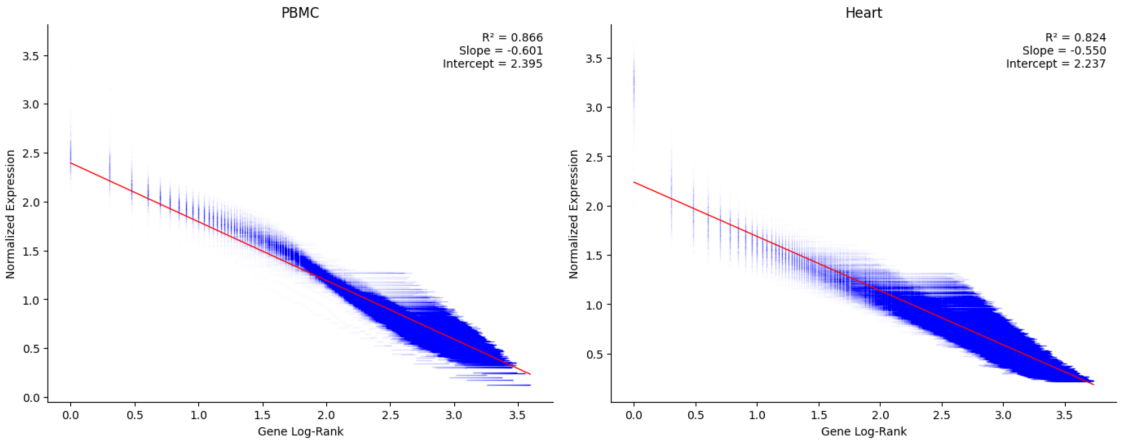

Figure 9: C2S allows for conversion from expression information into cell sentence format with minimal information loss. Using a linear model fitted between rank and original expression, cell sentences can be converted back to expression accurately.

High-throughput single-nucleus RNA sequencing of over three million nuclei from the entire adult human brain identified 461 clusters and 3313 subclusters. The analysis revealed area-specific cortical neurons, diverse midbrain and hindbrain neurons, and regional diversity in astrocytes and oligodendrocyte precursors. This study provides a comprehensive understanding of the molecular diversity of the human brain, offering insights into brain health and diseases.

Single-cell and single-nucleus assays were used to create a detailed atlas of healthy and diseased kidney cells, identifying rare populations and altered cellular states in kidney injury. This revealed biological pathways related to chronic kidney disease progression. The atlas, developed through collaborative efforts, aims to provide a valuable resource for kidney research.

Single-cell RNA sequencing of glioblastoma cells from four patients revealed genomic and transcriptomic variations within the tumor. Infiltrating neoplastic cells shared a consistent gene signature across patients, suggesting a common infiltration mechanism. Additionally, distinct myeloid cell populations were identified in the tumor core and surrounding peritumoral space. This study provides detailed insights into GBM cell types, shedding light on tumor formation and migration.

Figure 10: Example abstract summaries from scRNA-seq datasets collected from CellxGene [2].