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

Syed Asad Rizvi[*,†]
Yale University, Google Research

Daniel Levine[*]
Yale University

Aakash Patel[*]
Yale University

Shiyang Zhang[*]
Yale University

Eric Wang[*]
Google DeepMind

Curtis Jamison Perry[*]
Yale University

Nicole Mayerli Constante
Yale University

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

Hattie Chung
Yale University

Rahul M. Dhodapkar
University of Southern California

Bryan Perozzi
Google Research

Jeffrey Ishizuka[‡]
Yale University
jeffrey.ishizuka@yale.edu

Shekoofeh Azizi[‡]
Google DeepMind
shekazizi@google.com

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

October 10, 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. Scaling the model to 27 billion parameters yields consistent improvements in predictive and generative capabilities and supports advanced downstream tasks that require synthesis of information across multi-cellular contexts. Targeted fine-tuning with modern reinforcement learning techniques produces strong performance in perturbation response prediction, natural language interpretation, and complex biological reasoning. This predictive strength directly enabled a dual-context virtual screen that uncovered a striking context split for the kinase inhibitor silmitasertib (CX-4945), suggesting its potential as a synergistic, interferon-conditional amplifier of antigen presentation. Experimental validation in human cell models unseen during training confirmed this hypothesis, demonstrating that C2S-Scale can generate biologically grounded, testable discoveries of context-conditioned biology. C2S-Scale unifies transcriptomic and textual data at unprecedented scales, surpassing both specialized single-cell models and general-purpose LLMs to provide a platform for next-generation single-cell analysis and the development of "virtual cells."

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

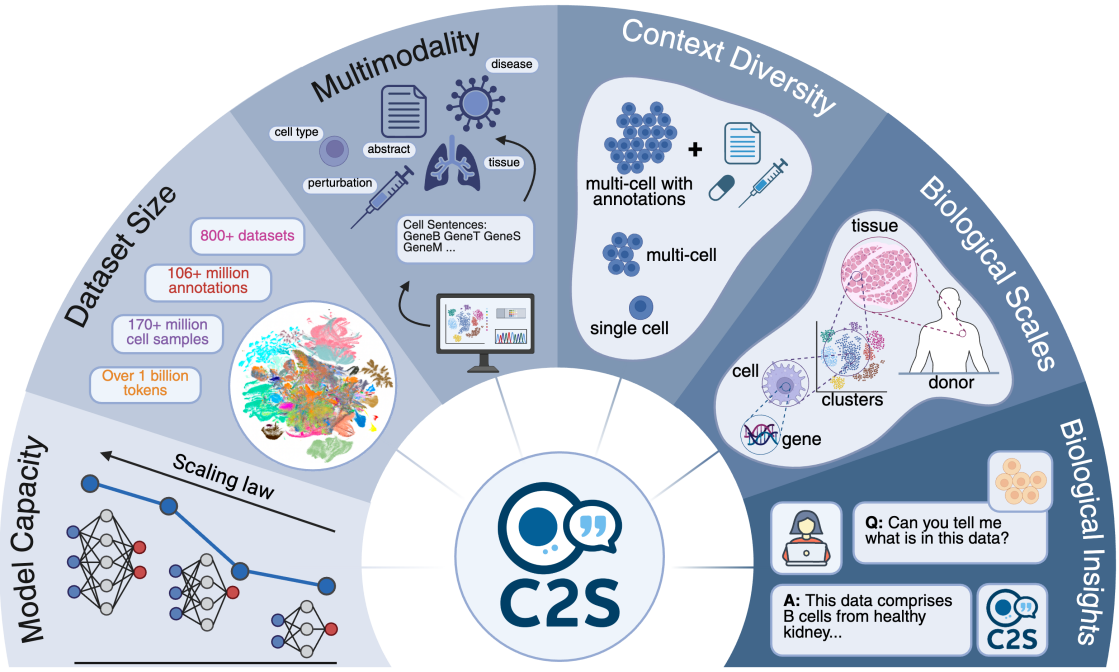

Figure 1: Scaling LLM-based single-cell analysis. A multidimensional expansion of the C2S [14] framework, demonstrating advances in model capacity, dataset size, multimodality, multi-cell support, and integration across biological scales, from single cells to organism-wide insights in natural language. 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 limited by custom architectures constrained to scRNA-seq data, 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 opportunity for new approaches that can integrate diverse data types, including the rich contextual information contained in biological text and metadata.

Large Language Models (LLMs) [8, 9, 10] offer a promising solution to these challenges. Widely used in natural language processing (NLP), LLMs exhibit consistent performance improvements with scale across 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] provides a framework to leverage LLMs for biology by transforming high-dimensional single-cell data into a textual format. 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 **C2S-Scale**, a new family of LLMs trained on a multimodal corpus of over 50 million cells and associated text. We show that scaling these models up to 27 billion parameters leads to consistent performance improvements across a range of predictive and generative tasks (Fig. 1). C2S-Scale's flexible context allows it to analyze cellular interactions and diverse biological information in multi-cell contexts, enabling sophisticated applications from predicting perturbation responses to answering complex biological questions. To further enhance the biological accuracy of model outputs, we developed refinement techniques with reinforcement learning (GRPO) to align model predictions

with key biological objectives. We also introduce a novel metric, single-cell Fréchet Inception Distance (scFID), for assessing generative performance.

To demonstrate this platform's capacity for novel biological discovery, we programmed a dual-context virtual screen designed to find interferon (IFN)-conditional amplifiers of antigen presentation. The screen revealed a pronounced context split for the kinase inhibitor silmitasertib, which has not been reported to enhance MHC-I expression. Our model predicted a strong effect in the context of low levels of IFN exposure, but no effect in the absence of IFN signaling. We validated this prediction in targeted wet lab experiments using neuroendocrine human cell models not seen during training.

By releasing our models and resources, we provide a powerful, open-source platform for next-generation single-cell analysis.

2 Results

2.1 C2S-Scale: A foundation model for single-cell analysis at scale

To create a model capable of jointly interpreting transcriptomic data and biological text, we developed **C2S-Scale**, a family of LLMs trained on a large-scale corpus of scRNA-seq data and associated text (Fig. 2). C2S-Scale builds on the Cell2Sentence framework [13, 14], which represents single-cell gene expression profiles as textual "cell sentences": lists of gene names ranked by their expression level (Fig. 2B). This representation preserves relative gene expression while also allowing the model to leverage its knowledge about genes learned from vast text corpora. 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 Fig. 10).

Training C2S-Scale consists of two phases: a self-supervised general pretraining phase on our large-scale corpus, followed by additional tuning for specific tasks. To assemble the pretraining 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 pretrained C2S-Scale on a variety of 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 C2S-Scale demonstrates broad predictive and generative capabilities

We evaluated C2S-Scale on a diverse spectrum of single-cell tasks, outperforming or matching existing state-of-the-art transcriptomic and natural language foundation models (Fig. 3). For traditional single-cell analysis tasks, C2S-Scale achieved results competitive with expression-only foundation models such as scGPT [4] and Geneformer [5] on immune [15] and lung [16] datasets. For example, on a diverse immune tissue dataset, C2S-Scale predicted cell type annotations in natural language with 95.43% accuracy, slightly better than scGPT (93.1%) and Geneformer (94.0%). C2S-Scale models also generated rich cell embeddings when given a cell sentence as input, capturing both transcriptional and contextual information from natural language. We also construct a multimodal integration task assessing the similarity of embeddings of paired single-cell and bulk data. Notably, C2S-Scale could accurately match single-cell profiles to their corresponding bulk RNA-seq profiles despite no prior exposure to bulk RNA-seq data, suggesting that C2S captures a more biologically meaningful representation of cellular states through cell sentences.

Beyond these predictive tasks, C2S-Scale supports complex generative and interpretive functions not present in most other transcriptomic foundation models. For instance, C2S-Scale accurately predicts cellular transcriptional responses to perturbations, even generalizing to combinatorial and previously unseen conditions (described further in Section 2.7). Furthermore, when tasked with interpreting scRNA-seq data using natural language, C2S-Scale outperformed even leading general-purpose LLMs such as Llama [17, 18], GPT-4o [19] and Gemini [20] at tasks such as generating descriptive captions for cell clusters and summarizing entire datasets. Remarkably, C2S-Scale generalizes effectively to completely unseen scRNA-seq studies (Fig. 3), demonstrating its interpretive capabilities on completely unseen datasets. On question answering in natural language, C2S-Scale outperformed the best public LLM model (GPT-4o) by 3% in BERTScore, highlighting its answer quality and natural language capabilities. 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 4.

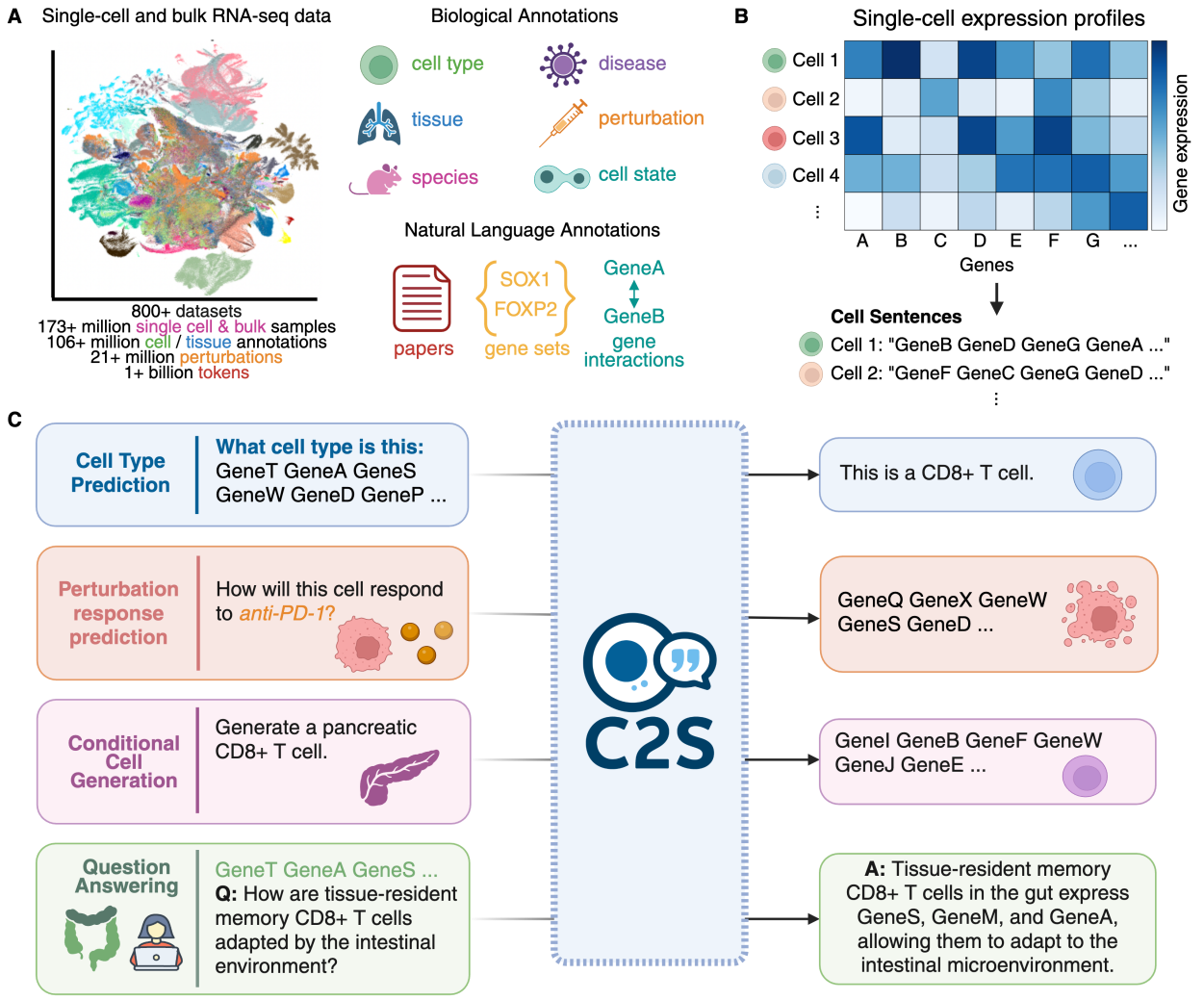

Figure 2: C2S-Scale bridges scRNA-seq data and natural language by training LLMs to perform single-cell analysis tasks on diverse, multimodal data. (A) A multimodal corpus of over 50 million human and mouse transcriptomes is gathered from public data atlases, encompassing cellular expression from a diverse 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.

Taken together, these results show that C2S-Scale is a uniquely versatile tool. It is the only model to our knowledge capable of spanning this entire range of single-cell analysis tasks, including prediction, generation, and natural language reasoning. This positions C2S-Scale as a comprehensive platform for next-generation biological discovery.

2.3 Scaling enhances the biological reasoning capabilities of C2S-Scale

A central principle of modern LLMs is that their performance improves predictably with increased scale [11, 12]. We analyzed the performance of C2S-Scale at a range of model capacities to test whether similar effects exist for LLMs in single-cell analysis. Our results show that similar scaling laws emerge when LLMs are trained on natural language representations of transcriptomic data: as model size increased from 410 million to 27 billion parameters, we observed consistent performance improvements across diverse biological tasks, including cell type annotation, tissue inference, and conditional cell generation (Fig. 4C).

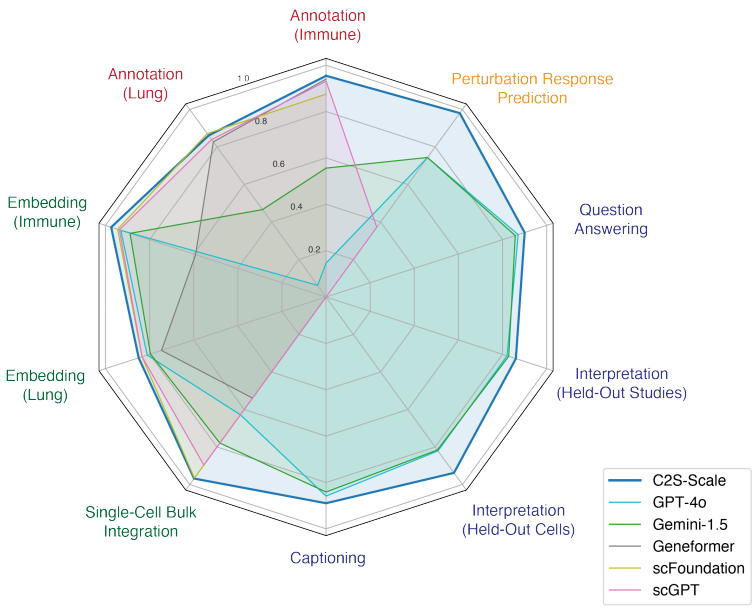

Figure 3: C2S-Scale outperforms both transcriptomic and natural language foundation models across diverse predictive and generative single-cell tasks. Tasks include standard single-cell analysis tasks such as cell type annotation (red) and cell embedding (green), a generative perturbation response prediction 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.

These scaling trends were consistent in both fully fine-tuned and parameter-efficient training regimes where only a fraction of model parameters were trained (Fig. 4D). Furthermore, for a fixed model size, performance also scaled with the amount of training data seen by the model (Fig. 4E). Together, these results establish that increasing both model and dataset size is a reliable strategy for enhancing the biological reasoning capabilities of cellular language models. This suggests that the full potential of this approach has not yet been reached and that future, larger models may yield even greater biological insights.

2.4 Interpreting single-cell data across biological scales using natural language

Natural language interpretation is an 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 [21] 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 pretraining and generative capabilities of LLMs, enabling natural language interpretation of scRNA-seq data at multiple scales of biology, illustrated in Fig. 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 [15] 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 (Fig. 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 [19] to generate natural language captions for cell clusters derived from annotated datasets (Methods Section 4.6). 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 [22], which quantifies semantic similarity between generated and ground-truth captions. As

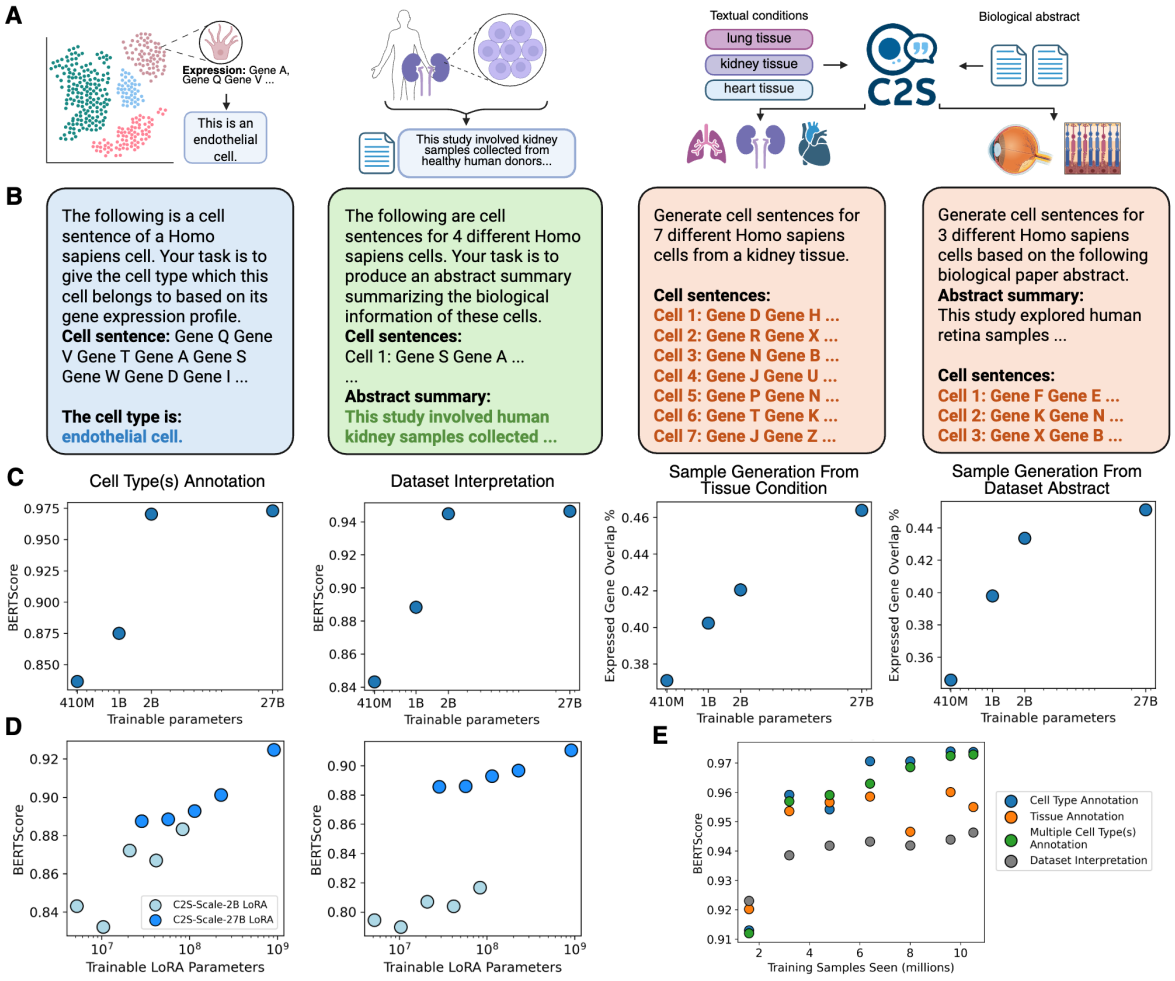

Figure 4: Cell2Sentence demonstrates consistent scaling in performance with increasing model capacity across diverse single-cell analysis tasks. (A) Examples of predictive and generative tasks on single-cell data. (B) Natural language prompts and responses for tasks in (A), colored by expression generation (red), predictive (blue), and language generation (green) tasks. (C) Performance scaling of fully fine-tuned C2S models on cell type annotation, dataset interpretation, and conditional sample generation tasks. (D) LoRA fine-tuned 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.

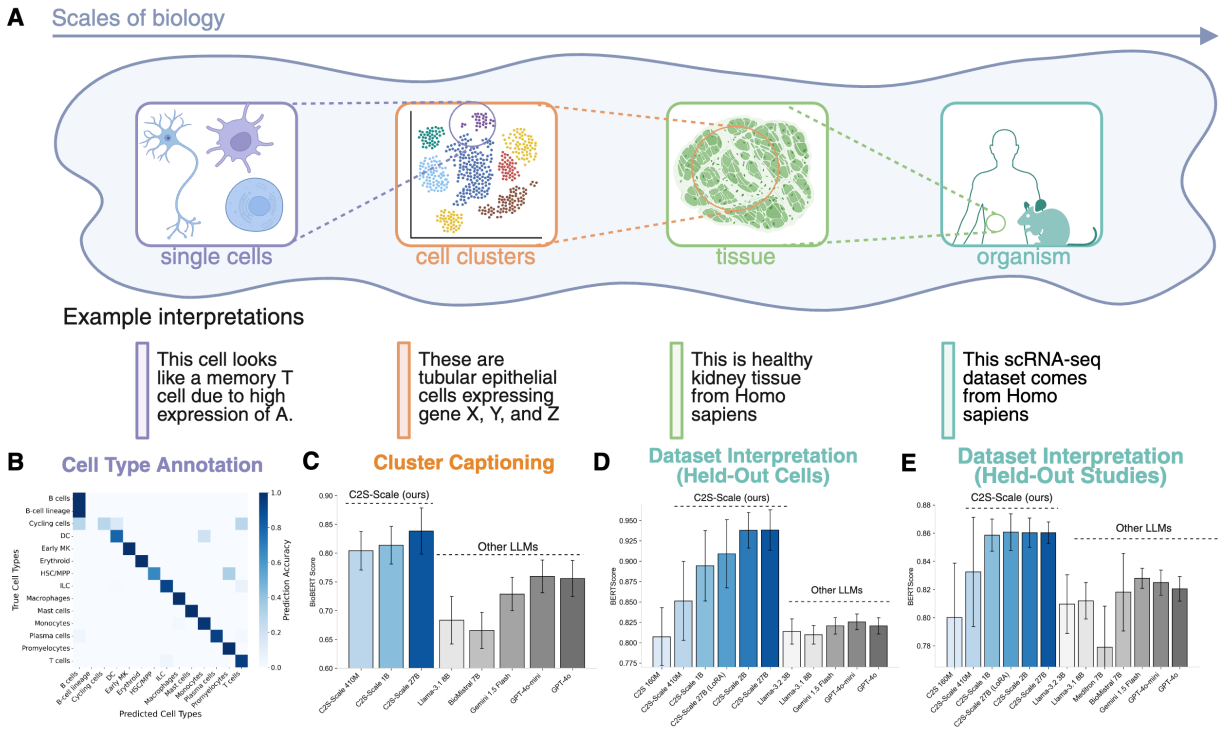

Figure 5: C2S-Scale enables natural language interpretation of scRNA-seq data at multiple scales, from single cells to entire datasets. (A) Different scales of biological data interpretation, from single cells to organism and dataset-level annotation. (B) Ground truth and predicted cell types for immune cells extracted from 16 different tissues of adult human donors [15], demonstrating the ability of C2S-Scale to annotate data at the single-cell level. (C) 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. (D-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 represent standard deviation across test set samples.

shown in Fig. 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 Fig. 11). Fig. 5D shows that C2S-Scale achieves the highest BERTScore among all evaluated models, including Llama [17, 18, 23], Meditron [24], BioMistral [25], Gemini [20], and GPT-4o [19]. Notably, C2S-Scale generalizes well to entirely unseen datasets, producing summaries that remain relevant and informative (Fig. 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 [26, 27, 28]. 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 [29, 30, 31].

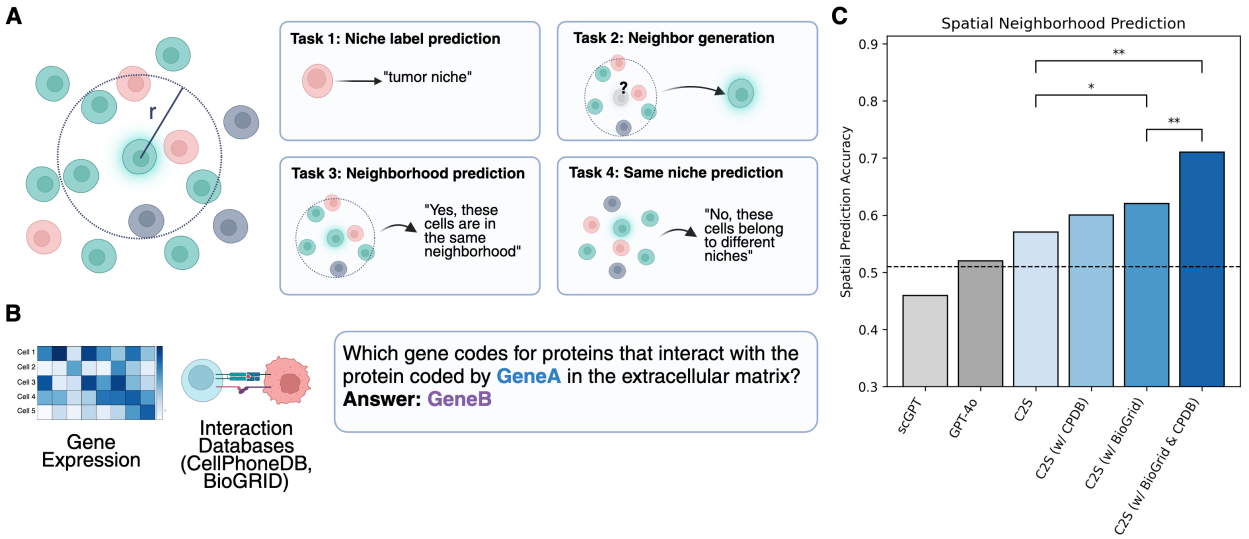

Figure 6: C2S-Scale can interpret multi-cellular spatial context and predict niche neighborhoods. (A) We fine-tune C2S-Scale on a variety of single and multi-cellular spatial tasks designed to enable C2S-Scale to perform spatial reasoning, including predicting niche labels, generating spatial neighbors, 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 (* = $P < 0.05$, ** = $P < 0.01$; McNemar's test). These results highlight the multi-task transfer learning potential of C2S-Scale for spatially-aware biological modeling.

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 [32]. 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 same niche (Fig. 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 (Fig. 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 [33] and protein interaction data from BioGRID [34], formatted as natural language prompts (Fig. 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 (Fig. 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 (Fig. 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 multi-cellular organization and interactions.

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

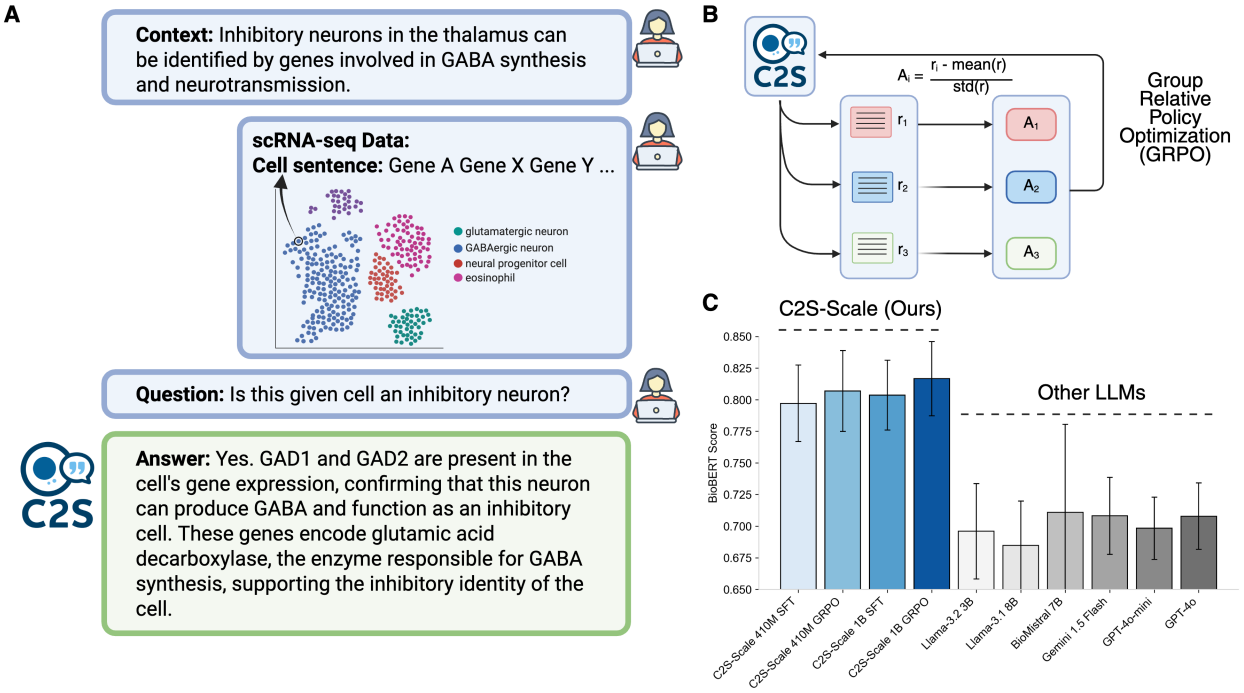

Figure 7: C2S-Scale demonstrates superior single-cell question answering performance compared to state-of-the-art (SOTA) LLMs. (A) Example QA scenario based on scRNA-seq data. (B) Overview of the GRPO framework [35], which further refines model performance by training on preference data. (C) Empirical 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.

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 measure a model's ability to understand information and apply reasoning [36, 37, 38, 39]. 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 medical [40, 41] and biological [42] 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 [19], and prompt it to generate meaningful questions (Fig. 7A).

After supervised fine-tuning (SFT), C2S-Scale surpasses the performance of state-of-the-art LLMs on scQA (Fig. 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) [43] through Group Relative Policy Optimization (GRPO) to further optimize the model to generate preferred responses to questions (Fig. 7B). By using BioBERTScore 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 generalization capabilities across unseen perturbations and cellular contexts, along with its broad applicability for modeling perturbation responses (Fig. 8A).

Training proceeds in two stages (Figure 8C): supervised fine-tuning (SFT) to predict gene-expression profiles of untreated cells—including L1000 cell lines—under specified perturbation conditions, followed by online reinforcement learning with GRPO [35] that optimizes biologically relevant objectives. We designed the reward function to prioritize the accurate prediction of key gene programs of interest. This includes apoptosis for L1000 [44] and interferon response for Dong et al. [45]. Concretely, the reward is computed over these targeted gene subsets (Figure 8F), which focuses optimization while preserving full-profile generation and improves out-of-distribution generalization (Figure 8G).

We introduce a new metric, scFID (Fig. 8B), an adaptation of the FID metric [46] widely used in computer vision to evaluate the realism of generated images. scFID adapts the FID metric by replacing the Inception-v3 model with a single-cell foundation model to embed transcriptomic data, enabling evaluation of generated cells in a representation space aligned with biological structure and functional gene programs. By assessing differences in this embedding space rather than at the level of individual genes, scFID captures higher-order variation across cell states, yielding stable model rankings (Fig. 8E) and aligning with distributional similarities evident in cell-state embeddings (Fig. 8D), while complementing expression-level metrics such as Kendall's τ and Pearson's r (Fig. 8G).

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, highlighting its ability to transfer to completely new contexts not seen during training (Fig. 8E). Compared to baselines, C2S-Scale performs best on fully unseen combinatorial perturbations, capturing nonlinear synergistic effects. Quantitative results (Fig. 8F) show superior MMD, Wasserstein, and scFID scores relative to competing models. GRPO further reduces scFID on interferon-related genes by 16%, thereby improving biological fidelity on immune pathways (Fig. 8G).

The L1000 results further underscore C2S-Scale's versatility in modeling perturbation responses across single-cell and bulk transcriptomic data. We evaluate performance on apoptosis-related genes, focusing on generalization to unseen compound treatments. Applying GRPO yields consistent gains (Fig. 8G), improving Kendall's τ by 9.2% for the 410M model and 4.9% for the 1B model, and Pearson's r by 6.6% for the 410M model and 3.6% for the 1B model. Rewards are defined on phenotype-linked gene programs (e.g., apoptosis in L1000 [44] and interferon response [45]; Fig. 8F), which yields context-aware scores well suited for virtual screening and candidate prioritization, while preserving full-profile prediction and enhancing out-of-distribution generalization (Fig. 8G).

2.8 Immune-context virtual screening reveals a cytokine-conditional amplifier of antigen presentation

A differentiating feature of C2S-Scale is its ability to connect complex transcriptional states across diverse biological contexts. To test whether C2S-Scale can uncover context-dependent determinants of immune visibility, we programmed a dual-context in-silico screen that predicts drug effects on MHC-I antigen-presentation programs in immune-context-positive versus immune-context-neutral cytokine signaling settings. Leveraging its demonstrated strength in perturbation response prediction, the model identified **silmitasertib**, a CK2 inhibitor, as one of the top candidates with a pronounced context split: a strong predicted increase in antigen-presentation programs in the immune-context-positive condition of low-level interferon (IFN) signaling (Fig. 9B; other drugs known to upregulate MHC-I highlighted in blue), but little to no effect in the neutral condition (Fig. 9C). We selected low-level IFN signaling as a tissue-specific regulator of immunity that is frequently present, but insufficient to drive maximal antigen presentation. We reasoned that enhanced antigen presentation in this context has the potential to drive increased T cell recognition, further IFN production, and positive feedback.

Our results were notable because silmitasertib has not been reported in the literature to enhance MHC-I expression, highlighting the novelty of both the effect itself and its context dependence. We confirmed that interferon response, quantified by a rank-based score for an interferon-stimulated gene set, was elevated in the immune-context-positive sample, but negligible in the neutral sample (Fig. 9D). Based on both the model's predictions and the known role of interferons in MHC-I regulation, we hypothesized that the compound acts as an interferon-conditional amplifier, lowering the response threshold to interferon rather than initiating antigen presentation de novo (Fig. 9E).

We validated this hypothesis in two human neuroendocrine cell models that were completely unseen in C2S-Scale's training data. In the first model (Merkel cell origin), silmitasertib alone did not alter HLA-A,B,C surface levels, whereas the combination of low-dose IFN-β and silmitasertib produced a marked increase in MHC-I mean fluorescence intensity

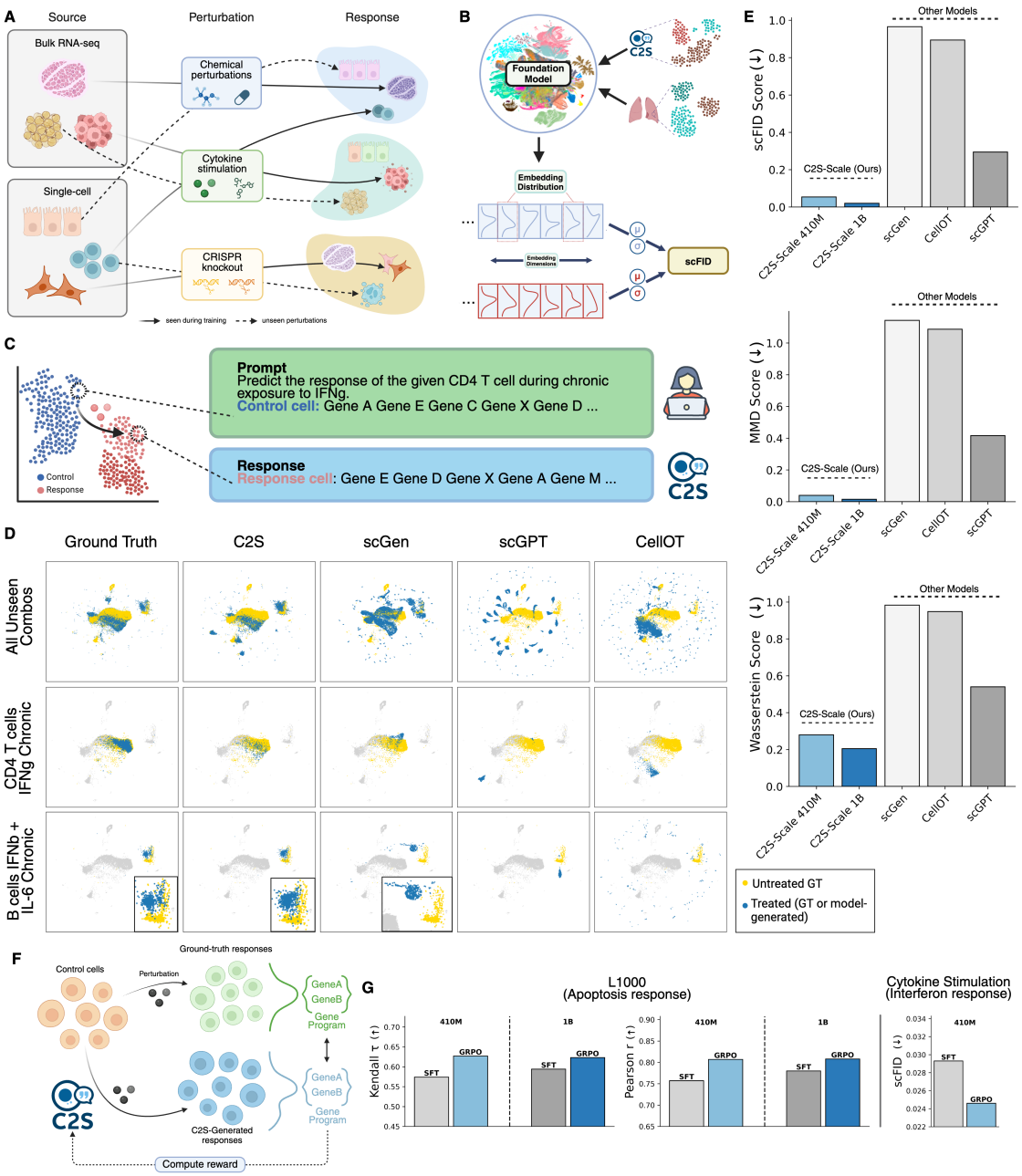

Figure 8: C2S-Scale models outperform existing methods in predicting cellular responses to unseen perturbations. (A) Overview of the C2S-Scale perturbation prediction framework, which supports diverse perturbation types including drugs, cytokines, and genetic knockouts. (B) Diagram of the scFID metric, 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.

(MFI) (Fig. 9F; 13.6% increase MHC-I MFI at 10nM and 34.9% at 1000nM). The amplification effect generalized across interferon subtypes (IFN-γ, Fig. 9G; 24.9% increase MHC-I MFI at 10nM and 37.3% at 300nM) and was reproduced in a second, independent human cell model (pulmonary origin, Fig. 9H; 17.1% increase MHC-I MFI at 10nM and 28.1% at 100nM). Notably, neuroendocrine cells were minimally represented in the training data for our model, with no representation of Merkel cells at all.

This discovery of a novel cytokine-conditional amplifier of antigen presentation demonstrates C2S's ability to perform high-throughput virtual screens to identify promising therapeutic candidates to validate experimentally. Additionally, it illustrates how C2S can reveal context-conditioned biology that is missed in context-neutral assays.

3 Discussion

Although artificial intelligence approaches including neural network models have achieved significant breakthroughs in protein structure and the prediction of molecular interactions, less progress in modeling multi-cellular tissues, pathologic states, and context-specific biology has been made. Principal challenges in this space include the underlying diversity, complexity, and pleiotropy of biological systems, which compounds across hierarchical organization from genes to transcriptional programs, and cells to tissues to organisms. Indeed, the semantic complexity and contextuality of biological systems seems unrivaled–outside of language itself. 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.

Context-specific decoding is a core task for both LLMs and biological systems alike. To test the ability of C2S-Scale to derive context-specific biological meaning, we conducted a conditional virtual screen, identifying an IFN-specific regulator of antigen presentation. We validated the effectiveness of silmitasertib in neuroendocrine Merkel cell and pulmonary cell models in which the downregulation of antigen presentation machinery is a well-established mechanism of resistance to immunotherapies. This success provides a blueprint for future screens targeting other complex biological contexts.

We anticipate that higher-capacity models and more diverse training corpora will 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 Methods

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

4.1 Data Collection

To construct the C2S-Scale pretraining corpus, we assembled over 50 million single-cell transcriptomic profiles from human and mouse tissues. Datasets were obtained from established public repositories, including the CELLxGENE [2] and Human Cell Atlas [3] data portals, and span a wide range of tissues, disease states, and experimental conditions. Each dataset was accompanied by author-provided metadata, such as cell type and tissue annotations, donor information,

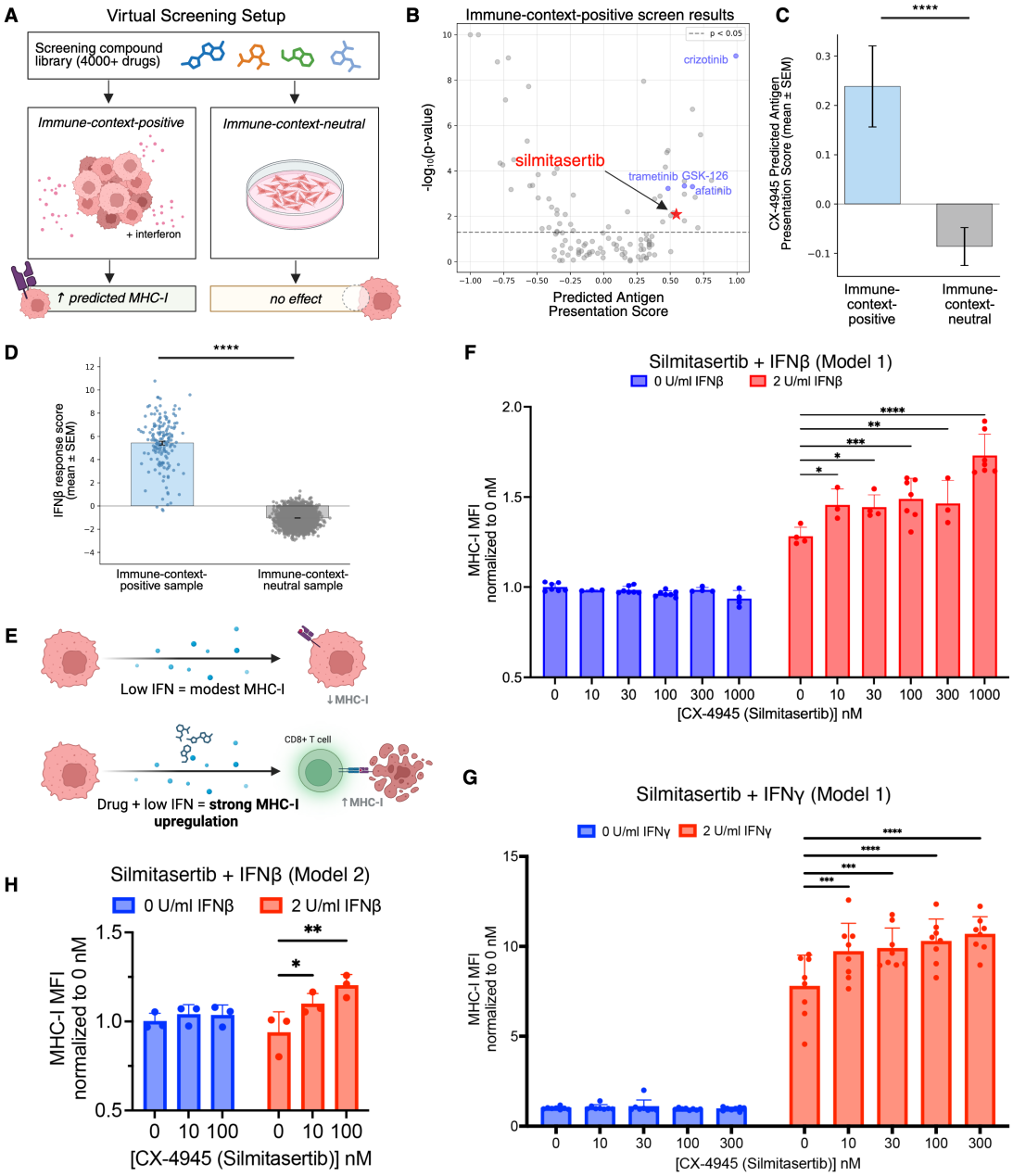

Figure 9: Immune-context virtual screening identifies a cytokine-conditional amplifier of antigen presentation. (A) Schematic of a dual-context virtual screen that predicts drug effects on immune visibility (MHC-I program) in immune-context-positive (primary human samples with endogenous interferon signaling) versus immune-context-neutral (isolated cell) settings. (B) Ranked predictions in the immune-context-positive screen nominate silmitasertib, a CK2 inhibitor, among top candidates to increase antigen-presentation programs (highlighted in red). Selected positive controls known to upregulate MHC-I are highlighted in blue. (C) Silmitasertib shows a context split, with a strong predicted effect in the immune-context-positive setting and negligible effect in the immune-context-neutral setting. (D) Interferon response was quantified by a rank-based score of a curated interferon-stimulated gene set (see Methods). Each point is one sample; bars = mean±SEM; ****, $P < 0.0001$ (Wilcoxon test). (E) Hypothesis: the compound is an interferon-conditional amplifier that lowers the response threshold for STAT1/IRF1 and thereby amplifies MHC-I upregulation. (F) Experimental validation in an unseen cell type shows no effect of CK2 inhibition alone and marked HLA-A,B,C upregulation in the presence of low-dose IFN-β (n=3 independent experiments; points = replicates; bars = mean±SD.; two-sided tests with multiple-comparison correction). (G) The amplification holds with IFN-γ, indicating robustness across interferon subtypes. (H) The same interferon-conditional amplification is observed in a second, independent human cell model, supporting generality.

developmental stage, and associated study identifiers. Where available, supplementary textual resources, including paper abstracts and study descriptions, were also retained.

Raw scRNA-seq data were processed using standard preprocessing pipelines, including quality control, library size normalization, and log-transformation, following established conventions [47]. For each dataset, the transcriptomic profiles were converted into cell sentences, and the accompanying annotations were preserved to construct natural language prompts. This resulted in a multimodal corpus linking expression profiles with textual descriptors of biological context. A complete list of datasets included in the corpus is provided in Supplementary Table 1.

4.2 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

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

The gene names are in natural language, forming a sentence interpretable by language models (exemplified in Fig. 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 pretraining on natural language data, which has been shown to be significant for large-scale pretrained LLMs [21].

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 Fig. 10 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 $R^2 = 0.85$, 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.

4.2.1 Multi-Task Prompt Formatting.

C2S-Scale was designed to operate in natural language, enabling a broad range of predictive and generative tasks in single-cell analysis. These tasks include cell type and tissue annotation, multi-cell generation, and dataset-level interpretation. The complete list of pretraining tasks, together with their inputs and outputs, is provided in Table 1.

Prompts were constructed by combining the cell sentence representation of one or more cells with task-specific natural language instructions. For predictive tasks, the input prompt included a cell sentence and an instruction, and the output corresponded to the metadata label of interest. For example, in the cell type annotation task, the input consisted of the cell sentence and the instruction "Predict the cell type of this cell", and the output was the corresponding cell type label. For generative tasks, this structure was inverted: metadata conditions were provided in the input prompt, and the model was trained to generate one or more cell sentences in response.

Metadata included in natural language prompts encompassed cell type, tissue annotations, perturbation conditions, disease states, and text from associated studies or abstracts, thereby providing additional biological context. This framework enables C2S-Scale to interpret instructions, integrate biological knowledge, and generalize across diverse applications.

4.3 C2S-Scale architecture and pretraining

4.3.1 Input representation

C2S-Scale employs large language models (LLMs) based on the Transformer architecture [8] to model cell sentences in natural language. Input sequences are represented as high-dimensional embeddings suitable for processing by neural networks. Each word in a cell sentence corresponds to a gene name, which is first tokenized using the pretrained

tokenizer associated with the backbone model. This approach avoids the introduction of new vocabulary and maintains compatibility with the LLM's pretraining knowledge.

Tokenized gene names are mapped into vector representations through an embedding layer trained alongside the model. These embeddings capture semantic properties of genes informed both by their biological context and by the pretrained model's prior knowledge. Positional encodings are added to preserve the rank order of genes within each cell sentence, allowing the model to learn dependencies across expression-ranked sequences.

4.3.2 Attention mechanism

The central component of the Transformer is the self-attention mechanism [48, 8], which enables the model to compute pairwise relationships between tokens. For single-cell tasks, this allows the model to dynamically prioritize genes that are most informative for a given context, such as lineage-defining markers for classification or perturbation-responsive genes for prediction. The attention mechanism also extends naturally to metadata tokens (e.g. cell type, tissue, disease state), enabling the model to integrate gene expression with contextual information in a shared representation.

4.3.3 Model architecture

C2S-Scale adopts a decoder-only Transformer design [19], chosen for its capacity to model sequential data and support generative tasks. The architecture consists of a stack of Transformer blocks, each containing a multi-head self-attention layer followed by a position-wise feedforward network. Residual connections and layer normalization are applied throughout to stabilize optimization and facilitate scaling to billions of parameters. This modular structure allows the model to capture long-range dependencies in gene expression data while remaining computationally efficient.

4.3.4 Pretraining objective

The model is pretrained with a next-token prediction objective [49], in which the model learns to predict the next token in a sequence given all preceding tokens. Applied to cell sentences, this involves predicting the next gene in the rank-ordered expression list, optionally conditioned on metadata tokens. This autoregressive formulation encourages the model to capture the hierarchical organization of gene expression programs and to integrate biological context during generation.

In contrast to masked-token objectives such as those used in Geneformer [5], which predict randomly masked genes in non-linguistic sequences, the autoregressive objective aligns naturally with downstream generative applications. Training the model in this way conditions it to produce coherent, biologically meaningful outputs for tasks such as cell generation, dataset-level interpretation, and question answering.

4.3.5 Training Setup

Pretraining was carried out on the C2S-Scale corpus of more than 50 million single-cell transcriptomes with associated metadata and textual annotations. A multi-task learning framework was used to jointly optimize across the pretraining tasks described in Table 1, enabling the model to integrate transcriptomic and contextual information.

The C2S-Scale 410M-parameter and 1B-parameter models were trained on one Nvidia A100/H100 GPU with the Transformers library (version 4.46.3) [50] and PyTorch (version 2.4.1) [51] on a High Performance Computing (HPC) cluster running Red Hat Enterprise Linux release 8.10. Models larger than 1B parameters were trained on 256 TPU v4s using the Jax library. We used a starting learning rate of 1e-5 with linear decay and weight decay of 0.01.

4.4 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 [52] and Pythia [53] 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, dataset interpretation, and conditional sample generation tasks. Both fully fine-tuned and LoRA fine-tuned variants [54] were evaluated to assess scaling behavior under different computational budgets.

Performance was measured using BERTScore [22] between generated and reference outputs for predictive tasks such as cell type annotation and dataset interpretation, providing a semantic measure of response quality. Let the reference output be $x = \langle x_1, \ldots, x_k \rangle$ and the generated output be $\hat{x} = \langle \hat{x}_1, \ldots, \hat{x}_l \rangle$, where tokens are represented by contextual embeddings. Pairwise similarity between tokens is given by the cosine similarity $s(x_i, \hat{x}_j) = \frac{x_i^\top \hat{x}_j}{\|x_i\| \|\hat{x}_j\|}$. BERTScore recall, precision, and F1 are then defined as

Table 1: Pretraining 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

$$R_{\text{BERT}} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} s(x_i, \hat{x}_j), \quad P_{\text{BERT}} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} s(x_i, \hat{x}_j),$$

$$F_{\text{BERT}} = \frac{2\, P_{\text{BERT}}\, R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}}.$$

This formulation captures semantic similarity even when exact lexical matches are absent. Unless otherwise noted, all reported BERTScore values correspond to the F1 variant.

For generative tasks such as conditional cell generation, we evaluated outputs by measuring gene overlap between generated and target cell sentences. This metric captures the proportion of ground truth genes recovered in the generated output, providing a direct measure of transcriptomic fidelity. Let G_{ref} denote the set of genes in the reference cell sentence and G_{gen} the set of genes in the generated cell sentence. Gene overlap is defined as

$$\text{Overlap}(G_{\text{gen}}, G_{\text{ref}}) \;=\; \frac{|\, G_{\text{gen}} \cap G_{\text{ref}} \,|}{|\, G_{\text{ref}} \,|}.$$

4.5 Post-training methods

4.5.1 Supervised fine-tuning

C2S-Scale was adapted to downstream applications through supervised fine-tuning on labeled datasets. Fine-tuning used the same autoregressive next-token prediction objective as pretraining, with prompts formatted to match each task. For example, a prompt might consist of a cell sentence followed by the instruction "Predict the tissue of origin for this cell:", and the model was trained to output the corresponding metadata label.

Parameter-efficient strategies were used to limit overfitting and reduce compute cost. Low-Rank Adaptation (LoRA) and lightweight adapter layers updated only a small subset of parameters, while the majority of pretrained weights remained frozen. This design allowed rapid task-specific adaptation with modest data requirements.

4.5.2 Reinforcement learning alignment

Reinforcement learning (RL) was used to further align model outputs with biological accuracy and interpretability. We employed Group Relative Policy Optimization (GRPO), a policy-gradient method that incorporates task-specific reward signals directly into parameter updates [43, 35].

The supervised fine-tuned model (policy π_θ) generated multiple candidate outputs $o = (o_1, \ldots, o_{|o|})$ for each input prompt q. Each token o_t was assigned probability $\pi_\theta(o_t \mid q, o_{<t})$, where $o_{<t}$ denotes the prefix. Rewards r_i were assigned to each candidate sequence o_i using automated evaluation metrics such as BERTScore [22] and domain-specific scores for tasks like perturbation response prediction.

Proximal Policy Optimization (PPO) maximizes a clipped surrogate objective, which requires estimating per-token advantages A_t using a value function:

$$J_{\text{PPO}}(\theta) = \mathbb{E}_{q\sim P(Q),\, o\sim\pi_{\theta_{\text{old}}}}\left[\frac{1}{|o|}\sum_{t=1}^{|o|}\min\left(\frac{\pi_\theta(o_t \mid q, o_{<t})}{\pi_{\theta_{\text{old}}}(o_t \mid q, o_{<t})}A_t,\ \text{clip}\left(\frac{\pi_\theta(o_t \mid q, o_{<t})}{\pi_{\theta_{\text{old}}}(o_t \mid q, o_{<t})}, 1-\epsilon, 1+\epsilon\right)A_t\right)\right],$$

where $\pi_{\theta_{\text{old}}}$ is the policy from the previous iteration, A_t is the advantage at step t, and ϵ is the clipping threshold. Maintaining a critic to estimate A_t increases computational cost and can destabilize training.

GRPO replaces the value function with a group-relative baseline. For each prompt q, the model samples G candidate outputs $\{o_1, \ldots, o_G\}$ with associated rewards $\{r_1, \ldots, r_G\}$. Relative advantages are defined by normalizing rewards across the group:

$$\tilde{r}_i = \frac{r_i - \text{mean}(r)}{\text{std}(r)}, \qquad \hat{A}_{i,t} = \tilde{r}_i \quad \forall t \in o_i.$$

The GRPO objective is

$$J_{\text{GRPO}}(\theta) = \mathbb{E}_{q,\{o_i\}_{i=1}^G}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left(\min\left(\frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})}\hat{A}_{i,t},\right.\right.\right.$$
$$\left.\left.\left.\text{clip}\left(\frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})}, 1-\epsilon, 1+\epsilon\right)\hat{A}_{i,t}\right) - \beta D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})\right)\right],$$

where π_{ref} is the frozen SFT model and β controls the KL regularization strength.

GRPO eliminates the critic network, reduces memory requirements, and yields stable optimization at scale. When trained with biologically relevant reward functions, C2S-Scale refined its predictions and aligned generative behavior with biological ground truth.

4.6 Downstream Tasks

4.6.1 Cell type annotation

For the cell type annotation task, we fine-tuned the model to predict cell type labels on an immune tissue dataset [55], pancreas dataset [56], and a lung dataset [16]. We used 80% of cells from each dataset for training and reserved 20% for evaluation. C2S-Scale was provided with a cell sentence and a natural language prompt, such as “Predict the cell type of this cell:”. C2S-Scale was fine-tuned for this task using the same next-token prediction objective [49] as the pretraining step, predicting cell type labels in natural language. Other scFMs were fine-tuned using prediction heads on top of the pretrained transformer weights in accordance with the recommended strategies for each model.

4.6.2 Cell generation

For cell generation tasks, we fine-tuned the model to unconditionally or conditionally generate cell expression on the immune tissue and lung datasets. The model was given a natural language prompt containing relevant metadata for conditional generation, or no information in the case of unconditional generation, and was 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."

4.6.3 Cell embedding

For cell embedding, we used 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 formatted input prompts for C2S-Scale in the same manner as in cell type prediction tasks. However, instead of decoding token predictions, we took the last hidden state from the last layer of the C2S-Scale model, and average pooled 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.

4.6.4 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 designed a simple single-cell and bulk RNA seq integration task. Using a single-cell lung tissue dataset [16], we constructed pseudo-bulk samples by aggregating over donor, cell type, and batch. For each pseudo-bulk sample, we randomly sampled ten single-cell samples from the same conditions to construct pairs. We embedded each single-cell and pseudo-bulk sample individually using each model and computed the cosine similarity between the paired single-cell and bulk samples. Following [57], we used 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.

4.6.5 Cluster captioning

To generate the cluster captioning dataset, we selected 30 scRNA-seq datasets and performed standard preprocessing, clustering, and differential expression analysis. We then prompted GPT-4o [19] 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 sampled 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.

4.6.6 Dataset interpretation

For the dataset-level interpretation task, we created 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) an 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, this serves as 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 Fig. 11.

To create the training corpus distribution dataset interpretation test set, we first gathered 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 one abstract, the held-out samples will still contain similar information to training set abstract generation samples that the model has seen. We sampled 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 incorporating 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 [56] and a human retina [58] 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.

4.6.7 Spatial niche prediction

For the spatial niche prediction task, we used the CosMx Spatial Molecular Imager Human Liver dataset [32], 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 180 mm^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 facilitate spatial analyses. We define a neighborhood to be a radius of 0.02 pixels (approximately 20 μm), 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:** Given 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. The data contained 19,754 training samples and 3,968 test samples.

Additionally, to enhance the model's understanding of cell communication, we included gene interaction metadata from CellPhoneDB [33] and BioGRID [34]. We restricted the data to only retain interactions involving the 1,000 genes in the CosMx data, and also only to genes coding for extracellular proteins (determined using MatrixDB [59]). We included 5,822 interaction samples from CPDB and 2,334 from BioGRID.

Models were evaluated on a held-out test set comprising 3,968 samples. Performance was measured as mean prediction accuracy across the spatial neighborhood prediction tasks. To compare models, paired differences in prediction outcomes were assessed using McNemar's test with continuity correction, which evaluates whether two classifiers differ significantly in their error distributions when applied to the same test set. Significance was reported as p-values from McNemar's test, with values below 0.05 considered statistically significant.

4.6.8 Question answering

We used 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 contributed 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 applied GRPO to further refine answer quality. To create the GRPO training set, we collected an additional 600 samples from unseen studies, with each sample prompting the SFT model to generate 32 candidate answers. We then used 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 served 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 evaluated 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 Fig. 7.

4.6.9 Perturbation prediction

The Dong et al. dataset [45] dataset includes immune cells exposed to individual and combinatorial cytokines, with each cell annotated by type, stimulation, and exposure length – yielding 133 conditions. We retained the 5000 most highly variable genes and evaluated models in the scGPT embedding space [4] using maximum mean discrepancy (MMD), Wasserstein distance, and scFID (Section 4.7). 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 [60] 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 [61] 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 [44], 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 τ 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 4 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.

For evaluation, we computed metrics differently across datasets. For the Dong et al. [45] dataset, we computed maximum mean discrepancy (MMD), Wasserstein distance, and scFID for each unique combination of condition labels (cell type, cytokine, and exposure duration), and averaged these values across all combinations to obtain the final metric. For the L1000 dataset [44], we computed Pearson's r against the Level 3 gene expression values and Kendall's τ on the ranks of the gene expression values for each test sample individually and then reported the average across all samples.

Kendall's τ measures rank correlation between two ordered lists. Given n genes, we consider all $\frac{1}{2}n(n-1)$ possible gene pairs. For any pair of genes (i, j), if their relative order (which gene is ranked higher) is the same in both the generated output and the ground-truth ranking, the pair is *concordant*; if their relative order is reversed, the pair is *discordant*. Tied pairs (where the genes share the same rank in either list) are handled by assigning them the same value. Kendall's τ is then defined as

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)},$$

where n_c and n_d denote the number of concordant and discordant pairs, respectively. In our application, the ranks of the 978 L1000 landmark genes are derived from the generated output of the model, where the cell sentence places genes in descending expression order (e.g., `GeneT GeneA GeneS GeneW` ...). Genes not present in the model's output are assumed to share the lowest possible rank (e.g., if 950 genes are generated, the remaining 28 share rank 951). The same ranking convention is applied to the L1000 ground-truth sample, where unexpressed genes also share the last rank. Kendall's τ is then computed between these two ranked lists, yielding a rank-based correlation that is robust to tied ranks and sparse expression. 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.

4.7 Single-Cell Fréchet Inception Distance

The scFID is an adaptation of the FID [46] tailored for evaluating generative models in single-cell transcriptomics. While the traditional FID employs the Inception v3 model [62] 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:

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

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 computed the scFID for each unique combination of test labels—such as specific cell types, perturbations, and exposure durations—and then averaged these individual scFID values.

4.8 Virtual Screen Setup

Datasets We analyzed drug responses in both primary tumor samples and an immortalized cell line in order to capture effects across distinct immune environments. The immune-context-positive data comprised bulk RNA-seq from a pan-cancer atlas [63], which includes 364 tumor specimens spanning 12 cancer types. Cells were sorted by flow cytometry, and we restricted our analysis to the "tumor" compartment, yielding 162 bulk samples. As an immune-context-neutral system, we used the Merkel cell WAGA cell line, as it was not part of the training data for the model. We obtained data from GEO [64], containing 4,199 cells. For the single-cell data, standard preprocessing was applied, including removal of genes expressed in fewer than three cells, removal of cells with fewer than 50 counts, normalization to a total count of 10^4 per cell, and `log1p` transformation.

To quantify type I interferon activity across bulk tumour samples and single cells, we computed a rank-based analytical z-score for a curated interferon-stimulated gene (ISG) set. For each expression profile, all detected genes were ranked by expression level. The mean rank of the ISG set was then compared to the null expectation of randomly distributed ranks using a two-sample Wilcoxon test.

Compound library The screening library was derived from the L1000 resource, which catalogs over 30,000 small molecules. Because our goal was to prioritize compounds that could feasibly be validated, we filtered this set using GPT-o3 to predict commercial availability. This step produced a working library of 4,266 drugs.

Perturbation inference Drug perturbations were simulated using our C2S-Scale perturbation response prediction model. Each bulk tumor sample was perturbed *in silico* with every drug in the library three times, for a total of $N = 486$ samples per drug. For the WAGA cell line, 20 representative cells were each perturbed 20 times with every drug for a total of $N = 400$ samples per drug. Replicates corresponded to independent forward passes through the model, with stochastic sampling at a temperature of 0.3 to introduce variability across predictions.

Scoring of antigen-presentation programs Antigen-presentation activity was quantified by calculating enrichment scores for each perturbed profile. We applied single-sample gene set enrichment analysis (ssGSEA) with the "Class I MHC mediated antigen processing and presentation" gene set from MSigDB [60], using the Python package `gseapy` (v1.1.8) with parameters `sample_norm_method='rank'` and `weight=0`. Scores were aggregated across replicates for each drug and normalized to the interval $[-1, 1]$. As a complementary metric, we also computed the average log-fold change for HLA-A,B,C, which produced results consistent with ssGSEA (Supplementary Fig. 12).

Top-ranked drugs were examined for prior evidence of involvement in antigen-presentation pathways. Manual inspection was used to flag compounds not previously reported in the literature, and these were prioritized for further analysis.

4.9 Experimental Validation of Interferon-Conditional Effects

To validate the interferon-conditional effects predicted in silico, we performed experiments in two tumor-derived cell lines: MDK-knockout WAGA (Merkel cell carcinoma, MCC) and DMS153 (small cell lung cancer, SCLC). Cells (600,000–2,500,000 cells/ml) were treated with Silmitasertib at the indicated concentrations for 24 hours, followed by

stimulation with 2 U/ml human IFNβ (PBL Assay Science, cat. #11415) or 2 U/ml human IFNγ (PBL Assay Science, cat. #11500) for an additional 24 hours. In parallel, dose–response assays were performed by titrating IFNβ across a range of 0.5–200 U/ml to characterize sensitivity to interferon signaling.

After treatment, cells were harvested and stained for surface expression of major histocompatibility complex class I molecules HLA-A,B,C (clone W6/32, BioLegend). Live tumor cells were gated using Zombie Aqua fixable viability dye (BioLegend) to exclude dead cells prior to analysis by flow cytometry using the CytoFLEX S running CytExpert 2.4 (all Beckman Coulter). All assays were performed in three independent biological replicates. For statistical comparisons, a two-way Brown–Forsythe and Welch ANOVA was applied, followed by Dunnett's T3 correction for multiple testing.

4.10 Data Availability

A list of HCA and CELLxGENE datasets used for pretraining is provided in Supplementary Table 1. Spatial transcriptomic data for the niche prediction task was obtained from CosMx [32]. Publicly available interaction databases were acquired from [33, 34, 59]. For the perturbation prediction task we used transcriptomic data from L1000 [44] and from [45]. For the virtual screen we used primary tumor data from [63] and cell line data from [64]. Model weights are available on Hugging Face.

4.11 Code Availability

Code for model training is publicly available at: https://github.com/vandijklab/cell2sentence

5 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.

6 Author Contributions

Project lead: S.A. Rizvi. Model training was done by D.L., A.P., S.Z., E.W., and B.P. Scaling evaluations were done by S.A.R., and benchmarking across predictive and generative single-cell tasks was done by S.A.R. and A.P. Perturbation prediction evaluations and scFID implementation were done by D.L. Question answering evaluations were done by S.Z., and wet-lab validation experiments were done by C.J.P. and N.M.C. Data curation was done by S.H., D.Z., Z.L., C.L., E.S., D.J., and L.Z. Reviewing and editing was done with guidance from C.T., J.K., D.B., B.H., R.D., H.C., R.M.D., B.P., J.I., S.Z., and D.v.D. All of the authors reviewed the manuscript.

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] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

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

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

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

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

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

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

[37] 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.

[38] 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.

[39] 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.

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

[41] 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.

[42] 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.

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

[44] 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.

[45] 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.

[46] 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.

[47] 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.

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

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

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

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

[61] 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.

[62] 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.

[63] Alexis J Combes, Bushra Samad, Jessica Tsui, Nayvin W Chew, Peter Yan, Gabriella C Reeder, Divyashree Kushnoor, Alan Shen, Brittany Davidson, Andrea J Barczak, et al. Discovering dominant tumor immune archetypes in a pan-cancer census. *Cell*, 185(1):184–203, 2022.

[64] K. Fan, J. Becker, and J. Gravemeyer. Waga single cell rna sequencing. Gene Expression Omnibus, NCBI, GEO Accession: GSE130346, 2019. Homo sapiens, Expression profiling by high throughput sequencing. BioProject: PRJNA535920. Accessed: 2025-08-14.

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

[66] 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.

[67] 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.

[68] 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.

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

[70] 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.

7 Supplementary

7.1 Limitations

7.1.1 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 paradigms from vision-language models, where arbitrary tokenization orders paired with causal attention still achieve state-of-the-art performance [65]. 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 [66, 67, 68]. 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) [69, 70]. 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 pretrained 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.

7.1.2 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.

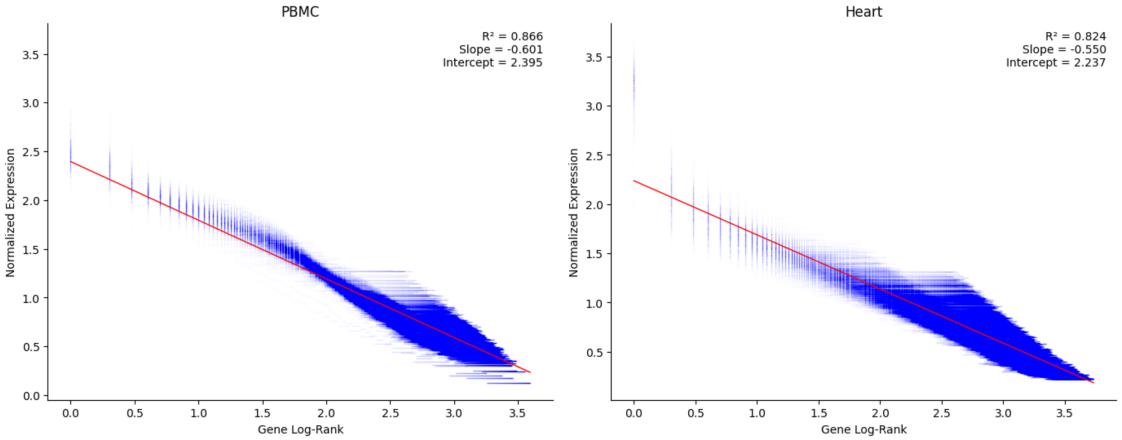

Figure 10: C2S allows for conversion from expression information into cell sentence format with minimal information loss. Using a linear model fitted between rank and original expression, cell sentences can be converted back to expression accurately.

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 11: Example abstract summaries from scRNA-seq datasets collected from CELLxGENE [2].

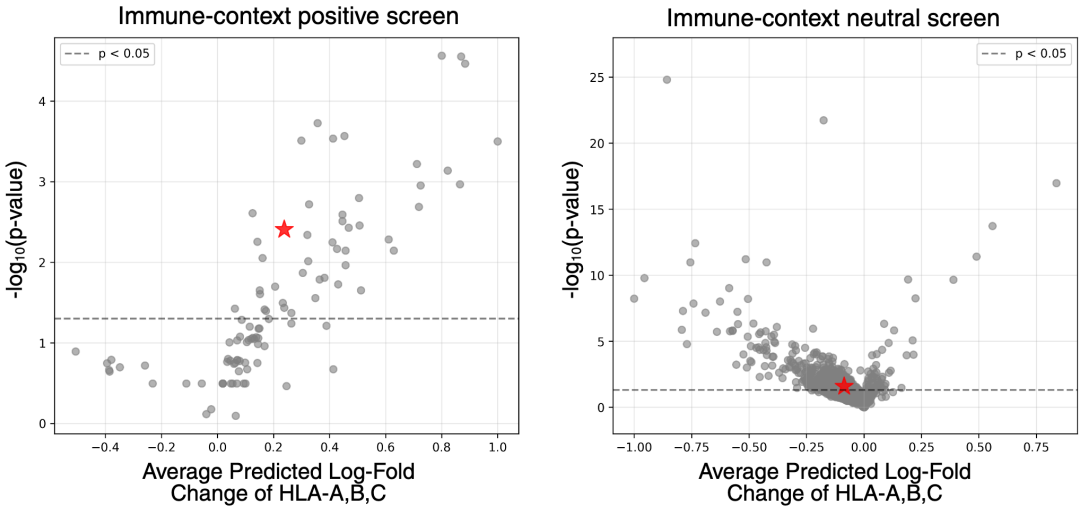

Figure 12: Predicted effects of silmitasertib on MHC-I antigen presentation in immune-context-positive (left) and immune-context-neutral (right) screens. Each point represents a compound, plotted by the average predicted log-fold change of HLA-A,B,C versus the corresponding significance level. Silmitasertib is highlighted in red. Results are consistent with the primary scoring approach using the antigen-presentation gene set (see Methods).