SINGLE-CELL FOUNDATION MODELS AS MODALITY TRANSLATORS: BRIDGING GENE PROGRAMS AND NATURAL LANGUAGE

Anonymous authors
Paper under double-blind review

ABSTRACT

Single-cell foundation models effectively capture gene expression patterns and cellular heterogeneity to support downstream analysis, but struggle to incorporate broader knowledge. Large language models possess such knowledge, yet cannot directly interpret single-cell expression data. Connecting these complementary capabilities often requires specialized model designs or modifications to existing frameworks. We introduce **CellDialect**, a framework that repurposes pretrained single-cell foundation models as modality translators. **CellDialect** converts expression profiles into cell-conditioned continuous tokens compatible with a frozen language model, without pretraining a new language backbone or changing its architecture. Joint numerical and language supervision supports cell representation learning, question answering, descriptive text generation, and expression generation conditioned on cellular metadata. Benchmark evaluations demonstrate improved representation quality over the evaluated baselines and competitive zero-shot cell task question answering. **CellDialect** connects existing single-cell and language foundation models within a shared framework. It offers a path to supporting diverse single-cell tasks and improving existing models with modest additional pretraining.

1 INTRODUCTION

Single-cell RNA sequencing reveals cellular diversity across tissues, development, and disease, while expanding cell atlases connect molecular measurements to biological organization (Elmentaite et al., 2022; Gulati et al., 2025). Extracting reusable knowledge from these resources requires methods that handle sparse expression measurements, technical variation, and differences between datasets (Luecken & Theis, 2019; Argelaguet et al., 2021). Representation learning and large-scale pretraining address this need by capturing transferable structure across diverse cell populations (Eraslan et al., 2019; Baek et al., 2025).

Numerical representations alone, however, do not provide a natural interface for interpreting cellular identity and context. Researchers reason through biological concepts, descriptions, and questions. Pretrained language models offer a complementary interface and knowledge acquired from large text corpora, motivating their integration with cellular measurements (Min et al., 2023; Ashyrmamatov et al., 2026). Such integration should support both language-based interaction and numerical prediction, whose distinct assumptions require task-specific evaluation (Dimitrov et al., 2026).

A central question is therefore how to introduce cellular omics into a general-purpose language-model framework efficiently while keeping its parameters frozen and architecture unchanged. A useful framework should retain transferable cellular structure, connect expression profiles with biological language, and produce numerical outputs from biological conditions, without requiring adaptation to each evaluation dataset.

We introduce CellDialect, a framework that integrates cellular omics into a frozen language model through a continuous cellular interface. Pretrained single-cell foundation models serve as modality translators, enabling cell representation learning, question answering, descriptive text generation,

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

and metadata-conditioned expression generation within a shared training framework. Zero-shot evaluation on held-out datasets shows improved representation quality over the evaluated baselines and competitive cell-type question answering. By reusing pretrained cellular and language models while keeping the language backbone frozen and its architecture unchanged, CellDialect provides an efficient, reusable paradigm for connecting cellular omics with general-purpose language models without pretraining a new language backbone.

2 Related work

2.1 Foundation Models for Single-Cell Omics

Single-cell Omics foundation models use large collections of molecular profiles to learn transferable representations of gene relationships and cellular heterogeneity. An early example, scBERT, combined gene embeddings with discretized expression embeddings and used masked pretraining on approximately 1.1 million cells (Yang et al., 2022). Geneformer subsequently scaled pretraining to approximately 30 million transcriptomes, representing each cell as a sequence of gene-identity tokens ordered by normalized expression and learning through masked gene prediction (Theodoris et al., 2023). scGPT combined gene-identity and expression-value embeddings with a generative attention-masking scheme for learning cellular context and predicting expression (Cui et al., 2024). scFoundation further introduced read-depth-aware pretraining and an asymmetric encoder–decoder, scaling to 100 million parameters and over 50 million cells (Hao et al., 2024). GeneCompass incorporated regulatory networks, promoter sequences, gene-family annotations, and co-expression relationships as prior knowledge in cross-species pretraining (Yang et al., 2024).

The scope of cellular pretraining has also expanded beyond human transcriptomes. EpiAgent extended this paradigm to single-cell epigenomics, tokenizing accessible candidate cis-regulatory elements and learning through cell–element alignment and accessibility-signal reconstruction (Chen et al., 2025). TranscriptFormer used protein-language-model embeddings to represent genes across species and autoregressively modeled gene identities and expression counts (Pearce et al., 2026). UCE likewise used protein-derived gene embeddings, but learned cellular representations through expression-based gene sampling and prediction of gene expression status (Rosen et al., 2026).

2.2 Multimodal cellular foundation models

Multimodal cellular models add metadata and language to broaden downstream applications through three main routes.

Joint expression–metadata modeling. scMulan trains a domain-specific generative model on structured cell sentences containing genes, expression values, metadata, and task indicators; entity and value prediction support annotation and conditional generation (Bian et al., 2024).

Language-model adaptation. Cell2Sentence (C2S) represents cells as expression-ranked gene-name sequences and adapts pretrained language models through next-token prediction (Levine et al., 2024). C2S-Scale extends this approach with larger backbones and multitask training on cell sentences, instructions, and metadata (Rizvi et al., 2025).

Cross-modal alignment and connection. LangCell aligns cell and text encoders through contrastive learning and cell–text matching for text-based identity prediction (Zhao et al., 2024). CellWhisperer contrastively aligns transcriptome and text embeddings and connects cellular representations to a conversational language model through an adapter (Schaefer et al., 2025). Cell2Text projects contextualized Geneformer embeddings into an instruction-tuned language model for description generation (Kharouiche et al., 2025).

These approaches connect cellular measurements with semantic information through domain-specific models or adapted language components. Training from scratch does not directly inherit a general-purpose language backbone, whereas fine-tuning or internal adaptation can change its pretrained parameters or computation. CellDialect keeps the language backbone frozen and its architecture unchanged, using a single-cell foundation model as a modality translator to connect cellular omics with general-purpose language knowledge.

3 METHODS

CellDialect integrates cellular omics into a general-purpose language model while keeping its backbone frozen and architecture unchanged. A pretrained single-cell foundation model translates expression profiles into continuous cell token embeddings, which serve as a cellular prompt (Figure 1). A dedicated representation token aggregates this information for numerical prediction, while the language model generates textual outputs. Training combines *omics signal self-prediction* and *autoregressive text generation*, with auxiliary regularization and conditional reconstruction. This shared interface provides an efficient, reusable paradigm for connecting cellular omics with general-purpose language models.

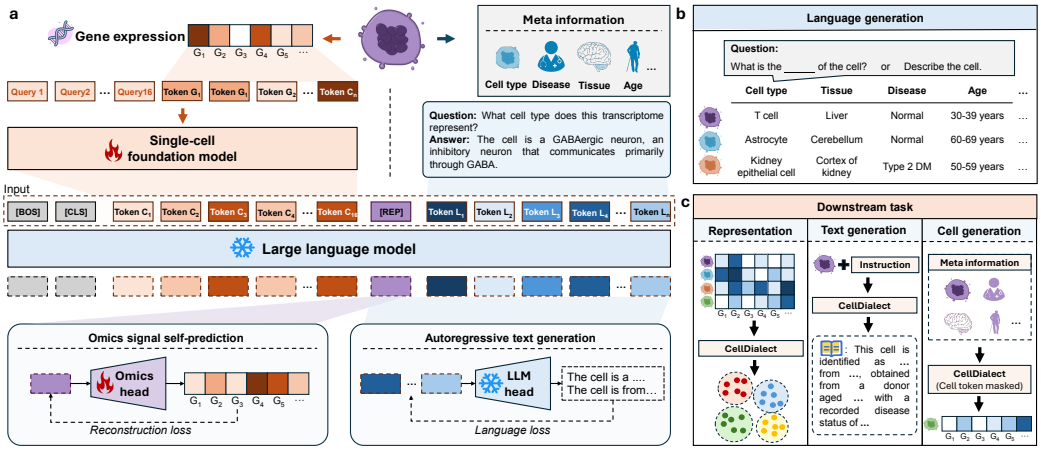

Figure 1: Overview of the proposed framework. **(a)** The work flow for CellDialect, integrating cellular omics data into large language models. **(b)** Examples of text generation. **(c)** Downstream task of CellDialect.

3.1 MODEL ARCHITECTURE

CellDialect comprises three modules: a single-cell foundation model that translates expression profiles into cell token embeddings, a frozen large language model that processes these embeddings together with text, and a signal-recovery module that predicts expression from the resulting cell representation.

Single-cell foundation model. We adopt scGPT (Cui et al., 2024), a representative foundation model for scRNA-seq analysis, and use its continually pretrained variant, scGPT-CP. Its encoder comprises 12 layers and 8 attention heads, with hidden dimension $d_c = 512$. For cell i, let g_{ij} and v_{ij} denote the identity and binned expression value of input gene j. The corresponding embedding is $\boldsymbol{e}_{ij} = E_g(g_{ij}) + E_v(v_{ij})$. We replace the original single classification position with K learnable query positions:

$$\begin{aligned} ([\mathrm{CLS}], g_{i1}, \ldots, g_{in_i}) &\longrightarrow (q_1, \ldots, q_K, g_{i1}, \ldots, g_{in_i}), \\ \boldsymbol{E}_i &= [\boldsymbol{q}_1 + \boldsymbol{\iota}_1; \ldots; \boldsymbol{q}_K + \boldsymbol{\iota}_K; \boldsymbol{e}_{i1}; \ldots; \boldsymbol{e}_{in_i}]. \end{aligned} \tag{1}$$

Here $\boldsymbol{q}_k$ and $\boldsymbol{\iota}_k$ are rows of two trainable arrays $Q, I \in \mathbb{R}^{K \times d_c}$ shared across cells. The first line denotes token positions; the second gives their actual input embeddings. Queries attend jointly with valid gene positions, while padding is masked. The default is $K = 16$, with 51 expression bins and a maximum input length of 1,200 positions including queries.

The query outputs are normalized and projected by a shared MLP P_η into the language embedding space:

$$\begin{aligned} \widetilde{H}_i &= F_\theta(\boldsymbol{E}_i)_{1:K}, \qquad \boldsymbol{h}_{ik} = \frac{\widetilde{\boldsymbol{h}}_{ik}}{\max(\|\widetilde{\boldsymbol{h}}_{ik}\|_2, \epsilon)}, \\ \boldsymbol{z}_{ik} &= P_\eta(\boldsymbol{h}_{ik}), \qquad Z_i = [\boldsymbol{z}_{i1}; \ldots; \boldsymbol{z}_{iK}]. \end{aligned} \tag{2}$$

Multiple query positions expose distributed cellular information to the language backbone, enabling it to extend the cell encoder in further extracting information about gene regulatory relationships and cellular heterogeneity.

Frozen language backbone. We adopt Llama (Grattafiori et al., 2024), a representative family of publicly available, open-weight language models, using Llama-3.1-8B-Instruct with approximately 8 billion parameters and embedding dimension $d_l = 4096$. Let $\boldsymbol{e}_{\mathrm{BOS}}$ be its beginning-of-sequence embedding, s_{cls} and s_{rep} two learned language-side embeddings, and $E_\phi(p_i)$ and $E_\phi(a_i)$ the embedded prompt and training answer. The input embedding sequence, representation readout, and numerical output are

$$\begin{aligned} u_i &= [\boldsymbol{e}_{\mathrm{BOS}}; s_{\mathrm{cls}}; \boldsymbol{z}_{i1}; \ldots; \boldsymbol{z}_{iK}; s_{\mathrm{rep}}; E_\phi(p_i); E_\phi(a_i)], \\ \boldsymbol{r}_i &= L_\phi(u_i)_{s_{\mathrm{rep}}}, \qquad \widehat{\boldsymbol{y}}_i = D_\psi(\boldsymbol{r}_i). \end{aligned} \tag{3}$$

Thus, Z_i is a continuous cellular prompt, not a sequence of gene names tokenized as text. The language-side s_{cls} is separate from the scGPT classification position replaced in Equation 1. Under causal attention, s_{rep} reads the preceding cellular prefix but cannot access the later question or answer. Its 4,096-dimensional output $\boldsymbol{r}_i$ is the cell representation. Text is generated through the backbone's native output head after the task prompt. The backbone, its text embeddings, and its native head remain frozen; gradients through the backbone train the cellular interface. We load the backbone with 4-bit NF4 quantization and bfloat16 computation.

Signal-recovery module. The signal-recovery module reconstructs the input cell's expression signal from the representation-token output $\boldsymbol{r}_i$. Specifically, D_ψ maps this representation to predicted expression over the target gene panel, providing quantitative supervision that encourages the cellular interface to retain expression information. Reconstruction targets are normalized and log-transformed as defined below. Decoder implementation details appear in Appendix B.1.

3.2 Model pretraining.

Omics signal self-prediction. The first principal task trains the cellular prompt and representation to retain quantitative expression information. For counts x_{ig} mapped to the target panel $\mathcal{G}$, we construct the target before truncating the encoder input:

$$y_{ig} = \log\left(1 + 10^4 \frac{x_{ig}}{\sum_{g' \in \mathcal{G}} x_{ig'}}\right). \tag{4}$$

Let o_{ig} indicate whether gene g is measured in the source feature panel. Reconstruction is supervised on observed entries, including measured zeros; genes absent from that panel are excluded. This self-prediction task uses the cell's own normalized, log-transformed expression as its reconstruction target.

Autoregressive text generation. The second principal task trains the cellular prompt to support answer and description generation. Questions cover cell identity, tissue of origin, and disease status; descriptive targets express cellular annotations in sentence form. The QA data were generated using ChatGPT and stored in a fixed target bank, containing 835 cell-type, 367 tissue, and 245 disease targets. Training samples cell-type QA, disease QA, tissue QA, and description tasks with a deterministic $7:1:1:2$ schedule, falling back to an available task when a target is missing. With teacher forcing, only answer tokens, including termination, contribute to the language loss. Prompt construction and illustrative examples are provided in Appendix B.2.

For batch $\mathcal{B}$ and T_i supervised answer tokens, the two principal losses are

$$\begin{aligned} \mathcal{L}_{\mathrm{rec}} &= \frac{\sum_{i \in \mathcal{B}, g} o_{ig} (\hat{y}_{ig} - y_{ig})^2}{\sum_{i \in \mathcal{B}, g} o_{ig}}, \\ \mathcal{L}_{\mathrm{lang}} &= -\frac{\sum_{i \in \mathcal{B}} \sum_{t=1}^{T_i} \log P_\phi(a_{it} \mid Z_i, p_i, a_{i,<t})}{\sum_{i \in \mathcal{B}} T_i}. \end{aligned} \tag{5}$$

The language conditioning notation omits the special embeddings in Equation 3; ϕ is fixed, while Z_i is trainable through the translator and bridge.

We additionally penalize redundant query states ($\mathcal{L}_{\text{div}}$), reconstruct expression with one query masked ($\mathcal{L}_{\text{mask}}$), and predict expression from metadata with all expression-derived positions masked ($\mathcal{L}_{\text{meta}}$):

$$\mathcal{L} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{lang}}\mathcal{L}_{\text{lang}} + \lambda_{\text{div}}\mathcal{L}_{\text{div}} + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}} + \lambda_{\text{meta}}\mathcal{L}_{\text{meta}}. \tag{6}$$

All weights are 1 except $\lambda_{\text{div}} = 0.1$. In the auxiliary passes, metadata precedes the representation token. The metadata-only readout therefore uses the supplied biological conditions without access to the target cell's expression. Auxiliary losses and scheduling details are given in Appendix B.2.

Training data and schedule. The training corpus is drawn from CELLxGENE Census release 2025-11-08 and comprises 444 primary-human datasets with 91,099,871 cells after exclusions. We first align the interface for one epoch with scGPT frozen, then unfreeze scGPT and jointly train the encoder, interface, and signal-recovery module until convergence, while keeping the language backbone frozen throughout. Training uses four H200 GPUs and AdamW (Loshchilov & Hutter, 2019); during joint training, the learning rates are 2×10^{-6} for scGPT and 10^{-5} for the other trainable components. By reusing pretrained cellular and language models, this design limits optimization to approximately 253 million parameters and avoids updating the full 8B language backbone. Optimization details are provided in Appendix B.3, and corpus exclusions and evaluation splits are described in Appendix C.

3.3 Downstream applications

All downstream applications use fixed model parameters. Clustering uses $\boldsymbol{r}_i$ extracted at s_{rep}, before any downstream text. QA and description generation append a task prompt to the cellular prefix. For metadata-conditioned expression prediction, metadata precedes s_{rep}, all expression-derived positions are masked, and the signal-recovery module decodes the resulting representation. Evaluation protocols appear in Appendix C.

4 Results

4.1 Experimental setup

We evaluate CellDialect under zero-shot transfer on 10 held-out datasets, with no dataset-specific fine-tuning or parameter updates during evaluation. The evaluated dataset IDs were excluded from CellDialect training. Representation quality is measured by normalized mutual information (NMI) and adjusted Rand index (ARI). Language evaluation covers question answering (QA) and descriptions; additional analyses examine metadata-conditioned expression estimation and query-token removal. We evaluate the model on 10 held-out datasets spanning 9 tissues: Brodmann area 46 (Brod46), cortex of kidney (CorKid), esophageal epithelium (EsoEpi), hippocampal formation (Hippo), kidney, labial gland (Labial), liver (Liver1 and Liver2), small intestine (SmaInt), and submandibular gland (SubMG). Evaluation details are organized in Appendix C. Representation evaluation uses all valid annotated cells. Across experiments, we assess how a compact cellular interface introduces omics information into a language-model framework while keeping the language backbone's pretrained weights and architecture intact.

4.2 Preserving cell structure under zero-shot transfer

The REP token precedes downstream text, and causal attention prevents it from accessing any subsequent metadata, question, or answer. Clustering therefore uses the expression-conditioned representation rather than later metadata. We construct cosine neighbor graphs from full-dimensional embeddings and apply Leiden clustering (Traag et al., 2019), selecting the resolution to match the number of annotated cell types. This uses the reference cluster count for clustering evaluation.

CellDialect achieves mean NMI of 0.699 and ARI of 0.530, exceeding the strongest evaluated baseline averages (Table 1; Appendix D). It improves NMI over scGPT-CP on nine of ten datasets and over scGPT-human on all ten. Uniform manifold approximation and projection(UMAP) (McInnes et al., 2018) on the kidney cortex dataset (39,176 cells; 13 annotated types) illustrates the resulting representation structure (Figure 2). These results show that the cellular interface supports transfer-

able representation structure within a frozen language-model framework, without requiring target-dataset fine-tuning.

Table 1: NMI across benchmark datasets.

Method	Brod46	CorKid	EsoEpi	Hippo	Kidney	Labial	Liver1	Liver2	SmaInt	SubMG	Avg
UCE	0.672	0.718	0.356	0.691	0.875	**0.618**	0.761	0.846	0.519	0.591	0.665
scGPT-CP	0.686	0.698	0.347	0.671	0.822	0.601	0.737	0.846	0.513	0.595	0.652
scGPT-human	0.688	0.667	0.350	0.669	0.818	0.516	0.691	0.837	0.456	0.511	0.620
C2S-Scale	0.658	0.683	0.300	0.733	0.744	0.591	0.737	0.753	0.386	0.576	0.616
TranscriptFormer	0.691	0.715	0.326	0.746	0.890	0.546	0.744	0.893	0.494	**0.670**	0.672
scfoundation	0.232	0.585	0.364	0.705	0.776	0.596	0.764	**0.917**	0.428	0.629	0.599
scmulan	0.550	0.494	0.318	0.792	0.542	0.582	0.705	0.727	0.220	0.375	0.531
CellDialect	**0.708**	**0.760**	**0.369**	**0.836**	**0.893**	0.595	**0.766**	0.876	**0.550**	0.639	**0.699**

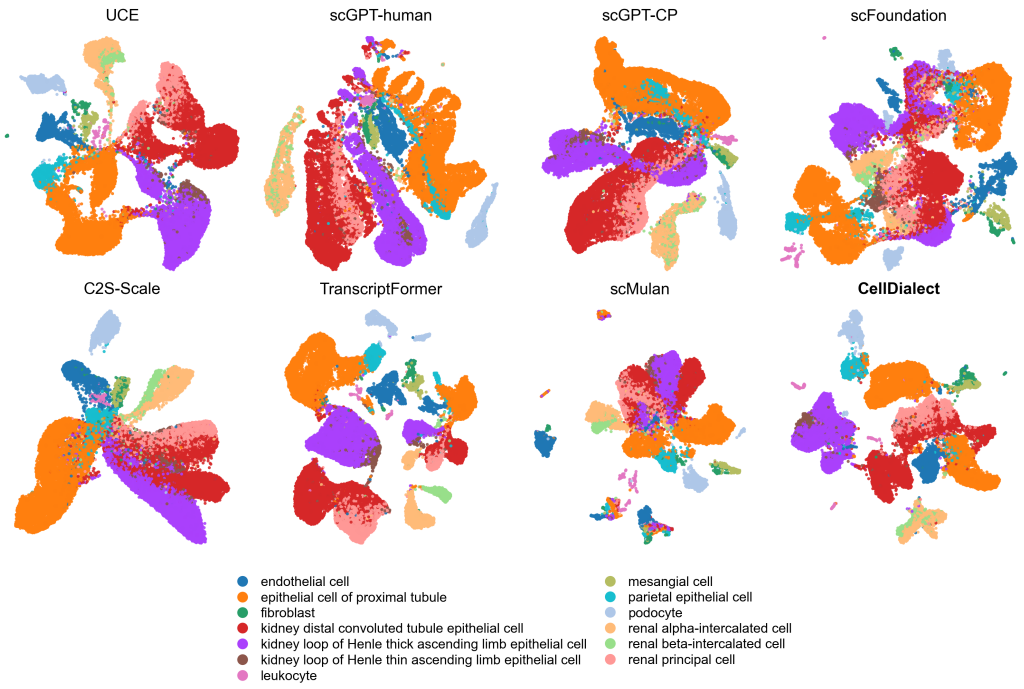

Figure 2: UMAP visualization of different methods.

4.3 Language Generation

CellDialect uses the same expression-derived cellular prefix to produce task-specific answers and integrated descriptions (Figure 3). The cellular prefix remains visible for these expression-to-text tasks. We evaluate cell type, tissue, and disease content by mapping predicted and reference concepts to the Cell Ontology (Diehl et al., 2016), UBERON (Mungall et al., 2012), and MONDO (Vasilevsky et al., 2026), respectively. Wu–Palmer similarity (Wu & Palmer, 1994) gives partial credit for hierarchically related concepts. Appendix C.3 defines the score. Figure 3 also summarizes task coverage under the evaluated checkpoints and prompting configurations. CellWhisperer and CellDialect cover all three question categories. Task-specific question-answering performance is examined in Section 4.4. The description-level results in Figure 3 support CellDialect's ability to communicate multiple aspects of cellular identity within a single description and identify variation in the reliability of individual attributes. Evaluations for age prediction are in Appendix E. Together, these observations show that the learned cellular interface makes omics information accessible to the frozen

language backbone for zero-shot text generation, extending its inputs without updating its pretrained weights or architecture.

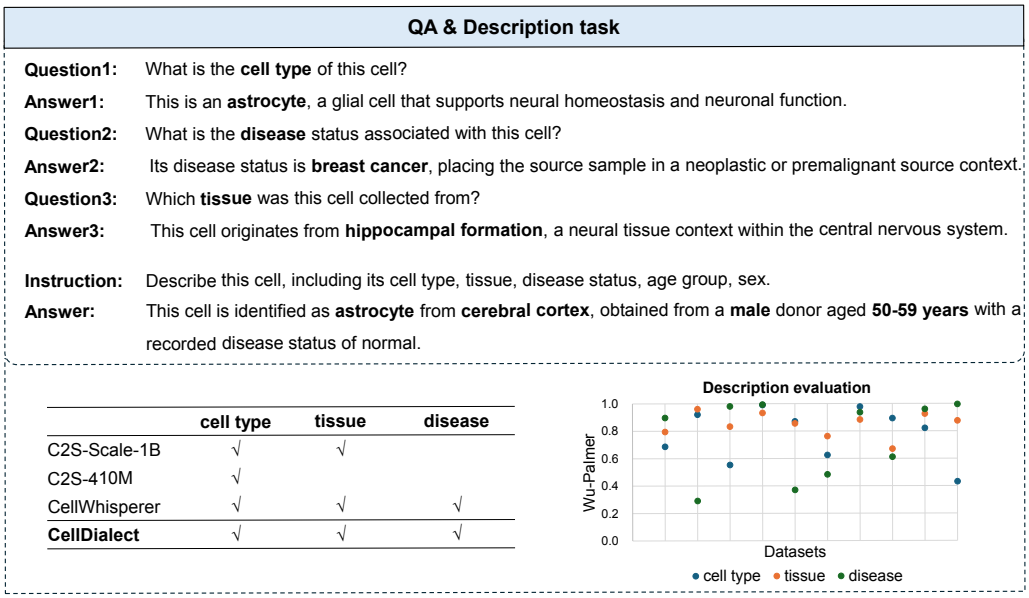

Figure 3: Language-related analysis.

4.4 CELL-TASK QUESTION ANSWERING

We evaluate cell type, tissue, and disease-status question answering across ten datasets under zero-shot transfer. CellDialect achieves the highest mean cell-type and tissue QA scores(Tables 2 and 3). These results support using the cellular interface to connect expression profiles to the frozen language model's semantic predictions across cellular identity and tissue context.

Table 2: Wu–Palmer scores for cell type QA across benchmark datasets.

Method	Brod46	CorKid	EsoEpi	Hippo	Kidney	Labial	Liver1	Liver2	SmaInt	SubMG	Avg
C2S-Scale-1B	0.606	0.857	0.739	0.997	0.852	0.655	0.509	0.836	0.569	0.362	0.698
C2S-410M	0.572	0.891	**0.896**	0.998	0.859	**0.714**	0.529	0.883	0.683	**0.561**	0.759
CellWhisperer	0.521	0.719	0.820	**0.999**	0.732	0.599	0.483	0.803	0.569	0.415	0.666
CellDialect	**0.707**	**0.922**	0.630	0.992	**0.886**	0.669	**0.976**	**0.884**	**0.720**	0.425	**0.781**

Table 3: Wu–Palmer scores for tissue QA across benchmark datasets.

Method	Brod46	CorKid	EsoEpi	Hippo	Kidney	Labial	Liver1	Liver2	SmaInt	SubMG	Avg
C2S-Scale-1B	**0.886**	**0.999**	**0.850**	0.897	**0.913**	0.596	0.698	0.575	0.931	0.626	0.797
CellWhisperer	0.800	0.869	0.799	0.773	0.762	0.510	0.554	**0.667**	**0.942**	0.511	0.719
CellDialect	0.828	0.951	0.665	**0.947**	0.857	**0.716**	**0.881**	0.664	0.931	**0.885**	**0.833**

Under the same zero-shot protocol, mean disease-status similarity is 0.598 on the five disease-associated datasets (Table 4) and 0.981 on the five healthy-only datasets (Table 12). These comparisons depend on the inference protocol: CellWhisperer's published disease benchmark uses

candidate-label matching, whereas here its released chat model generates free-form answers without candidate labels. Its low healthy-only scores therefore characterize this chat configuration, rather than its published candidate-label benchmark. Together, the QA results demonstrate that the cellular interface connects expression information to the semantic capabilities of a large language model while preserving the language backbone's pretrained weights and architecture.

Table 4: Wu–Palmer scores for disease QA across disease benchmark datasets.

Method	CorKid	Kidney	Labial	Liver1	Liver2	Avg
CellWhisperer	0.024	0.220	0.154	0.721	0.323	0.288
CellDialect	**0.558**	**0.398**	**0.519**	**0.929**	**0.585**	**0.598**

4.5 METADATA-CONDITIONED EXPRESSION GENERATION

We evaluate metadata-conditioned expression generation in a zero-shot setting on held-out data excluded from CellDialect training, with no dataset-specific fine-tuning. All expression-derived cell token embedding positions are masked in language-model attention. The REP token can therefore use the supplied metadata but cannot access the target cell's expression through the cellular prefix. The shared decoder maps this metadata-conditioned representation to a normalized expression vector. Figure 4 provides a qualitative comparison with real cells through UMAP placement and marker-expression patterns in a hippocampal formation dataset (Hu et al., 2023). This qualitative analysis illustrates how the shared cellular interface extends the frozen language-model framework to metadata-conditioned numerical outputs while preserving the language backbone's pretrained weights and architecture.

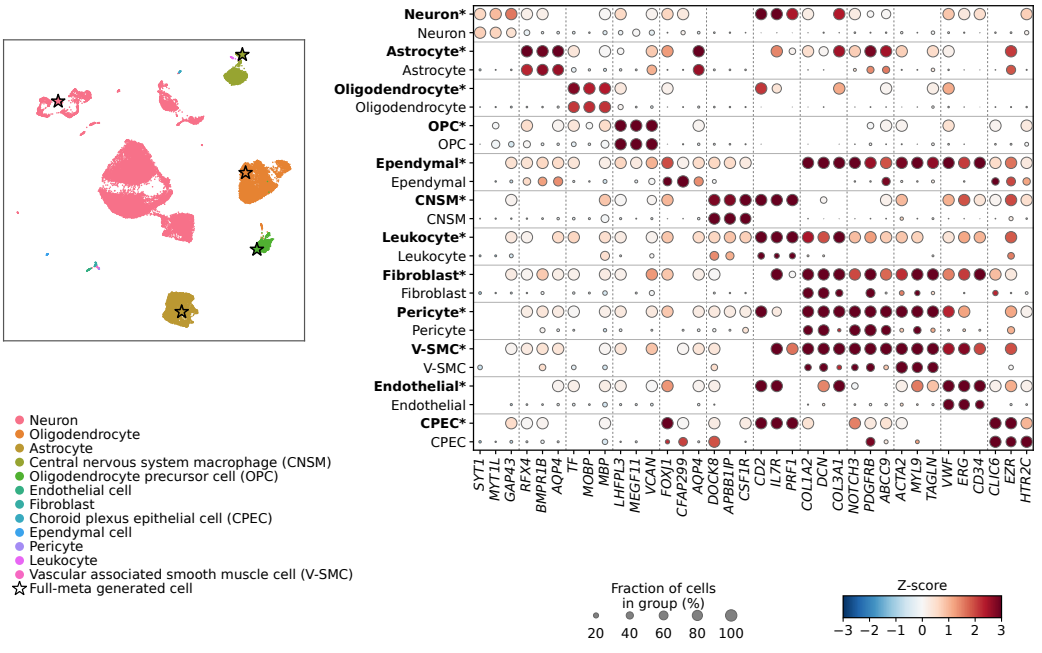

Figure 4: Generation performance comparison.

4.6 Ablation studies

Number of query tokens. We remove query tokens at inference, retaining 12, 8, 4, or 2 tokens with ten subset trials per reduced setting (Figure 5). Mean NMI is 0.699, 0.694, and 0.696 with 16, 12, and 8 tokens, respectively, decreasing to 0.680 with 2 tokens. Cell-type QA is more sensitive to removal, suggesting that additional tokens support semantic prediction beyond the information sufficient for clustering. Per-dataset results and trial variability appear in Appendix F.

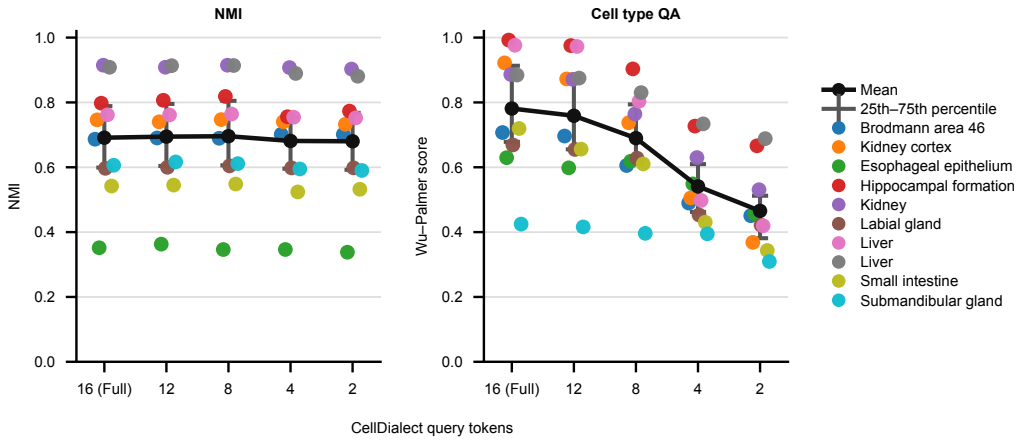

Figure 5: Ablation studies on number of query tokens.

Language-backbone comparison. We compare CellDialect with two frozen language backbones: Sheared-LLaMA-1.3B (Xia et al., 2023) and Llama-3.1-8B-Instruct (Grattafiori et al., 2024). As shown in Table 5, representation quality is similar across the two configurations. The 8B configuration performs better on cell-type, tissue, and disease QA. These results suggest that the choice of language backbone has a greater effect on language-based predictions than on clustering quality in this comparison. This comparison reflects the combined influence of backbone size, pretraining, and instruction tuning.

Table 5: Language-backbone comparison.

Language backbone	NMI	ARI	Cell type QA	Tissue QA	Disease QA
Sheared-LLaMA-1.3B	0.696	**0.536**	0.750	0.710	0.546
Llama-3.1-8B-Instruct	**0.699**	0.530	**0.781**	**0.833**	**0.598**

5 Conclusion

We present CellDialect, a framework that rethinks single-cell foundation models as modality translators connecting gene programs to biological language. CellDialect transforms gene expression profiles into cell-conditioned continuous tokens—a "cell dialect" that enables a frozen large language model to access cellular information. Joint numerical and language supervision supports representation learning, question answering, and descriptive text generation. Evaluations demonstrate improved average representation quality and strong zero-shot question-answering performance. These findings extend the role of single-cell foundation models from learning cellular representations to translating cellular information into a form accessible to language models, supporting diverse tasks with modest additional training.

AI USE STATEMENT

We used generative AI tools to polish the manuscript, identify relevant literature, assist with code debugging for engineering and repetitive implementation tasks, and generate synthetic question–answer pairs for constructing the QA dataset. The authors take responsibility for the final manuscript, references, code, and datasets, including all AI-assisted content and the scientific claims based on it.

REFERENCES

Ricard Argelaguet, Anna S. E. Cuomo, Oliver Stegle, and John C. Marioni. Computational principles and challenges in single-cell data integration. *Nature Biotechnology*, 39:1202–1215, 2021. doi: 10.1038/s41587-021-00895-7. URL https://doi.org/10.1038/s41587-021-00895-7.

Islambek Ashyrmamatov, Su Ji Gwak, Su-Young Jin, Ikhyeong Jun, Umit V. Ucak, Jay-Yoon Lee, and Juyong Lee. A survey on large language models in biology and chemistry. *Experimental & Molecular Medicine*, 58:970–980, 2026. doi: 10.1038/s12276-025-01583-1. URL https://doi.org/10.1038/s12276-025-01583-1.

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

Haiyang Bian, Yixin Chen, Xiaomin Dong, Chen Li, Minsheng Hao, Sijie Chen, Jinyi Hu, Maosong Sun, Lei Wei, and Xuegong Zhang. scMulan: A Multitask Generative Pre-Trained Language Model for Single-Cell Analysis. In Jian Ma (ed.), *Research in Computational Molecular Biology (RECOMB)*, volume 14758 of *Lecture Notes in Computer Science*, pp. 479–482. Springer, 2024. doi: 10.1007/978-1-0716-3989-4_57. URL https://doi.org/10.1007/978-1-0716-3989-4_57.

Xiaoyang Chen, Keyi Li, Xuejian Cui, Zian Wang, Qun Jiang, Jiacheng Lin, Zhen Li, Zijing Gao, Hairong Lv, and Rui Jiang. EpiAgent: foundation model for single-cell epigenomics. *Nature Methods*, 22:2316–2327, 2025. doi: 10.1038/s41592-025-02822-z. URL https://doi.org/10.1038/s41592-025-02822-z.

Jo-Fan Chien, Hanqing Liu, Bang-An Wang, Chongyuan Luo, et al. Cell-type-specific effects of age and sex on human cortical neurons. *Neuron*, 112(15):2524–2539.e5, 2024. doi: 10.1016/j.neuron.2024.05.013.

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

Alexander D. Diehl, Terrence F. Meehan, Yvonne M. Bradford, Matthew H. Brush, Wasila M. Dahdul, David S. Dougall, Yongqun He, David Osumi-Sutherland, Alan Ruttenberg, Sirarat Sarntivijai, et al. The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. *Journal of Biomedical Semantics*, 7(1):44, 2016. doi: 10.1186/s13326-016-0088-7. URL https://doi.org/10.1186/s13326-016-0088-7.

Daniel Dimitrov, Stefan Schrod, Martin Rohbeck, and Oliver Stegle. Interpretation, extrapolation and perturbation of single cells. *Nature Reviews Genetics*, 27:349–370, 2026. doi: 10.1038/s41576-025-00920-4. URL https://doi.org/10.1038/s41576-025-00920-4.

Rasa Elmentaite, Cecilia Domínguez Conde, Lu Yang, and Sarah A. Teichmann. Single-cell atlases: shared and tissue-specific cell types across human organs. *Nature Reviews Genetics*, 23:395–410, 2022. doi: 10.1038/s41576-022-00449-w. URL https://doi.org/10.1038/s41576-022-00449-w.

Gökcen Eraslan, Žiga Avsec, Julien Gagneur, and Fabian J. Theis. Deep learning: new computational modelling techniques for genomics. *Nature Reviews Genetics*, 20:389–403, 2019. doi: 10.1038/s41576-019-0122-6. URL https://doi.org/10.1038/s41576-019-0122-6.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 Herd of Models. arXiv preprint arXiv:2407.21783, 2024. URL https://arxiv.org/abs/2407.21783.

Gunsagar S. Gulati, Jeremy Philip D'Silva, Yunhe Liu, Linghua Wang, and Aaron M. Newman. Profiling cell identity and tissue architecture with single-cell and spatial transcriptomics. *Nature Reviews Molecular Cell Biology*, 26:11–31, 2025. doi: 10.1038/s41580-024-00768-2. URL https://doi.org/10.1038/s41580-024-00768-2.

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

E. Horeth, J. Bard, M. Che, T. Wrynn, E. A. C. Song, et al. High-resolution transcriptomic landscape of the human submandibular gland. *Journal of Dental Research*, 102(5):525–535, 2023. doi: 10.1177/00220345221147908.

Congxue Hu, Tengyue Li, Yingqi Xu, Xinxin Zhang, Feng Li, Jing Bai, Jing Chen, Wenqi Jiang, Kaiyue Yang, Qi Ou, et al. CellMarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data. *Nucleic Acids Research*, 51(D1): D870–D876, 2023. doi: 10.1093/nar/gkac947. URL https://doi.org/10.1093/nar/gkac947.

Oussama Kharouiche, Aris Markogiannakis, Xiao Fei, Michail Chatzianastasis, and Michalis Vazirgiannis. Cell2Text: Multimodal LLM for Generating Single-Cell Descriptions from RNA-Seq Data. arXiv preprint arXiv:2509.24840, 2025. URL https://arxiv.org/abs/2509.24840.

Johanna Klughammer, Daniel L. Abravanel, Åsa Segerstolpe, Timothy R. Blosser, et al. A multi-modal single-cell and spatial expression map of metastatic breast cancer biopsies across clinicopathological features. *Nature Medicine*, 30(11):3236–3249, 2024. doi: 10.1038/s41591-024-03215-z.

Daniel Levine, Syed A 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. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 27299–27325. PMLR, 2024. URL https://proceedings.mlr.press/v235/levine24a.html.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Malte D. Luecken and Fabian J. Theis. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, 15(6):e8746, 2019. doi: 10.15252/msb.20188746. URL https://doi.org/10.15252/msb.20188746.

E. Madissoon, A. Wilbrey-Clark, R. J. Miragaia, et al. scrna-seq assessment of the human lung, spleen, and esophagus tissue stability after cold preservation. *Genome Biology*, 21:1, 2020. doi: 10.1186/s13059-019-1906-x.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv preprint arXiv:1802.03426, 2018. URL https://arxiv.org/abs/1802.03426.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):30:1–30:40, 2023. doi: 10.1145/3605943. URL https://doi.org/10.1145/3605943.

Christopher J Mungall, Carlo Torniai, Georgios V Gkoutos, Suzanna E Lewis, and Melissa A Haendel. Uberon, an integrative multi-species anatomy ontology. *Genome Biology*, 13 (1):R5, 2012. doi: 10.1186/gb-2012-13-1-r5. URL https://doi.org/10.1186/gb-2012-13-1-r5.

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

Thomas J. F. Pranzatelli, Paola Perez, Anson Ku, Bruno Matuck, et al. GZMK+CD8+ t cells target a specific acinar cell type in sjögren's disease. *Annals of the Rheumatic Diseases*, 85(5):880–898, 2026. doi: 10.1016/j.ard.2025.08.014.

Maximilian Reck, David P. Baird, Stefan Veizades, et al. Multiomic analysis of human kidney disease identifies a tractable inflammatory and pro-fibrotic tubular cell phenotype. *Nature Communications*, 16:4745, 2025. doi: 10.1038/s41467-025-59997-4.

Syed Asad Rizvi, Daniel Levine, Aakash Patel, Shiyang Zhang, Eric Wang, Curtis Jamison Perry, Ivan Vrkic, Nicole Mayerli Constante, Zirui Fu, Sizhuang He, et al. Scaling Large Language Models for Next-Generation Single-Cell Analysis. bioRxiv preprint, 2025. URL https://doi.org/10.1101/2025.04.14.648850.

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

Moritz Schaefer, Peter Peneder, Daniel Malzl, Salvo Danilo Lombardo, Mihaela Peycheva, Jake Burton, Anna Hakobyan, Varun Sharma, Thomas Krausgruber, Celine Sin, et al. Multimodal learning enables chat-based exploration of single-cell data. *Nature Biotechnology*, 2025. doi: 10.1038/s41587-025-02857-9. URL https://doi.org/10.1038/s41587-025-02857-9.

Kimberly Siletti, Rebecca Hodge, Alejandro Mossi Albiach, Ka Wai Lee, Song-Lin Ding, et al. Transcriptomic diversity of cell types across the adult human brain. *Science*, 382(6667):eadd7046, 2023. doi: 10.1126/science.add7046.

Christina V. Theodoris, Ling Xiao, Anant Chopra, Mark D. Chaffin, Zeina R. Al Sayed, Matthew C. Hill, Helene Mantineo, Elizabeth M. Brydon, Zexian Zeng, X. Shirley Liu, and Patrick T. Ellinor. Transfer learning enables predictions in network biology. *Nature*, 618: 616–624, 2023. doi: 10.1038/s41586-023-06139-9. URL https://doi.org/10.1038/s41586-023-06139-9.

V. A. Traag, L. Waltman, and N. J. van Eck. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1):5233, 2019. doi: 10.1038/s41598-019-41695-z. URL https://doi.org/10.1038/s41598-019-41695-z.

Manuel Trebo, Thomas Maurer, Felix J. Krendl, Stefan Salcher, et al. Ex vivo modelling of human colorectal cancer liver metastasis by normothermic machine perfusion. *Molecular Cancer*, 24: 264, 2025. doi: 10.1186/s12943-025-02430-7.

Nicole A Vasilevsky, Sabrina Toro, Nicolas Matentzoglu, Joseph E Flack, Kathleen R Mullen, Harshad Hegde, Sarah Gehrke, Patricia L Whetzel, Yousif Shwetar, Nomi L Harris, et al. Mondo: integrating disease terminology across communities. *GENETICS*, 232(4):iyaf215, 2026. doi: 10.1093/genetics/iyaf215. URL https://doi.org/10.1093/genetics/iyaf215.

Parker C. Wilson, Yoshiharu Muto, Haojia Wu, Anil Karihaloo, Sushrut S. Waikar, and Benjamin D. Humphreys. Multimodal single cell sequencing implicates chromatin accessibility and genetic background in diabetic kidney disease progression. *Nature Communications*, 13:5253, 2022. doi: 10.1038/s41467-022-32972-z.

Zhibiao Wu and Martha Palmer. Verb Semantics and Lexical Selection. In *32nd Annual Meeting of the Association for Computational Linguistics*, pp. 133–138. Association for Computational Linguistics, 1994. doi: 10.3115/981732.981751. URL https://aclanthology.org/P94-1019/.

Mengzhou Xia, Tianyu Gao, Zhiyuan Zeng, and Danqi Chen. Sheared LLaMA: Accelerating language model pre-training via structured pruning. *arXiv preprint arXiv:2310.06694*, 2023. URL https://arxiv.org/abs/2310.06694.

Fan Yang, Wenchuan Wang, Fang Wang, Yuan Fang, Duyu Tang, Junzhou Huang, Hui Lu, and Jianhua Yao. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nature Machine Intelligence*, 4:852–866, 2022. doi: 10.1038/s42256-022-00534-z. URL https://doi.org/10.1038/s42256-022-00534-z.

Xiaodong Yang, Guole Liu, Guihai Feng, Dechao Bu, Pengfei Wang, Jie Jiang, Shubai Chen, Qinmeng Yang, Hefan Miao, Yiyang Zhang, Zhenpeng Man, Zhongming Liang, Zichen Wang, Yaning Li, Zheng Li, Yana Liu, Yao Tian, Wenhao Liu, Cong Li, Ao Li, Jingxi Dong, Zhilong Hu, Chen Fang, Lina Cui, Zixu Deng, Haiping Jiang, Wentao Cui, Jiahao Zhang, Zhaohui Yang, Handong Li, Xingjian He, Liqun Zhong, Jiaheng Zhou, Zijian Wang, Qingqing Long, Ping Xu, The X-Compass Consortium, Hongmei Wang, Zhen Meng, Xuezhi Wang, Yangang Wang, Yong Wang, Shihua Zhang, Jingtao Guo, Yi Zhao, Yuanchun Zhou, Fei Li, Jing Liu, Yiqiang Chen, Ge Yang, and Xin Li. GeneCompass: deciphering universal gene regulatory mechanisms with a knowledge-informed cross-species foundation model. *Cell Research*, 34:830–845, 2024. doi: 10.1038/s41422-024-01034-y. URL https://doi.org/10.1038/s41422-024-01034-y.

Suyuan Zhao, Jiahuan Zhang, Yushuai Wu, Yizhen Luo, and Zaiqing Nie. LangCell: Language-Cell Pre-training for Cell Identity Understanding. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 61159–61185. PMLR, 2024. URL https://proceedings.mlr.press/v235/zhao24u.html.

Rachel K. Zwick, Petr Kasparek, Brisa Palikuqi, Sara Viragova, et al. Epithelial zonation along the mouse and human small intestine defines five discrete metabolic domains. *Nature Cell Biology*, 26(2):250–262, 2024. doi: 10.1038/s41556-023-01337-z.

A CODE AVAILABILITY

All data processing and training code, as well as the checkpoints used for final evaluation, will be made available through the anonymous project website: `https://osf.io/ph92x/overview?view_only=980252e087a045f8bdc47a402427dab9`.

B IMPLEMENTATION, HYPERPARAMETERS, AND TRAINING

B.1 ARCHITECTURE AND SEQUENCE CONSTRUCTION

We use scGPT-CP, the continually pretrained variant of scGPT. The scGPT input at position j is $e_{ij} = E_g(g_{ij}) + E_v(v_{ij})$. Two trainable arrays $Q, I \in \mathbb{R}^{16\times 512}$ are added before encoding; both are shared across cells. Queries and valid genes participate in the same self-attention computation, with padding masked. The first 16 output states are normalized using the implementation's numerical safeguard and passed through

$$z_{ik} = W_2\,\mathrm{LN}(\mathrm{GELU}(W_1 h_{ik} + b_1)) + b_2. \tag{7}$$

The bridge has widths $512 \to 2048 \to 4096$. The signal-recovery module reconstructs normalized, log-transformed expression from the representation-token output $\boldsymbol{r}_i$. The numerical decoder has widths $4096 \to 8192 \to 19520$, with GELU and layer normalization between its linear layers.

In Equation 3, BOS is the beginning-of-sequence token, p_i includes the instruction and chat control tokens, and a_i includes the answer and termination token. Representation extraction uses only the prefix through s_{rep}, yielding a 4,096-dimensional embedding. Teacher forcing predicts each answer token from the preceding state; prompt and padding positions do not contribute to the language loss.

Table 6: Architecture and numerical-target configuration.

Component	Configuration
Cell encoder	scGPT-CP; 12 layers, 8 heads, width 512
Language backbone	Llama-3.1-8B-Instruct
Query count / encoder width	16 / 512
Query parameters	Two arrays $Q, I \in \mathbb{R}^{16\times 512}$
Projection widths	$512 \to 2048 \to 4096$
Representation dimension	4,096
Expression decoder widths	$4096 \to 8192 \to 19520$
Target gene panel size	19,520
Input representation	Gene identities and binned expression values
Expression binning / input length	51 bins / length 1,200 (including 16 queries)
Target normalization / observation mask construction	$\log(1+10^4 x/\sum_g x_g)$ over the target panel; source-feature availability mask
Backbone precision / quantization	4-bit NF4, double quantization, bfloat16 compute

B.2 AUXILIARY OBJECTIVES

For a set of cells $\mathcal{U}$, define the observed-entry reconstruction error

$$\mathcal{E}(\hat{Y}, Y, O; \mathcal{U}) = \frac{\sum_{i\in\mathcal{U},g} o_{ig}(\hat{y}_{ig} - y_{ig})^2}{\sum_{i\in\mathcal{U},g} o_{ig}}. \tag{8}$$

This pools observed cell–gene entries.

Query diversity. A margin penalty discourages strong correlation or anticorrelation:

$$\mathcal{L}_{\mathrm{div}} = \frac{1}{|\mathcal{B}|K(K-1)} \sum_{i\in\mathcal{B}} \sum_{k\neq l} \left[\max(0, |h_{ik}^\top h_{il}| - \mu)\right]^2. \tag{9}$$

Masked-query reconstruction. A rotating schedule selects a subset $\mathcal{S}$ of batch cells and one query position k_i per selected cell. The auxiliary sequence is $[s_{\text{cls}}; Z_i; m_i; s_{\text{rep}}]$, where m_i is metadata after dropout. The selected query position is masked in language attention. The resulting prediction is supervised by

$$\mathcal{L}_{\text{mask}} = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \frac{\sum_g o_{ig} (\hat{y}_{ig}^{(-k_i)} - y_{ig})^2}{\sum_g o_{ig}}. \tag{10}$$

Unlike $\mathcal{E}$, this first averages within a cell and then equally across selected cells. The rotating schedule distributes removed positions across the queries. Visible cell tokens and metadata jointly support this training objective; it is distinct from the evaluation-time removal of multiple query tokens in Appendix F.

Metadata-conditioned reconstruction. A second pass uses the same selected cells and metadata but masks all 16 expression-derived positions. The representation readout accesses only the classification embedding and visible metadata:

$$\mathcal{L}_{\text{meta}} = \mathcal{E}(\hat{Y}^{\text{meta}}, Y, O; \mathcal{S}). \tag{11}$$

Complete metadata occupies two of four scheduling slots; each predefined partial subset occupies one. One partial subset contains cell type, tissue, and disease status. Whole-prompt metadata dropout has probability 0.15. The representation token is last, and position indices cumulatively count visible positions. The other partial subset contains tissue, donor age group, and profiling assay.

Autoregressive text-generation targets. Cell-type, disease, and tissue answers come from a fixed prebuilt target bank; descriptions express annotations through sentence variants. The deterministic task schedule has weights $7 : 1 : 1 : 2$ for these three QA tasks and description, respectively, with fallback when a requested target is missing. The static target bank contains 835 cell-type, 245 disease, and 367 tissue targets across the three QA tasks. The QA training data were generated using ChatGPT to provide a consistent set of question–answer pairs across tasks.

Prompt and answer format. The following examples specify the question format documented in Appendix C; bracketed answer fields are placeholders, not extracted training records. A training instance pairs the cellular prefix with one question and its corresponding target-bank answer:

> **Cell-type QA.** What is the cell type of this cell?
> **Answer:** [This is an astrocyte, a glial cell that supports neural homeostasis and neuronal function.]
>
> **Tissue QA.** Which tissue was this cell collected from?
> **Answer:** [This cell originates from hippocampal formation, a neural tissue context within the central nervous system.]
>
> **Disease QA.** What is the disease status associated with this cell?
> **Answer:** [Its disease status is breast cancer, placing the source sample in a neoplastic or premalignant source context.]

The bank contains category-level targets that can be reused across cells; its category counts are not counts of unique cell–question pairs. Descriptions express available cell annotations through sentence variants.

B.3 Training stages and optimizer

Training begins with one epoch of interface alignment, during which scGPT is frozen and the interface and signal-recovery module are optimized. We then unfreeze scGPT and jointly train these components until convergence. The language backbone remains frozen throughout, and training uses four H200 GPUs.

The language backbone and native head stay frozen. Numerical reconstruction and language supervision update the expression translator through the frozen backbone; the diversity term acts on query states. Metadata-only supervision updates the numerical decoder and visible language-side special embeddings without access to expression-derived positions.

Table 7: Loss weights and auxiliary-task settings.

Setting	Value
Omics self-prediction weight λ_{rec}	1.0
Autoregressive text-generation weight λ_{lang}	1
Query-diversity weight λ_{div}	0.1
Masked-query weight λ_{mask}	1.0
Metadata weight λ_{meta}	1
Diversity margin μ	0.2
Masked queries per selected training cell	1, rotating across query positions
Metadata whole-prompt dropout	0.15
Cell-type : disease : tissue : description schedule	$7:1:1:2$
Complete : partial-1 : partial-2 metadata schedule	$2:1:1$

Table 8: Training schedule and compute configuration. Missing duration and stopping-criterion details are marked for completion.

Stage	Cell encoder	Training schedule	Hardware / duration
Interface alignment	Frozen	1 epoch	4 H200 GPUs / about 110h
Joint training	Trainable	Until convergence	4 H200 GPUs / about 230h

Table 9: Optimization settings for joint training.

Setting	Value
Optimizer	AdamW (Loshchilov & Hutter, 2019)
Local batch size	16
scGPT learning rate	2×10^{-6}
Other trainable-component learning rate	10^{-5}
Weight decay / Adam betas / epsilon	0.01 / $(0.9, 0.999)$ / 10^{-8}
LR schedule / warmup	Fixed learning rates; no scheduler or LR warmup in the joint training loop
Gradient accumulation / clipping	One optimizer step per batch / norm 1.0
Training seeds / independent runs	Training seed 42
Hardware / GPU-hours / precision	4 H200 GPUs; about 1360; bfloat16 autocast

Trainable parameters and data exclusions. The model has 253,256,768 trainable parameters. Held-out dataset IDs are excluded from CellDialect training; the exclusion scope and corpus construction are described in Appendix C.

C Data and evaluation protocols

C.1 Training corpus and held-out datasets

The main tables report ten datasets from the selected-10 inventory. The training corpus uses CELLxGENE Census release 2025-11-08, 444 post-exclusion primary-human datasets, and 91,099,871 cells. The exclusion list covers 131 selected datasets and the named held-out collections; these exclusions prevent overlap with the held-out dataset IDs during CellDialect training. Exclusion concerns additional CellDialect training, not inherited pretraining. Table 10 gives source studies and dataset information.

Table 10: Held-out datasets.

Dataset	Tissue	Source study	Number of cells	Number of cell types
Brod46	Brodmann area 46	Chien et al. (2024)	37,121	15
CorKid	Cortex of Kidney	Wilson et al. (2022)	39,176	13
EsoEpi	Epithelium of Esophagus	Madissoon et al. (2020)	87,947	15
Hippo	Hippocampal Formation	Siletti et al. (2023)	40,191	12
Kidney	Kidney	Reck et al. (2025)	46,957	15
Labial	Labial Gland	Pranzatelli et al. (2026)	94,227	19
Liver1	Liver	Klughammer et al. (2024)	12,494	10
Liver2	Liver	Trebo et al. (2025)	75,104	14
SmaInt	Small Intestine	Zwick et al. (2024)	19,200	10
SubMG	Submandibular Gland	Horeth et al. (2023)	15,684	16

C.2 Representation evaluation

All annotated cells are embedded with fixed model parameters. Cosine nearest-neighbor graphs are constructed from the full-dimensional embeddings. Leiden resolution is selected by bisection to match the reference number of cell types. Thus, the reference cluster count is used in evaluation; the protocol is not fully label-free clustering. NMI and ARI compare clusters with reference annotations and are averaged equally across datasets. UMAP is computed from the embedding-based neighbor graph. The evaluator use 8 bisection steps and 21 local-grid points, with resolution bounds 0.01–64. If the search skips the target count, graph-adjacent communities are merged deterministically. NMI uses scikit-learn's arithmetic normalization; PCA is disabled for this representation benchmark. UMAP uses random initialization. Neighbor count and random seed are 15 and 42 in the earlier joint configuration.

C.3 Ontology mapping and semantic similarity

Cell type, tissue, and disease concepts are mapped to the Cell Ontology (Diehl et al., 2016), UBERON (Mungall et al., 2012), and MONDO (Vasilevsky et al., 2026). Wu–Palmer similarity (Wu & Palmer, 1994) is

$$\mathrm{WP}(\hat{c}, c) = \frac{2d(a^*)}{d(\hat{c}) + d(c)}, \qquad a^* = \arg \max_{a \in A(\hat{c}) \cap A(c)} d(a). \tag{12}$$

$A(c)$ includes a concept and its ancestors. Root depth is 1; for multiple-parent hierarchies, depth follows the longest path from a root. The selected common ancestor has greatest depth. Identical concepts receive 1, and hierarchically related concepts receive partial credit.

C.4 Language and expression-generation evaluation

QA uses task-specific prompts with fixed model parameters; descriptions are evaluated separately for cell type, tissue, and disease content. Dataset scores receive equal weight. CellWhisperer's released chat model is evaluated with free-form generation and no candidate label list, unlike its published candidate-label disease benchmark (Schaefer et al., 2025). The archived evaluator defines

the prompts "What is the cell type of this cell?", "Which tissue was this cell collected from?" and "What is the disease status associated with this cell?".

Metadata-conditioned expression estimation masks every expression-derived position and uses the shared decoder. The current evidence is qualitative UMAP placement and marker-expression comparison. Marker selection references CellMarker 2.0 (Hu et al., 2023).

C.5 Mapping descriptive text to Cell Ontology

We map generated descriptions to Cell Ontology nodes using a rule-based procedure. Matching prioritizes the cell-identity clause preceding the first punctuation mark. Text is lowercased, with hyphens, underscores, and punctuation normalized, before matching official ontology names and synonyms as contiguous phrases. Multiple matches are resolved by prioritizing longer phrases, official names over synonyms, and deeper ontology nodes.

When no exact phrase match is found, fuzzy matching requires a similarity score of at least 92 and a margin of at least 3 points over the runner-up. Long descriptions without an exact match are not forcibly mapped through fuzzy matching.

The mapping success rate is 99.43%, indicating high coverage. Successful mapping denotes assignment to an ontology node, rather than a correct cell-type prediction. Final predictions are evaluated using Wu–Palmer similarity between the mapped and reference nodes; unmapped outputs receive a score of zero and remain included in the reported averages.

D Supplementary benchmark results and visualizations

D.1 Adjusted Rand index

Table 11 reproduces the original ARI results without changing entries, formatting, or emphasis. Per-dataset NMI and QA results remain in the main text.

Table 11: ARI across benchmark datasets.

Method	Brod46	CorKid	EsoEpi	Hippo	Kidney	Labial	Liver1	Liver2	SmaInt	SubMG	Avg
UCE	0.591	0.547	0.118	0.339	0.802	0.325	0.528	0.788	0.356	0.351	0.475
scGPT-CP	0.580	0.500	0.119	0.286	0.719	0.262	0.442	0.766	0.301	0.317	0.429
scGPT-human	**0.592**	0.540	0.128	0.280	0.694	0.242	**0.617**	0.738	0.302	0.271	0.440
C2S-Scale	0.546	0.494	0.103	0.471	0.598	0.306	0.539	0.575	0.206	0.346	0.418
TranscriptFormer	0.574	0.544	0.126	0.532	**0.867**	0.312	0.453	0.918	0.309	**0.484**	0.512
scfoundation	0.103	0.365	0.150	0.348	0.681	0.266	0.568	**0.940**	0.277	0.441	0.414
scmulan	0.529	0.360	**0.234**	0.685	0.326	**0.416**	0.562	0.638	0.112	0.183	0.405
CellDialect	0.520	**0.604**	0.149	**0.722**	0.850	0.258	0.556	0.822	0.348	0.474	**0.530**

D.2 UMAPs for the remaining benchmarks

The main text shows kidney cortex. Figure 6-Figure 14 are the remaining benchmark visualizations, including brodmann area 46, epithelium of esophagus, hippocampal formation, Kidney, labial gland, Liver1, Liver2, small intestine, and submandibular gland, with method and cell-type labels and a shared legend within each dataset. Seeds and UMAP settings belong in Appendix C.

D.3 Wu–Palmer scores for disease QA across normal benchmark datasets.

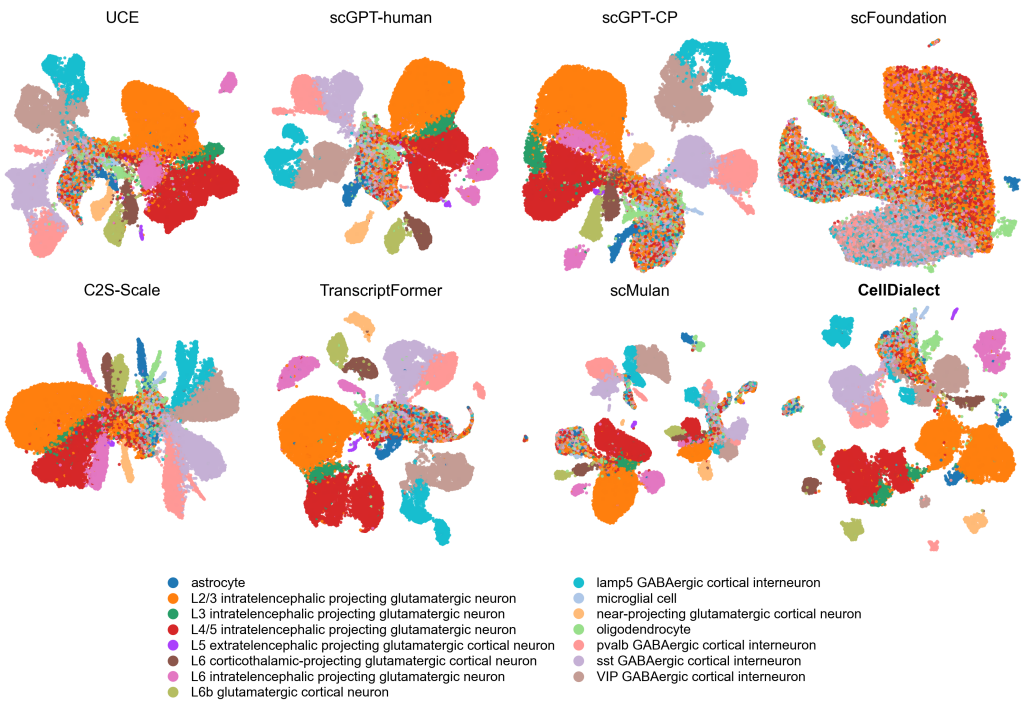

Figure 6: UMAP visualization of brodmann area 46.

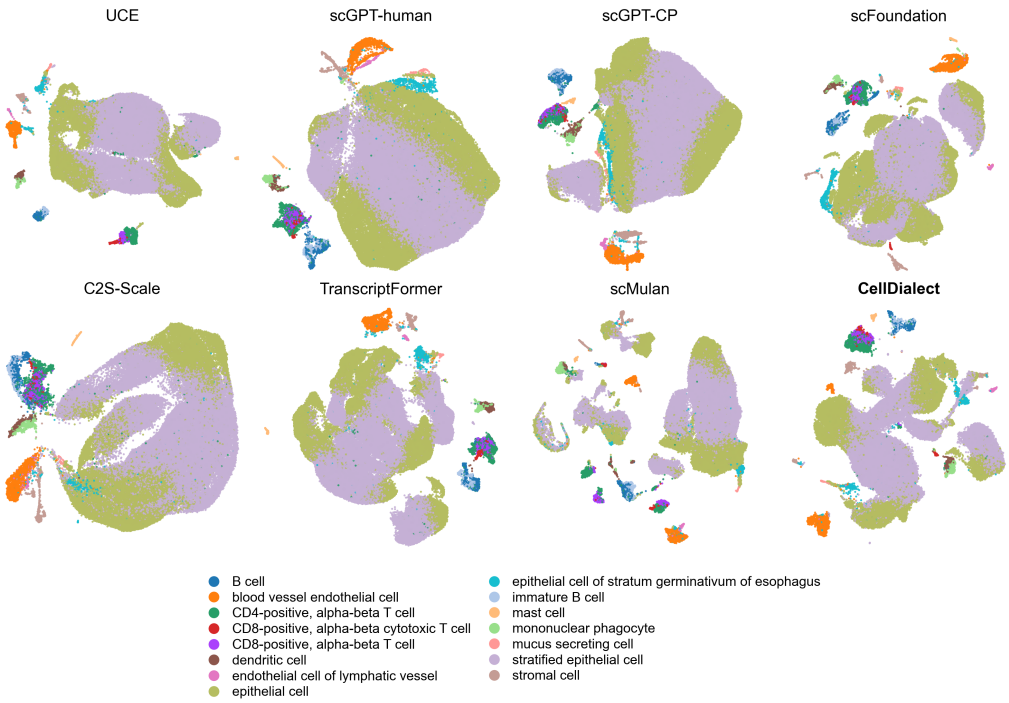

Figure 7: UMAP visualization of epithelium of esophagus.

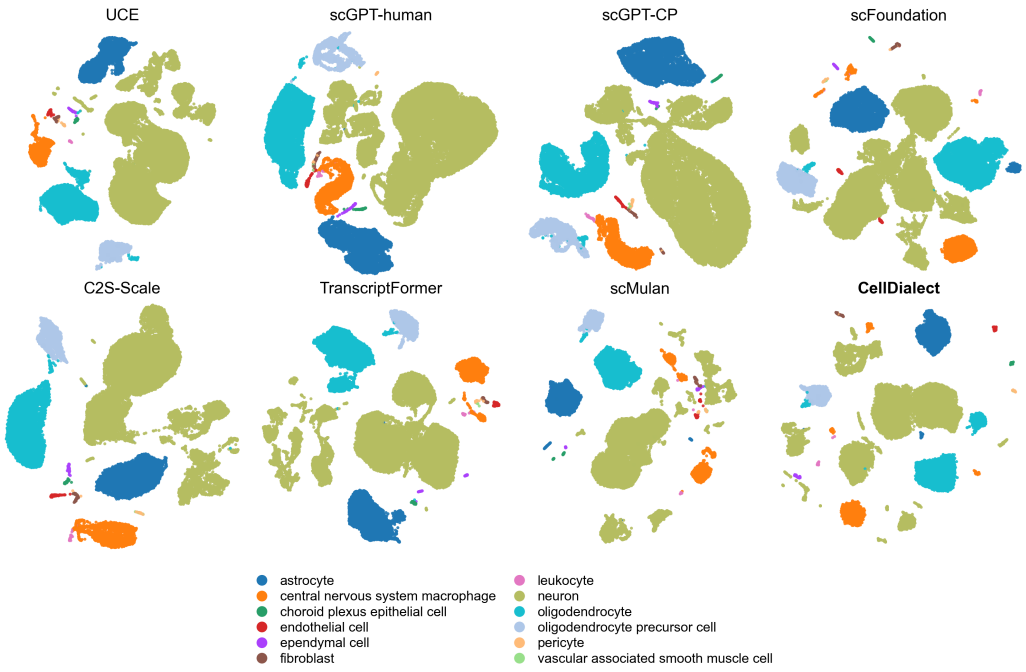

Figure 8: UMAP visualization of hippocampal formation.

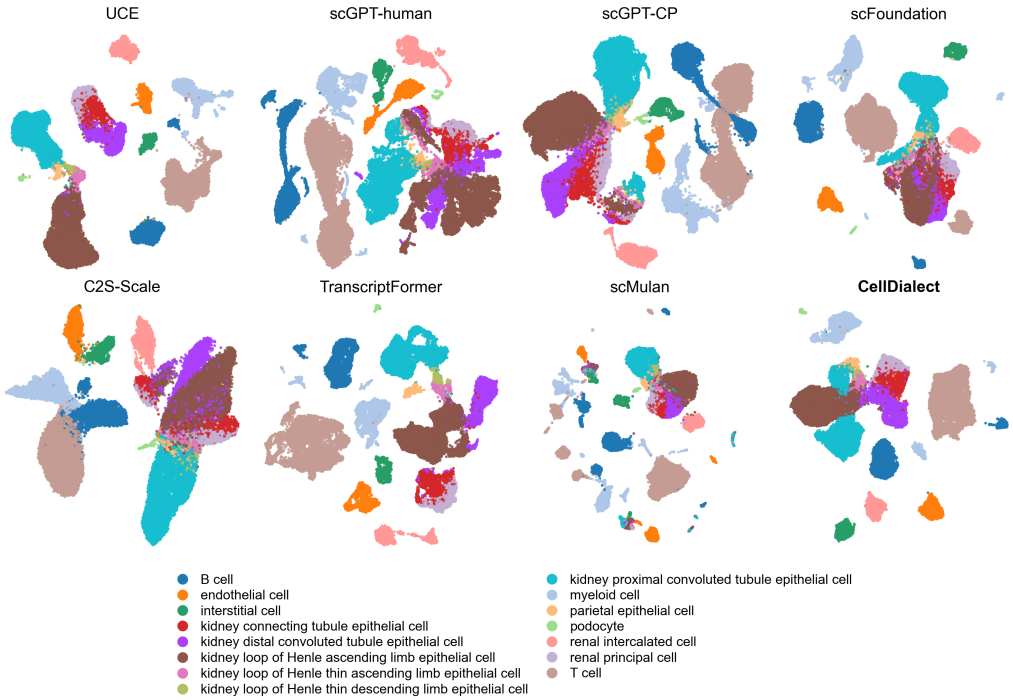

Figure 9: UMAP visualization of kidney.

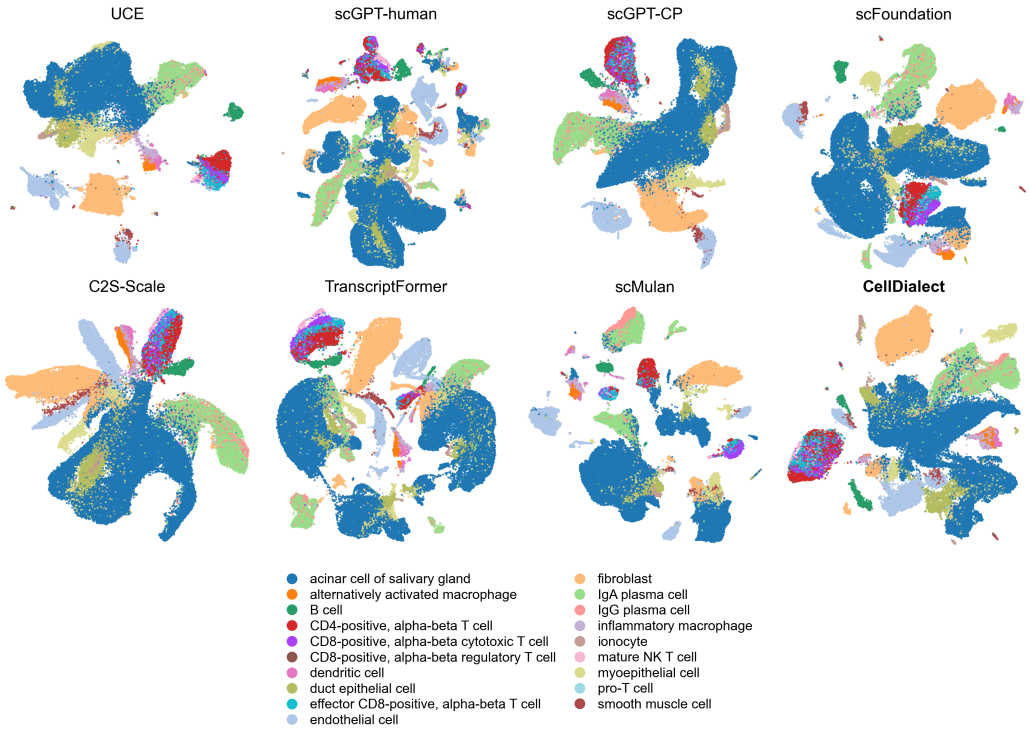

Figure 10: UMAP visualization of labial gland.

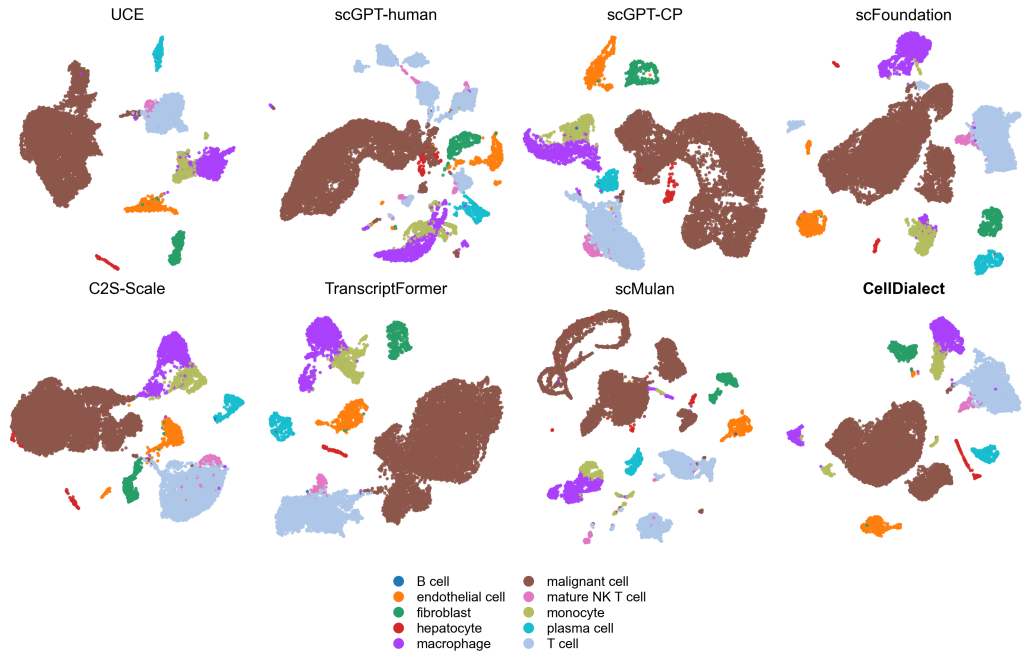

Figure 11: UMAP visualization of liver1.

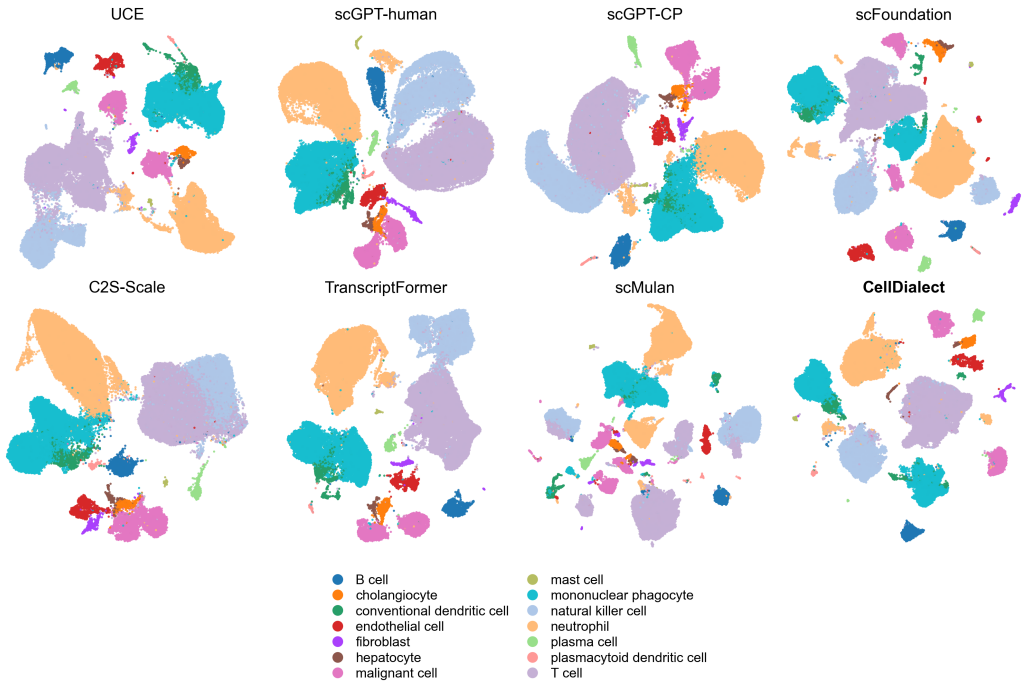

Figure 12: UMAP visualization of liver2.

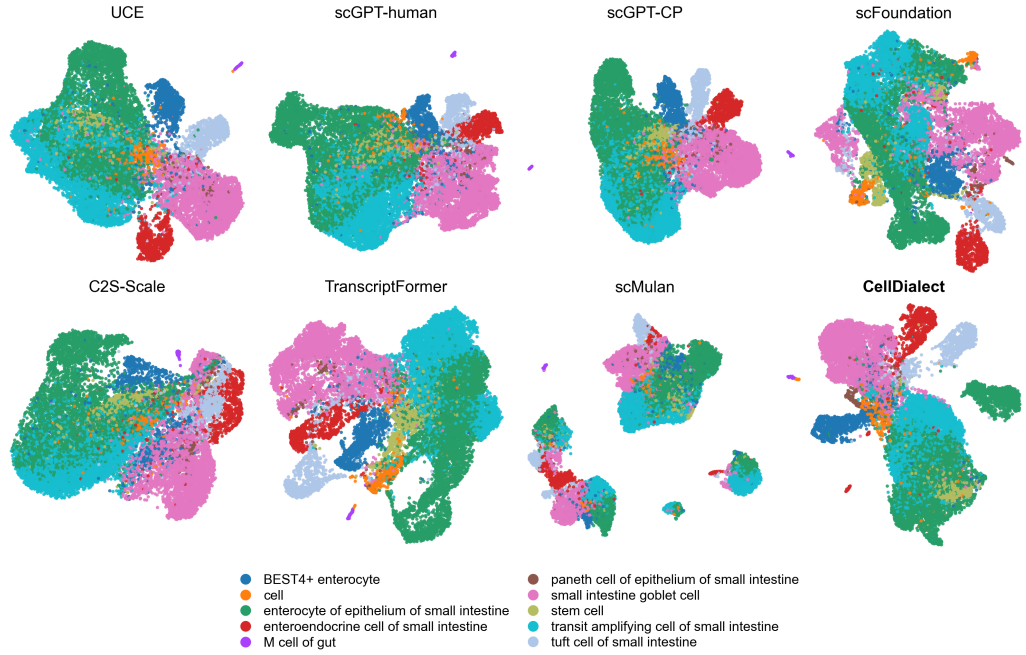

Figure 13: UMAP visualization of small intestine.

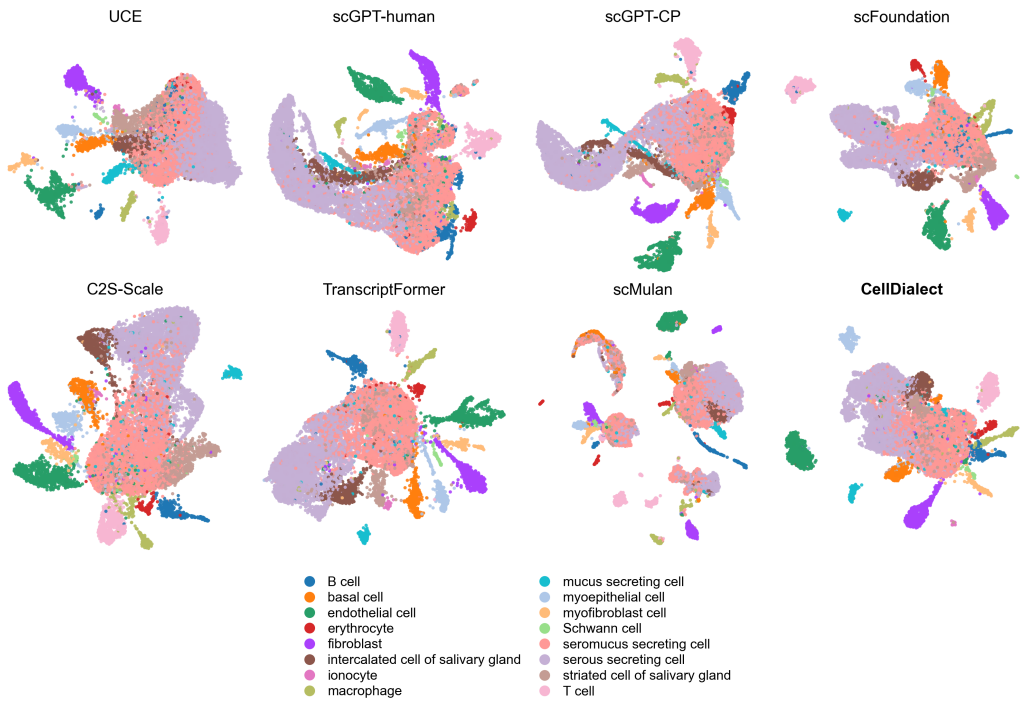

Figure 14: UMAP visualization of submandibular gland.

Table 12: Wu–Palmer scores for disease QA across normal benchmark datasets.

Method	**Brod46**	**EsoEpi**	**Hippo**	**SmaInt**	**SubMG**	**Avg**
CellWhisperer	0.000	0.001	0.000	0.001	0.002	0.001
CellDialect	**0.913**	**0.995**	**0.999**	**0.998**	**1.000**	**0.981**

E AGE PREDICTION

Age-group prediction and evaluation. We evaluate donor age-group prediction by extracting age intervals from model-generated cellular descriptions and comparing them with the reference dataset annotations. Reference and predicted intervals are mapped to the same ordered ten-year age groups, such as 30–39 and 40–49 years.

Let b_i and $\hat{b}_i$ denote the ordinal indices of the reference and predicted age groups for cell i. The age-group mean absolute error (MAE) on dataset d is

$$\mathrm{MAE}_{\mathrm{age},d} = \frac{1}{N_d} \sum_{i=1}^{N_d} \left| \hat{b}_i - b_i \right|, \tag{13}$$

where N_d is the number of evaluated cells. An error of zero indicates an exact match, whereas an error of one indicates a prediction in an adjacent age group. The overall score is the unweighted mean across the D evaluated datasets:

$$\mathrm{MAE}_{\mathrm{age}} = \frac{1}{D} \sum_{d=1}^{D} \mathrm{MAE}_{\mathrm{age},d}\,. \tag{14}$$

Lower values indicate more accurate predictions. All errors are expressed in age-group units rather than years.

Table 13 reports results on the ten held-out datasets, evaluated without dataset-specific fine-tuning. The average MAE is 1.718 age groups, with the lowest error on kidney cortex (0.830) and the highest on small intestine (2.448). These results extend the frozen language-model framework to donor age-group prediction through the same cellular interface.

Table 13: Mean absolute error(MAE) of age prediction across benchmark datasets.

	Brod46	**CorKid**	**EsoEpi**	**Hippo**	**Kidney**	**Labial**	**Liver1**	**Liver2**	**SmaInt**	**SubMG**	**Avg**
MAE (10-y bins)	2.236	0.830	1.460	1.491	1.746	1.737	1.271	1.662	2.448	2.300	1.718

F COMPLETE QUERY-TOKEN ABLATION RESULTS

F.1 RECOVERED PER-DATASET NMI EXPORT

Table 14 reports the recovered export with its original retained-token convention. Reduced settings show mean $\pm$ sample standard deviation over ten trials. The average row is the unweighted mean across the ten main-table datasets, with no pooled SD.

Table 14: NMI under query-token removal

Dataset	16 tokens	12 tokens	8 tokens	4 tokens	2 tokens
Brod46	0.708	0.690 ± 0.007	0.689 ± 0.011	0.702 ± 0.011	0.701 ± 0.016
CorKid	0.760	0.740 ± 0.003	0.747 ± 0.010	0.740 ± 0.008	0.733 ± 0.007
EsoEpi	0.369	0.363 ± 0.022	0.346 ± 0.036	0.346 ± 0.021	0.338 ± 0.033
Hippo	0.836	0.807 ± 0.036	0.818 ± 0.037	0.756 ± 0.056	0.773 ± 0.041
Kidney	0.893	0.909 ± 0.008	0.915 ± 0.008	0.908 ± 0.015	0.903 ± 0.010
Labial	0.595	0.600 ± 0.007	0.604 ± 0.005	0.598 ± 0.012	0.598 ± 0.012
Liver1	0.766	0.761 ± 0.002	0.764 ± 0.010	0.754 ± 0.012	0.752 ± 0.015
Liver2	0.876	0.913 ± 0.015	0.914 ± 0.016	0.889 ± 0.024	0.881 ± 0.039
SmaInt	0.550	0.545 ± 0.008	0.548 ± 0.010	0.524 ± 0.024	0.532 ± 0.023
SubMG	0.639	0.616 ± 0.019	0.611 ± 0.014	0.595 ± 0.028	0.590 ± 0.017
Average	0.699	0.694	0.696	0.681	0.680

F.2 PER-DATASET CELL TYPE QA UNDER QUERY-TOKEN REMOVAL

Table 15 reserves the Wu–Palmer matrix. Report reduced settings as mean $\pm$ standard deviation across ten selections; this measures token-selection variability, not training-seed variability.

Table 15: Cell type QA Wu-Palmer scores under query-token removal.

Dataset	16 tokens	12 tokens	8 tokens	4 tokens	2 tokens
Brod46	0.707	0.696 ± 0.035	0.605 ± 0.031	0.490 ± 0.095	0.451 ± 0.129
CorKid	0.922	0.873 ± 0.046	0.738 ± 0.113	0.506 ± 0.141	0.368 ± 0.098
EsoEpi	0.630	0.598 ± 0.033	0.619 ± 0.131	0.549 ± 0.124	0.457 ± 0.040
Hippo	0.992	0.975 ± 0.016	0.903 ± 0.034	0.727 ± 0.160	0.666 ± 0.182
Kidney	0.886	0.871 ± 0.017	0.764 ± 0.078	0.630 ± 0.108	0.531 ± 0.091
Labial	0.669	0.655 ± 0.031	0.628 ± 0.036	0.453 ± 0.044	0.421 ± 0.030
Liver1	0.976	0.973 ± 0.010	0.804 ± 0.242	0.498 ± 0.149	0.420 ± 0.046
Liver2	0.884	0.875 ± 0.032	0.830 ± 0.051	0.734 ± 0.032	0.688 ± 0.078
SmaInt	0.720	0.656 ± 0.083	0.610 ± 0.154	0.431 ± 0.115	0.343 ± 0.056
SubMG	0.425	0.416 ± 0.016	0.396 ± 0.029	0.395 ± 0.129	0.309 ± 0.032
Average	0.781	0.759	0.690	0.541	0.465