SPORE: SEMANTIC-AWARE GLOBAL POST-TRAINING FOR REUSABLE SINGLE-CELL REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Pretrained single-cell foundation models encode broad biological variation, yet their representations are not explicitly organized to make cell identity, biological context, and gene semantics accessible for downstream use. We introduce SPORE, a semantic-aware global post-training framework that adapts a pretrained single-cell foundation model once with heterogeneous biological supervision and yields a reusable checkpoint for deployment across studies and tasks. SPORE couples gene-semantic embedding alignment, ontology-aware cell supervision, and structured biological-context alignment to expose complementary biological semantics across scales. We evaluate the resulting checkpoint without target-specific backbone fine-tuning across complementary tests of downstream transfer, representation structure, and direct semantic inference. Across controlled evaluations, SPORE improves downstream transfer, makes biological information more accessible in frozen representations, and supports direct cell-type inference in new studies. These results support semantic-aware post-training as a means of organizing complementary biological semantics into a reusable single-cell representation.

1 INTRODUCTION

Single-cell foundation models (scFMs) aim to provide reusable representations across cell identity and state, disease-related phenotypes, biological mechanisms, and perturbational responses (Bunne et al., 2024). Recent scFMs are pretrained on large single-cell corpora using diverse learning objectives and biological priors (Cui et al., 2024; Theodoris et al., 2023; Hao et al., 2024; Rosen et al., 2026; Pearce et al., 2026). Yet large-scale pretraining does not by itself guarantee a representation that transfers readily across downstream settings (Kedzierska et al., 2025; DenAdel et al., 2026). Recent work has explored biologically guided adaptation and parameter-efficient fine-tuning (Weerasekara et al., 2026; He et al., 2026), but these approaches still target particular downstream settings, requiring separate adapted checkpoints or adapters (Figure 1). Freezing the backbone avoids repeated adaptation, but makes downstream performance depend more directly on the quality of the pretrained representation. This motivates our central question: *Can biological supervision improve a pretrained scFM once while preserving reuse across studies and tasks?*

This tension points to a representation-learning challenge. A pretrained representation may encode cell identity, biological context, or gene function while leaving these factors entangled or difficult for lightweight downstream predictors to access (Bengio et al., 2013; Xu et al., 2020). Prior work shows that biological supervision and external priors can explicitly shape single-cell representations: scCello incorporates Cell Ontology structure during foundation-model pretraining (Yuan et al., 2024); sciLaMA introduces external gene-level priors (Hu et al., 2025); and TEDDY uses tissue and disease annotations as supervised pretraining signals (Chevalier et al., 2025). These approaches introduce semantic information during model construction or pretraining, while post-training offers a practical complement: existing checkpoints can incorporate evolving semantic resources without repeating corpus-scale pretraining or specializing to a target downstream task.

We therefore study a global reuse setting in which a pretrained scFM is post-trained once on a heterogeneous corpus and then reused across target studies and tasks without target-specific backbone adaptation. In this setting, SPORE introduces supervision at three complementary semantic levels (gene, cell, and biological context) through gene-semantic alignment, ontology-aware cell supervision (Tan et al., 2026), and biological-context alignment, yielding a single reusable checkpoint. We

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

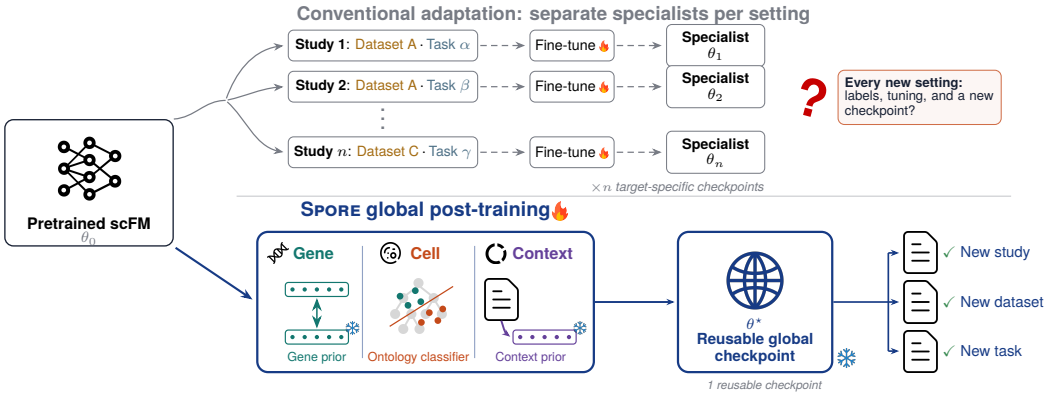

Figure 1: **One reusable checkpoint across studies and tasks.** Target-specific adaptation produces separate specialist checkpoints for individual study/task settings, whereas SPORE globally post-trains a pretrained scFM once to obtain a reusable checkpoint. The same checkpoint supports lightweight task-specific readouts on frozen representations and direct semantic inference through the learned interfaces.

evaluate both the resulting frozen representation and the semantic interfaces learned during post-training through downstream transfer, representation analysis, and direct semantic inference.

The contributions of this study are summarized as follows:

1. We formulate semantic-aware global post-training as a reusable-representation setting, where a pretrained scFM is adapted once with biological supervision and the resulting checkpoint is reused across studies and tasks without target-specific backbone adaptation.
2. We introduce SPORE, a multi-level semantic post-training framework that jointly supervises gene semantics, ontology-aware cell identity, and biological context in a shared representation.
3. Across transfer, representation analysis, and direct inference, SPORE improves hierarchy-compatible cell-type transfer across all evaluated backbones and supports direct annotation on study-disjoint data, while improving the label-efficient accessibility of cell representations.

2 Related Work

Single-cell foundation models and representation learning. scFMs incorporate biological structure into representation learning in different ways: scBERT, scGPT, Geneformer, and scFoundation learn representations primarily from large-scale transcriptomic data through self-supervised or generative objectives (Yang et al., 2022; Cui et al., 2024; Theodoris et al., 2023; Hao et al., 2024). TEDDY and scCello introduce biological annotations or ontology structure during model training (Chevalier et al., 2025; Yuan et al., 2024), while UCE and TranscriptFormer incorporate gene- or protein-level priors (Rosen et al., 2026; Pearce et al., 2026). Beyond foundation-model pretraining, SCimilarity uses metric learning to construct reusable cell representations for cross-study cell search (Heimberg et al., 2025). These approaches provide complementary ways of incorporating biological structure into learned single-cell representations.

Post-training and adaptation. Recent work also adapts single-cell models after pretraining. CellRefine uses marker and ontology guidance before downstream fine-tuning, with particular attention to long-tailed cell populations and covariate shift (Weerasekara et al., 2026). scPEFT instead studies parameter-efficient adaptation of pretrained single-cell models for downstream tasks (He et al., 2026). Both modify pretrained models toward particular target settings; SPORE instead performs a single global semantic post-training stage and reuses the resulting checkpoint without target-specific backbone adaptation.

Supervised cell annotation. Supervised cell annotation provides a complementary route to learning cell identity from labeled reference data. CellTypist and scTab learn supervised cell-type predictors from large annotated reference datasets (Domínguez Conde et al., 2022; Fischer et al., 2024), while HCE-scTab further incorporates Cell Ontology structure through hierarchical cross-entropy (Cultrera di Montesano et al., 2026). These methods treat cell identity as the primary supervised endpoint; in SPORE, ontology-aware cell supervision is one component of a broader reusable representation.

Evaluating reusable single-cell representations. Recent evaluations assess scFMs across multiple downstream tasks and transfer settings (Wu et al., 2025; Kedzierska et al., 2025; Liu et al., 2026), showing that their advantages over task-specific methods are neither uniform nor task-independent (Kedzierska et al., 2025; Liu et al., 2026). This motivates evaluating representation reuse across heterogeneous tasks rather than a single endpoint. Gene-function, regulatory-network, and perturbation benchmarks such as FRoGS (Chen et al., 2024), universal GRN inference (Qi et al., 2026), and PertEval-scFM (Wenteler et al., 2025) further extend evaluation beyond cell identity.

3 METHODS

3.1 PROBLEM SETUP AND DEPLOYMENT PROTOCOL

We consider a pretrained single-cell foundation model f_{θ_0}. For a single-cell expression profile x_i, we write its cell-level and contextualized gene-token representations as

$$z_i^{(0)} = f_{\theta_0}^{\text{cell}}(x_i), \qquad H_i^{(0)} = f_{\theta_0}^{\text{gene}}(x_i). \tag{1}$$

Global post-training uses a heterogeneous corpus $\mathcal{D}_{\text{global}} = \bigcup_{m=1}^{M} \mathcal{D}_m$ assembled from multiple source studies or datasets. Each $\mathcal{D}_m = \{(x_i, y_i, s_i)\}_{i=1}^{N_m}$ contains human single-cell transcriptomic profiles x_i together with available cell-identity annotations y_i and biological-context metadata s_i, which may be partially observed. Using $\mathcal{D}_{\text{global}}$ together with fixed auxiliary semantic resources $\mathcal{R}_{\text{sem}}$ for semantic supervision, a single global post-training stage adapts the model and learns semantic interfaces ϕ, yielding a reusable global checkpoint

$$\Theta^\star = (\theta^\star, \phi^\star) = \text{PostTrain}\big(\theta_0, \mathcal{D}_{\text{global}}; \mathcal{R}_{\text{sem}}\big). \tag{2}$$

A downstream setting consists of a target dataset–task pair $\mathcal{S}_j = (\mathcal{D}_j^{\text{tar}}, \mathcal{T}_j)$. The same checkpoint $\Theta^\star$ is reused across such settings: changing $\mathcal{D}_j^{\text{tar}}$ corresponds to a new dataset, while changing $\mathcal{T}_j$ corresponds to a new downstream task. No target-specific backbone adaptation is performed. We evaluate this reuse through frozen-feature transfer, where lightweight task-specific readouts may be fitted on the fixed representation, and probe-free semantic inference, where the learned semantic interfaces are used directly without fitting a target-specific predictor. Full evaluation protocols and model-specific interfaces are given in Appendices C–D.

3.2 MULTI-LEVEL SEMANTIC SUPERVISION

To shape the reusable representation during global post-training, SPORE uses three complementary supervision branches: gene-semantic alignment, ontology-aware cell identity, and biological-context alignment (Figure 2).

Ontology-aware cell identity. For each cell with Cell Ontology label y_i, the cell-type head produces logits $\ell_i \in \mathbb{R}^{|\mathcal{C}|}$ over the observed output classes $\mathcal{C}$. We represent the Cell Ontology (Tan et al., 2026) `is_a` relations as a directed acyclic graph $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{C} \subseteq \mathcal{V}$ and intermediate ontology nodes are retained. Let $\text{Anc}^+(c)$ denote the set containing class c and all of its ancestors, and let $d_\uparrow(c, a)$ be the shortest upward path length from c to an ancestor a.

For a candidate output class $c \in \mathcal{C}$ and true label $y \in \mathcal{C}$, we define the asymmetric common-ancestor cost

$$\kappa_\beta(c, y) = \min_{a \in \text{Anc}^+(c) \cap \text{Anc}^+(y)} \left[\beta\, d_\uparrow(c, a) + d_\uparrow(y, a)\right], \qquad 0 \le \beta \le 1. \tag{3}$$

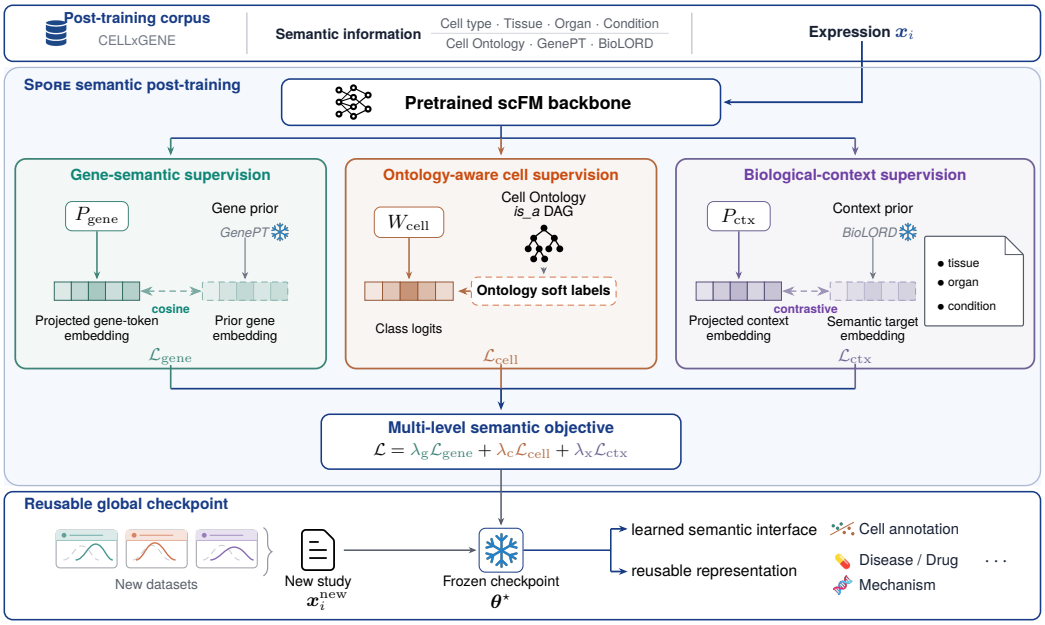

Figure 2: **Method overview.** SPORE globally post-trains a shared scFM backbone with gene-semantic, ontology-aware cell-identity, and biological-context supervision, yielding one reusable checkpoint across downstream settings.

The coefficient β controls the asymmetry between additional specificity and loss of specificity. If the candidate class c is a descendant of the annotated label y, its additional specificity is weighted by β; if c is an ancestor of y, the lost specificity retains full cost. Because Cell Ontology is a DAG with multiple inheritance, we minimize over all shared ancestors rather than assuming a unique lowest common ancestor. The resulting behavior for coarse, refining, and wrong-branch predictions is illustrated in Appendix Figure A1.

We convert these costs into soft targets over the observed output classes. For temperature $\tau > 0$, the target probability assigned to candidate c is $q_y(c) \propto \exp[-\kappa_\beta(c, y)/\tau]$. The resulting Asymmetric Ontology Soft-Label loss (AOSL) applies standard soft-label cross-entropy, building on hierarchy-aware classification objectives (Bertinetto et al., 2020; Cultrera di Montesano et al., 2026):

$$q_y(c) = \frac{\exp[-\kappa_\beta(c, y)/\tau]}{\sum_{c' \in \mathcal{C}} \exp[-\kappa_\beta(c', y)/\tau]},$$
$$\mathcal{L}_{\text{cell}} = -\frac{1}{B} \sum_{i=1}^{B} \sum_{c \in \mathcal{C}} q_{y_i}(c) \log p_{\theta,\phi}(c \mid x_i), \qquad p_{\theta,\phi}(c \mid x_i) = \text{softmax}(\ell_i)_c. \tag{4}$$

The cost and soft-target matrices are computed offline from a pinned ontology release; ontology construction and implementation checks are reported in Appendix A.

Biological-context alignment. Let s_i denote the canonicalized biological context of cell i, comprising tissue, organ-level context, and disease or physiological state, instantiated from the CELLxGENE `tissue`, `tissue_general`, and `disease` fields. Donor and dataset identifiers are excluded because they are not intended semantic targets and may encode study-specific shortcuts.

We serialize s_i into a fixed textual description and obtain its semantic embedding with a frozen biomedical encoder T, writing $T(s_i)$ for the encoding of this description. In the primary configuration, T is BioLORD-2023 (Remy et al., 2024). A shared projector P_{ctx} maps the current cell representation $z_i = f_\theta^{\text{cell}}(x_i)$ into the same semantic space:

$$\hat{e}_i^{\text{ctx}} = \text{Normalize}\big(P_{\text{ctx}}(z_i)\big), \qquad e_{s_i}^{\text{ctx}} = \text{Normalize}\big(T(s_i)\big). \tag{5}$$

We consider two objectives over this shared semantic space. Cosine regression aligns each cell with its positive context, $\mathcal{L}_{\text{ctx}}^{\cos} = B^{-1}\sum_{i=1}^{B}[1-\cos(\hat{e}_i^{\text{ctx}}, e_{s_i}^{\text{ctx}})]$. Our primary contrastive context objective scores the positive target against a fixed bank of canonical context embeddings from the training split of the post-training corpus. Let $\mathcal{U}_{\text{train}}$ denote these contexts, with frozen embeddings e_u^{ctx}. Its objective is

$$\mathcal{L}_{\text{ctx}}^{\text{ctr}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\big(\cos(\hat{e}_i^{\text{ctx}}, e_{s_i}^{\text{ctx}})/\tau_c\big)}{\sum_{u\in\mathcal{U}_{\text{train}}}\exp(\cos(\hat{e}_i^{\text{ctx}}, e_u^{\text{ctx}})/\tau_c)}. \tag{6}$$

The bank contains one embedding per unique training context, providing a fixed candidate set across optimization steps. We use $L_{\text{ctx}}^{\text{ctr}}$ by default; UCE uses $L_{\text{ctx}}^{\cos}$, with backbone-specific settings detailed in Appendix C.5.

Gene-semantic alignment. For backbones exposing a trainable model-level gene representation, the gene-semantic branch operates on this interface rather than on the cell-specific contextualized token states H_i. Let $\mathcal{G}_{\text{sem}}$ contain the genes with both a validated model-level representation v_g and a frozen GenePT target $r_g \in \mathbb{R}^{d_r}$ (Chen & Zou, 2025). At optimization step t, we sample $S_t \subseteq \mathcal{G}_{\text{sem}}$ with $|S_t| = M$.

A trainable projector maps each sampled gene representation into the semantic prior space:

$$\hat{r}_g = \text{Normalize}\big(P_{\text{gene}}(v_g)\big). \tag{7}$$

We align the projected representation with its fixed semantic target using normalized cosine regression:

$$\mathcal{L}_{\text{gene}} = \frac{1}{|S_t|}\sum_{g\in S_t}\left[1-\cos(\hat{r}_g, r_g)\right]. \tag{8}$$

Unlike the cell and context objectives, this loss is defined over sampled gene representations and is independent of the cells in the current expression mini-batch. Prior coverage, normalization, and gene-row sampling are detailed in Appendix A.3.

3.3 Joint Objective and Optimization

The primary objective combines the three semantic branches:

$$\mathcal{L}_{\text{sem}} = \lambda_{\text{gene}}\mathcal{L}_{\text{gene}} + \lambda_{\text{cell}}\mathcal{L}_{\text{cell}} + \lambda_{\text{ctx}}\mathcal{L}_{\text{ctx}}. \tag{9}$$

Each branch loss is averaged over its valid supervision units before weighting: cell and context losses use valid cells, whereas the gene-semantic loss uses sampled covered gene rows. Missing annotations mask the corresponding objective rather than providing negative supervision. Branch-weight settings are selected using the post-training validation split and fixed before downstream evaluation; optimization schedules are reported in Appendix C.

The active objectives also depend on the gene interface exposed by each backbone. Geneformer, scCello, and TEDDY-G use the full three-branch objective. For UCE and TF-Sapiens, we use the cell and context objectives only and set $\lambda_{\text{gene}} = 0$, because their native gene interfaces are fixed ESM-derived vocabulary embeddings rather than the trainable model-level gene representations targeted by the GenePT alignment. This preserves the same semantic post-training framework while respecting the native representation interface of each backbone.

Two-stage semantic post-training. Training separates semantic-interface initialization from representation adaptation. In Stage 1, the pretrained backbone is frozen and the semantic interfaces ϕ are trained on its fixed representations. In Stage 2, these interfaces are initialized from Stage 1 and jointly optimized with the trainable model parameters:

$$\begin{aligned}\phi^{(1)} &= \arg\min_{\phi}\mathcal{L}_{\text{sem}}(\theta_0, \phi),\\ (\theta^\star, \phi^\star) &= \arg\min_{\theta,\phi}\mathcal{L}_{\text{sem}}(\theta, \phi), \qquad (\theta, \phi)_0 = (\theta_0, \phi^{(1)}).\end{aligned} \tag{10}$$

Trainable model parameters use a smaller learning rate than the semantic interfaces. Geneformer, scCello, and TEDDY-G use full-model adaptation in Stage 2. UCE updates only the top eight backbone layers while keeping its gene-embedding interface frozen, whereas TF-Sapiens uses low-learning-rate full-backbone adaptation with its gene-embedding interface frozen. Parameter groups, learning rates, and checkpoint details are reported in Appendix C.

4 EXPERIMENTS

We evaluate whether global semantic post-training improves reuse without target-specific backbone adaptation. We consider frozen transfer across backbones and tasks, representation geometry and low-label accessibility, direct inference with the learned cell-identity interface on new studies, and supervision comparisons that examine which branches support transfer. Evaluation protocols and additional results appear in Appendices D–E.

4.1 FROZEN DOWNSTREAM TRANSFER

We post-train five representative backbones spanning diverse pretraining designs (Geneformer, scCello, TEDDY-G, UCE, and TranscriptFormer-Sapiens; see Appendix C.2). Post-training uses the training partition of a donor-split 100k-cell human CELLxGENE Census LTS 2025-11-08 (CZI Cell Science Program et al., 2025), with held-out donors used for cell-type evaluation. One selected checkpoint for each backbone is frozen and reused across all applicable tasks in Table 1, with matched downstream readouts, splits, and protocols for frozen and post-trained models.

The tasks span cell-type transfer, Cancer-Finder and drug sensitivity (Wu et al., 2025), tissue-specificity prediction (Chen et al., 2024), UGRN inference (Qi et al., 2026), and Norman perturbation prediction (Wenteler et al., 2025), using hAcc, AUROC, AUC, accuracy, AUPRC, and PertEval Norman_1 ΔAUSPC (10^{-2}), respectively. Cell-type hAcc accepts either the reference label or a valid Cell Ontology descendant; full protocols, additional metrics, and per-task results are reported in Appendix D and Tables A5–A10.

Table 1: **Frozen downstream transfer.** Boldface scores denote improvements over the paired frozen backbone; shading and underlining mark the best and second-best results in the table. Additional task-specific references are reported in Table A4. † UCE and TF-Sapiens use cell and context supervision only, with their native fixed ESM-derived interfaces retained for gene-level tasks.

Method	Setting	Cell type ↑	Cancer ↑	Drug sens. ↑	Tissue spec. ↑	UGRN ↑	Perturb. ↑
Simple non-FM baseline	task-specific	0.778	0.735	0.631	0.353	N/A	0.092
scGPT (Cui et al., 2024)	frozen	0.772	0.555	0.793	0.769	0.677	0.039
scFoundation (Hao et al., 2024)	frozen	0.654	0.806	0.726	0.319	0.622	0.078
Geneformer (Theodoris et al., 2023)	frozen	0.649	0.684	0.592	0.803	0.677	-0.058
	SPORE PT	**0.822**	**0.743**	**0.692**	**0.814**	0.675	**-0.028**
scCello (Yuan et al., 2024)	frozen	0.747	0.773	0.629	0.613	0.661	-0.013
	SPORE PT	**0.833**	**0.787**	**0.651**	**0.633**	0.661	**-0.010**
TEDDY-G (Chevalier et al., 2025)	frozen	0.665	0.734	0.650	0.576	0.676	-0.053
	SPORE PT	**0.824**	**0.748**	**0.750**	**0.627**	**0.680**	**-0.003**
UCE (Rosen et al., 2026)	frozen	0.839	0.797	0.547	0.563	0.703	0.077
	SPORE PT†	**0.874**	**0.799**	0.523	0.563	0.703	**0.087**
TF-Sapiens (Pearce et al., 2026)	frozen	0.871	0.726	0.535	0.525	0.710	0.152
	SPORE PT†	**0.894**	**0.777**	**0.552**	0.525	0.710	**0.156**

Across all five paired backbones, post-training improves cell-type hAcc and Cancer-Finder AUROC, while drug-sensitivity AUC improves for four (Table 1). Cancer-Finder and drug sensitivity are not direct post-training objectives, making these gains evidence of transfer beyond the supervised semantic targets. Tissue specificity improves for the three backbones with active GenePT alignment, whereas UGRN changes little and perturbation performance is largely preserved. Overall, the clearest gains are on cell-level transfer, especially cell type and Cancer-Finder, while gene-level, regulatory, and perturbation results are more mixed.

All five post-trained backbones outperform the simple non-FM references on cell type, Cancer-Finder, and tissue specificity, and three do so on drug sensitivity. Stronger task-specific baselines give a more mixed picture: several SPORE backbones exceed scVI (0.765 AUROC) on Cancer-Finder, whereas FRoGS-ArchS4 remains higher than the best tissue-specificity result (0.859 vs. 0.814; Table A4). Model size alone also does not track these results closely. After post-training, the 10M-parameter Geneformer and scCello are competitive with much larger backbones on several tasks.

Geneformer gives a concrete example of checkpoint reuse: the same checkpoint raises hAcc from 0.649 to 0.822, Cancer AUROC from 0.684 to 0.743, and drug-sensitivity AUC from 0.592 to 0.692. We use this backbone for the representation analysis below.

The task-specific pattern also depends on which representation is evaluated. Cancer-Finder and drug sensitivity use frozen cell embeddings, whereas tissue specificity and UGRN use model-level gene representations. UCE and TF-Sapiens retain their fixed ESM-derived gene interfaces because their SPORE configurations omit GenePT alignment, so their gene-level scores are unchanged.

4.2 Representation Accessibility and Geometry

We next examine how post-training changes the frozen Geneformer representation under a matched comparison. The frozen and post-trained models use the same 6-layer, 256-dimensional backbone family and identical extraction interfaces, so the comparison isolates the effect of post-training rather than architecture or representation dimensionality (Figure 3).

We first examine cross-dataset representation geometry. Within each of 39 tissue strata, we compare pairwise embedding similarity with negative Cell Ontology distance and macro-average the resulting Spearman correlations across tissues. At the first-token readout used for downstream transfer, macro Spearman ρ increases from 0.255 to 0.386, indicating a stronger correspondence between embedding geometry and ontology structure across datasets. In contrast, the mean-pool control changes only from 0.075 to 0.097. This interface contrast suggests that the change is concentrated in the representation actually used for downstream transfer rather than appearing uniformly across pooling choices.

We next test how readily task-relevant information can be extracted across downstream label budgets. A linear Cancer-Finder readout is trained on increasing fractions of the available labels and evaluated on the same four held-out domains. The post-trained representation achieves higher held-out AUROC at every label fraction, with the largest gain, 0.055, at 1% labels. The geometry and label-efficiency analyses therefore provide complementary views of representation reuse: one measures cross-dataset biological organization, while the other measures how readily a simple readout can recover task-relevant signal, especially under limited supervision. Numerical endpoints and full protocols are reported in Table A11 and Appendix D.

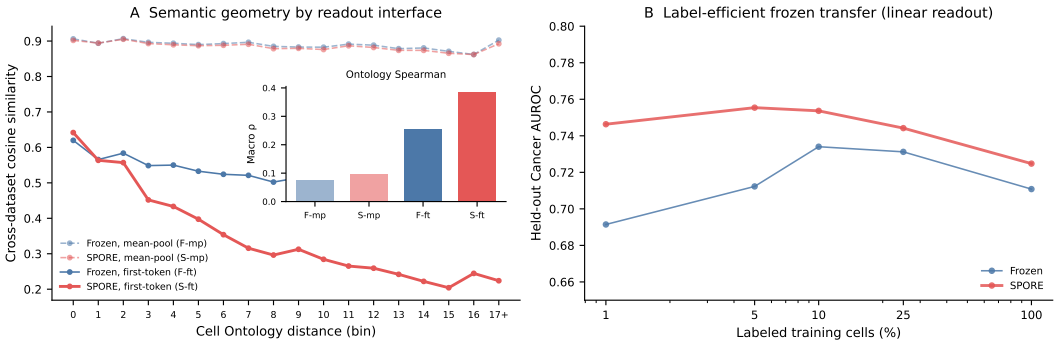

Figure 3: **Matched Geneformer representation analysis.** (A) Cross-dataset cosine similarity versus Cell Ontology distance for first-token and mean-pool readouts; the inset reports macro Spearman ρ between cosine similarity and negative ontology distance. (B) Held-out Cancer-Finder AUROC of a linear readout across increasing label fractions; points report the mean across four held-out domains.

4.3 Probe-Free Semantic Inference

The learned ontology head provides cell identities directly without a target-specific classifier or backbone update. We evaluate donor-held-out ID cells and 10,500 cells from 11 temporally screened, study-disjoint OOD studies, with no dataset-ID, collection-ID, or normalized-DOI overlap with post-training data. ID hMacro-F1 uses the fixed 164-class candidate universe; OOD hMacro-

Table 2: **Direct cell-type inference on donor-held-out and study-disjoint cells.** All methods use frozen interfaces. hAcc and hMacro-F1 are hierarchy-compatible; hF measures ancestor overlap and LCAE measures symmetric ontology displacement among strict errors (lower is better).

	Donor-held-out ID				Study-held-out OOD			
Method	hAcc ↑	hMacro-F1 ↑	hF ↑	LCAE ↓	hAcc ↑	hMacro-F1 ↑	hF ↑	LCAE ↓
CellTypist	0.6930	0.5305	0.9302	2.7920	0.5538	0.4755	0.8039	3.7314
scTab–Linear	0.7492	0.6021	0.9390	2.8564	0.5671	0.4694	0.8318	3.3042
scTab–MLP	0.7714	0.6102	0.9414	2.7766	0.6093	0.5259	0.8360	3.3674
scTab–TabNet	0.7567	0.5955	0.9373	2.9303	0.5894	0.4800	0.8093	3.8482
HCE-scTab TabNet	0.8668	0.6489	0.8920	2.8874	0.6531	0.5877	0.7862	3.9444
SPORE–Geneformer	0.8404	0.5359	0.9263	2.9274	0.6025	0.5030	0.8330	3.4205
SPORE–UCE	0.8852	0.6428	**0.9695**	3.1903	0.6150	0.5281	**0.8593**	3.1240
SPORE–TF-Sapiens	**0.9087**	**0.6688**	0.9660	**2.7515**	**0.6623**	**0.5912**	0.8560	**2.9577**

F1 averages over the 56 reference classes observed in the held-out studies. Cohort construction and metric definitions are given in Appendix D.7.

SPORE–TF-Sapiens achieves the highest hAcc and hMacro-F1 among the compared methods on both ID and OOD cells (Table 2), with hAcc of 0.909 and 0.662 versus 0.867 and 0.653 for HCE-scTab TabNet. SPORE–UCE has the highest ancestor-overlap hF, while SPORE–TF-Sapiens has the lowest LCAE; Appendix E.3 provides the reference-side audit.

As a downstream case, we test whether direct annotations preserve regional cell composition in a held-out developing human neocortex atlas (Wang et al., 2025). The fine-area subset contains 699 nuclei from 13 donors with BA17, BA9, or BA10 labels. Without target-specific fitting, predictions are collapsed into five prespecified biological compartments for regional comparison (Figure 4). Before aggregation, direct annotation reaches 72.7% exact and 81.7% descendant-compatible accuracy.

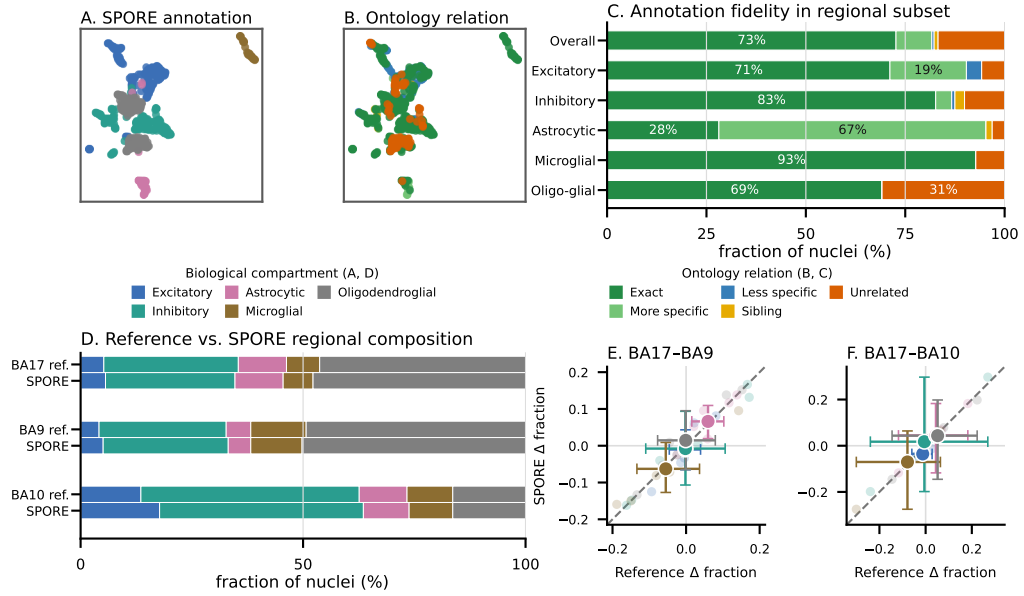

Figure 4: **Direct annotation preserves donor-paired regional composition.** Held-out developing human neocortex atlas. (A–C) Direct annotations and ontology relations. (D) Regional composition. (E–F) Donor-paired regional contrasts; error bars show bootstrap 95% CIs.

Donor-matched regional contrasts remain closely aligned with the study annotations: BA17–BA9 has six paired donors, Pearson $r = 0.980$, and mean absolute error (MAE) 0.009; BA17–BA10 has

three paired donors, $r = 0.945$, and MAE 0.012 (Figure 4E–F). Despite differences in cell-level ontology granularity, direct annotations preserve the regional composition signal after aggregation. Full sample accounting, bootstrap intervals, and paired results are reported in Appendix F.1.

4.4 Supervision Ablations and Data Efficiency

We next examine how the supervision branches contribute to frozen transfer on Geneformer. Table 3 compares the joint SPORE objective with flat cell supervision, ontology-aware cell supervision, context supervision, and a gene-semantic control under the same downstream tasks and splits. The Gene semantic condition trains only the gene vocabulary and projector, isolating gene-semantic alignment from broader backbone adaptation.

Table 3: **Supervision-component comparison on Geneformer.** Bold entries indicate improvements of SPORE over Frozen; shading and underlining denote the best and second-best results in each task, respectively.

Task	Frozen	Cell (flat)	Cell (ontology)	Context	Gene semantic	SPORE
Cell type (hAcc)	0.649	0.795	0.821	0.807	0.665	**0.822**
Cancer (AUROC)	0.684	0.729	0.742	0.738	0.673	**0.743**
Drug sensitivity (AUC)	0.592	0.660	0.687	0.716	0.655	**0.692**
Tissue spec. (Acc.)	0.803	0.881	0.864	0.881	0.847	**0.814**
UGRN (AUPRC)	0.677	0.674	0.677	0.679	0.677	0.675

The comparison first isolates the contribution of ontology-aware cell supervision. Relative to Cell (flat), Cell (ontology) increases hAcc from 0.795 to 0.821 and Cancer AUROC from 0.729 to 0.742, while the joint configuration reaches 0.822 and 0.743. Other branches favor different readouts: Context gives the highest drug-sensitivity AUC (0.716), whereas the Gene semantic projector-only control raises tissue-specificity accuracy from the frozen 0.803 to 0.847. The joint checkpoint exceeds Frozen on four of the five displayed endpoints, combining the strongest cell-type and cancer results with gains on drug sensitivity and tissue-specificity transfer. The branches therefore affect different downstream readouts, while the joint objective provides gains across several of them.

We also vary the amount of post-training data under a fixed 2,000-step budget. Cell hAcc increases across the nested 1%–100% subsets, exceeding the frozen reference at 10% (0.679 versus 0.649) and reaching 0.778 at 100%. Cancer AUROC increases from 0.674 at 1% to 0.716 at 100% and exceeds its frozen reference from 3% onward, whereas drug-sensitivity AUC varies across fractions. Thus the endpoints respond differently to additional post-training data under fixed compute, with the clearest data-efficiency trend in ontology-compatible cell identity (Appendix D.8, Table A15). Appendix E.4.2 reports an auxiliary one-at-a-time sensitivity analysis over the context and gene loss weights and the backbone learning rate.

5 Conclusion

SPORE studies semantic-aware global post-training for reusable single-cell representations. Across multiple backbones, it improves hierarchy-compatible cell-type transfer and cancer and drug-sensitivity performance while largely preserving gene-level, regulatory, and perturbation performance. Representation analyses show improved cross-dataset ontology organization and low-label accessibility, while the learned ontology interface supports direct inference on donor-held-out and study-disjoint cells.

Our evaluation focuses on human transcriptomic data and depends on the quality of biological annotations and semantic resources. More broadly, the results motivate a representation-centric view of scFM development: beyond model or pretraining-data scale, an important question is which biological information should shape the representation, when it should be introduced, and how different supervision signals affect downstream transfer. Understanding the relationship among pretraining, semantic post-training, representation geometry, and downstream transfer may be as important as further scaling, and may guide extensions across species, modalities, and richer supervision.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on research methodology and experiments, implement methods, assist with translation, clean and reformat datasets, and interpret results.

We did not use generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, or propose or refine research hypotheses. Research hypotheses and the final methodological decisions and scientific interpretations were determined by the authors.

Generating synthetic datasets, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, and qualitative or thematic data analysis were not applicable to this work.

Additionally, we used generative AI tools for brainstorming, sourcing and searching for information, identifying relevant literature, drafting portions of the manuscript, editing the manuscript for readability, and creating or editing software code. We reviewed all AI-assisted work. For literature discovery, generative AI tools provided candidate references and search suggestions; the authors independently searched for, carefully read, critically evaluated, and selected the cited works against the original sources. AI-assisted code was reviewed and tested for correctness. All AI-assisted manuscript text was reviewed, verified, and revised sentence by sentence by the authors, who determined the final wording. Research hypotheses, final methodological decisions, and scientific interpretations were determined by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work relies exclusively on publicly released, de-identified single-cell datasets and does not involve the collection of new biological samples or participant-level data. The resulting models are intended solely for research use and should not be interpreted as clinical diagnostic systems.

REPRODUCIBILITY STATEMENT

The proposed objectives and training protocol are specified in Section 3 and Appendices A–C. Appendix D defines the benchmark tasks and metrics, while Appendices B–C record preprocessing, splits, optimization settings, external-resource provenance, and model-specific interfaces; supplementary results are provided in Appendix E.

We provide an anonymized code snapshot as supplementary material, including reference post-training YAML configurations for all evaluated backbones, downstream evaluation and result-collection scripts, and environment instructions. An identical browsable mirror is available at `https://anonymous.4open.science/r/SPORE-0B21/`.

Public datasets and pretrained checkpoints are obtained from their original providers and are not redistributed; large derived embeddings and trained checkpoints are likewise not included in the submission snapshot. Appendices B–D and the code README document the external resources, experimental settings, and evaluation procedures for reconstructing the pipeline.

REFERENCES

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

Luca Bertinetto, Romain Mueller, Konstantinos Tertikas, Sina Samangooei, and Nicholas A. Lord. Making Better Mistakes: Leveraging Class Hierarchies With Deep Networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12503–12512, 2020.

Charlotte Bunne, Yusuf Roohani, Yanay Rosen, Ankit Gupta, Xikun Zhang, Marcel Roed, Theo Alexandrov, Mohammed AlQuraishi, Patricia Brennan, Daniel B. Burkhardt, Andrea Califano, Jonah Cool, Abby F. Dernburg, Kirsty Ewing, Emily B. Fox, Matthias Haury, Amy E. Herr,

Eric Horvitz, Patrick D. Hsu, Viren Jain, Gregory R. Johnson, Thomas Kalil, David R. Kelley, Shana O. Kelley, Anna Kreshuk, Tim Mitchison, Stephani Otte, Jay Shendure, Nicholas J. Sofroniew, Fabian Theis, Christina V. Theodoris, Srigokul Upadhyayula, Marc Valer, Bo Wang, Eric Xing, Serena Yeung-Levy, Marinka Zitnik, Theofanis Karaletsos, Aviv Regev, Emma Lundberg, Jure Leskovec, and Stephen R. Quake. How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell*, 187(25):7045–7063, 2024.

Hao Chen, Frederick J. King, Bin Zhou, Yu Wang, Carter J. Canedy, Joel Hayashi, Yang Zhong, Max W. Chang, Lars Pache, Julian L. Wong, Yong Jia, John Joslin, Tao Jiang, Christopher Benner, Sumit K. Chanda, and Yingyao Zhou. Drug target prediction through deep learning functional representation of gene signatures. *Nature Communications*, 15(1):1853, 2024.

Yiqun Chen and James Zou. Simple and effective embedding model for single-cell biology built from ChatGPT. *Nature Biomedical Engineering*, 9(4):483–493, 2025.

Alexis Chevalier, Soumya Ghosh, Urvi Awasthi, James Watkins, Julia Bieniewska, Nichita Mitrea, Olga Kotova, Kirill Shkura, Andrew Noble, Michael J. Steinbaugh, Vijay Sadashivaiah, George Dasoulas, Julien Delile, Christoph Meier, Leonid Zhukov, Iya Khalil, Srayanta Mukherjee, and Judith Mueller. Teddy: A FAMILY OF FOUNDATION MODELS FOR UNDERSTANDING SINGLE CELL BIOLOGY. In *ICML 2025 Generative AI and Biology (GenBio) Workshop*, 2025.

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.

Sebastiano Cultrera di Montesano, Davide D'Ascenzo, Srivatsan Raghavan, Ava P. Amini, Peter S. Winter, and Lorin Crawford. Improving atlas-scale single-cell annotation models with hierarchical cross-entropy loss. *Nature Computational Science*, 6(3):243–249, 2026.

CZI Cell Science Program, Shibla Abdulla, Brian Aevermann, Pedro Assis, Seve Badajoz, Sidney M Bell, Emanuele Bezzi, Batuhan Cakir, Jim Chaffer, Signe Chambers, J Michael Cherry, Tiffany Chi, Jennifer Chien, Leah Dorman, Pablo Garcia-Nieto, Nayib Gloria, Mim Hastie, Daniel Hegeman, Jason Hilton, Timmy Huang, Amanda Infeld, Ana-Maria Istrate, Ivana Jelic, Kuni Katsuya, Yang Joon Kim, Karen Liang, Mike Lin, Maximilian Lombardo, Bailey Marshall, Bruce Martin, Fran McDade, Colin Megill, Nikhil Patel, Alexander Predeus, Brian Raymor, Behnam Robatmili, Dave Rogers, Erica Rutherford, Dana Sadgat, Andrew Shin, Corinn Small, Trent Smith, Prathap Sridharan, Alexander Tarashansky, Norbert Tavares, Harley Thomas, Andrew Tolopko, Meghan Urisko, Joyce Yan, Garabet Yeretssian, Jennifer Zamanian, Arathi Mani, Jonah Cool, and Ambrose Carr. CZ CELLxGENE Discover: A single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *Nucleic Acids Research*, 53(D1):D886–D900, 2025.

Alan DenAdel, Madeline Hughes, Akshaya Thoutam, Anay Gupta, Andrew W. Navia, Nicolo Fusi, Srivatsan Raghavan, Peter S. Winter, Ava P. Amini, and Lorin Crawford. Evaluating the role of pretraining dataset size and diversity on single-cell foundation model performance. *Nature Methods*, 23(7):1447–1457, 2026.

C. Domínguez Conde, C. Xu, L. B. Jarvis, D. B. Rainbow, S. B. Wells, T. Gomes, S. K. Howlett, O. Suchanek, K. Polanski, H. W. King, L. Mamanova, N. Huang, P. A. Szabo, L. Richardson, L. Bolt, E. S. Fasouli, K. T. Mahbubani, M. Prete, L. Tuck, N. Richoz, Z. K. Tuong, L. Campos, H. S. Mousa, E. J. Needham, S. Pritchard, T. Li, R. Elmentaite, J. Park, E. Rahmani, D. Chen, D. K. Menon, O. A. Bayraktar, L. K. James, K. B. Meyer, N. Yosef, M. R. Clatworthy, P. A. Sims, D. L. Farber, K. Saeb-Parsy, J. L. Jones, and S. A. Teichmann. Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science*, 376(6594):eabl5197, 2022.

Felix Fischer, David S. Fischer, Roman Mukhin, Andrey Isaev, Evan Biederstedt, Alexandra-Chloé Villani, and Fabian J. Theis. scTab: Scaling cross-tissue single-cell annotation models. *Nature Communications*, 15(1):6611, 2024.

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.

Fei He, Ruixin Fei, Jordan E. Krull, Yang Yu, Xinyu Zhang, Xianyu Wang, Hao Cheng, Mingyue Gao, Li Su, Yibo Chen, Jinpu Li, Baichuan Jin, Yuzhou Chang, Anjun Ma, Qin Ma, and Dong Xu. Harnessing the power of single-cell large language models with parameter-efficient fine-tuning using scPEFT. *Nature Machine Intelligence*, 8(1):118–133, 2026.

Graham Heimberg, Tony Kuo, Daryle J. DePianto, Omar Salem, Tobias Heigl, Nathaniel Diamant, Gabriele Scalia, Tommaso Biancalani, Shannon J. Turley, Jason R. Rock, Héctor Corrada Bravo, Josh Kaminker, Jason A. Vander Heiden, and Aviv Regev. A cell atlas foundation model for scalable search of similar human cells. *Nature*, 638(8052):1085–1094, 2025.

Hongru Hu, Shuwen Zhang, Yongin Choi, Venkat S. Malladi, and Gerald Quon. sciLaMA: A Single-Cell Representation Learning Framework to Leverage Prior Knowledge from Large Language Models. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 24850–24868. PMLR, 2025.

Kasia Z. Kedzierska, Lorin Crawford, Ava P. Amini, and Alex X. Lu. Zero-shot evaluation reveals limitations of single-cell foundation models. *Genome Biology*, 26(1):101, 2025.

Svetlana Kiritchenko, Stan Matwin, Richard Nock, and A. Fazel Famili. Learning and Evaluation in the Presence of Class Hierarchies: Application to Text Categorization. In Luc Lamontagne and Mario Marchand (eds.), *Advances in Artificial Intelligence*, pp. 395–406, Berlin, Heidelberg, 2006. Springer. ISBN 978-3-540-34630-2.

Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods*, 16(12):1289–1296, December 2019. ISSN 1548-7105. doi: 10.1038/s41592-019-0619-0.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-Distribution Generalization via Risk Extrapolation (REx). In *Proceedings of the 38th International Conference on Machine Learning*, pp. 5815–5826. PMLR, 2021.

Xubin Li, Kartik Singhal, Qing Deng, Dai Chihara, David Russler-Germain, R. Andrew Harkins, Jared Henderson, Kotaro Arita, Atish Kizhakeyil, Ryan Sun, Priya Lakra, Usama Hussein, Jennifer A. Foltz, Ashley Wilson, Evelyn Schmidt, Imran Nizamuddin, Tommy Dinh, Akhil Kesaraju, Mark P. Hamilton, Carl Allen, Maher K. Gandhi, Joshua Tobin, Aixiang Jiang, Laura Hilton, David W. Scott, Francisco Vega, Christopher R. Flowers, Jason R. Westin, Obi L. Griffith, Todd A. Fehniger, Malachi Griffith, and Michael R. Green. Large B cell lymphoma microenvironment archetype profiles. *Cancer Cell*, 43(7):1347–1364.e13, 2025.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Tianyu Liu, Kexing Li, Yuge Wang, Hongyu Li, and Hongyu Zhao. Evaluating the Utilities of Foundation Models in Single-Cell Data Analysis. *Advanced Science*, 13(27):e14490, 2026.

Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, December 2018. ISSN 1548-7105. doi: 10.1038/s41592-018-0229-2.

Thomas M. Norman, Max A. Horlbeck, Joseph M. Replogle, Alex Y. Ge, Albert Xu, Marco Jost, Luke A. Gilbert, and Jonathan S. Weissman. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, 365(6455):786–793, 2019.

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

Jiaxin Qi, Hang Li, Yan Cui, Yuhua Zheng, and Jianqiang Huang. Towards Universal Gene Regulatory Network Inference: Unlocking Generalizable Regulatory Knowledge in Single-cell Foundation Models. In *Forty-Third International Conference on Machine Learning*, 2026.

François Remy, Kris Demuynck, and Thomas Demeester. BioLORD: Learning Ontological Representations from Definitions for Biomedical Concepts and their Textual Descriptions. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 1454–1465, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics.

François Remy, Kris Demuynck, and Thomas Demeester. BioLORD-2023: Semantic textual representations fusing large language models and clinical knowledge graph insights. *Journal of the American Medical Informatics Association*, 31(9):1844–1855, 2024.

Yanay Rosen, Yusuf Roohani, Ayush Agrawal, Leon Samotorčan, Stephen R. Quake, and Jure Leskovec. Universal cell embedding provides a foundation model for cell biology. *Nature*, pp. 1–9, 2026.

Carlos N. Silla and Alex A. Freitas. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, 22(1):31–72, 2011.

Dongqing Sun, Jin Wang, Ya Han, Xin Dong, Jun Ge, Rongbin Zheng, Xiaoying Shi, Binbin Wang, Ziyi Li, Pengfei Ren, Liangdong Sun, Yilv Yan, Peng Zhang, Fan Zhang, Taiwen Li, and Chenfei Wang. TISCH: A comprehensive web resource enabling interactive single-cell transcriptome visualization of tumor microenvironment. *Nucleic Acids Research*, 49(D1):D1420–D1430, January 2021. ISSN 0305-1048. doi: 10.1093/nar/gkaa1020.

Shawn Zheng Kai Tan, Aleix Puig-Barbe, Damien Goutte-Gattat, Caroline Eastwood, Brian Aevermann, Alida Avola, James P. Balhoff, Ismail Ugur Bayindir, Jasmine Belfiore, Anita Reane Caron, David S. Fischer, Nancy George, Benjamin M. Gyori, Melissa A. Haendel, Charles Tapley Hoyt, Huseyin Kir, Tiago Lubiana, Nicolas Matentzoglu, James A. Overton, Beverly Peng, Bjoern Peters, Ellen M. Quardokus, Patrick L. Ray, Paola Roncaglia, Andrea D. Rivera, Ray Stefancsik, Wei Kheng Teh, Sabrina Toro, Nicole Vasilevsky, Chuan Xu, Yun Zhang, Richard H. Scheuermann, Christopher J. Mungall, Alexander D. Diehl, and David Osumi-Sutherland. The Cell Ontology in the age of single-cell omics. *Scientific Data*, 13(1):946, 2026.

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(7965):616–624, 2023.

Li Wang, Cheng Wang, Juan A. Moriano, Songcang Chen, Guolong Zuo, Arantxa Cebrián-Silla, Shaobo Zhang, Tanzila Mukhtar, Shaohui Wang, Mengyi Song, Lilian Gomes de Oliveira, Qiuli Bi, Jonathan J. Augustin, Xinxin Ge, Mercedes F. Paredes, Eric J. Huang, Arturo Alvarez-Buylla, Xin Duan, Jingjing Li, and Arnold R. Kriegstein. Molecular and cellular dynamics of the developing human neocortex. *Nature*, 647(8088):169–178, November 2025. ISSN 1476-4687. doi: 10.1038/s41586-024-08351-7.

Sachini Weerasekara, Natasha Darras, Sagar Kamarthi, Colles Price, and Jacqueline Isaacs. Prototype Guided Post-pretraining for Single-Cell Representation Learning. *arXiv preprint arXiv.2605.07938*, 2026.

Aaron Wenteler, Martina Occhetta, Nikhil Branson, Victor Curean, Magdalena Huebner, William Dee, William Connell, Siu Pui Chung, Alex Hawkins-Hooker, Yasha Ektefaie, César Miguel Valdez Córdova, and Amaya Gallagher-Syed. PertEval-scFM: Benchmarking Single-Cell Foundation Models for Perturbation Effect Prediction. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 66633–66677. PMLR, 2025.

Jialu Wu, Qing Ye, Yilin Wang, Renling Hu, Yiheng Zhu, Mingze Yin, Tianyue Wang, Jike Wang, Chang-Yu Hsieh, and Tingjun Hou. Biology-driven insights into the power of single-cell foundation models. *Genome Biology*, 26(1):334, 2025.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A Theory of Usable Information under Computational Constraints. In *International Conference on Learning Representations*, 2020.

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(10):852–866, 2022.

Xinyu Yuan, Zhihao Zhan, Zuobai Zhang, Manqi Zhou, Jianan Zhao, Boyu Han, Yue Li, and Jian Tang. Cell ontology guided transcriptome foundation model. In *Advances in Neural Information Processing Systems*, volume 37, pp. 6323–6366. Curran Associates, Inc., 2024.

Zetian Zheng, Junyi Chen, Xingjian Chen, Lei Huang, Weidun Xie, Qiuzhen Lin, Xiangtao Li, and Ka-Chun Wong. Enabling Single-Cell Drug Response Annotations from Bulk RNA-Seq Using SCAD. *Advanced Science*, 10(11):2204113, 2023. ISSN 2198-3844. doi: 10.1002/advs.202204113.

Zhixing Zhong, Junchen Hou, Zhixian Yao, Lei Dong, Feng Liu, Junqiu Yue, Tiantian Wu, Junhua Zheng, Gaoliang Ouyang, Chaoyong Yang, and Jia Song. Domain generalization enables general cancer cell annotation in single-cell and spatial transcriptomics. *Nature Communications*, 15(1):1929, 2024.

APPENDIX

CONTENTS

A Semantic Supervision Details **16**

B Post-training Data and Semantic Resources **17**

C Models, Baselines, and Implementation **19**

D Evaluation Protocols and Metrics **21**

D.1 Cell-type Annotation Transfer 21

D.2 Cancer-Finder 22

D.3 Drug Sensitivity 22

D.4 Gene-level Evaluation 22

D.5 Perturbation Prediction 23

D.6 Representation Analyses 23

D.7 Direct Cell-type Inference 24

D.8 Ablation and Data Scaling 24

E Extended Results and Analyses **25**

E.1 Extended Frozen Downstream Transfer 25

E.1.1 Cell-type Annotation Transfer 25

E.1.2 Cancer-Finder 25

E.1.3 Drug Sensitivity 26

E.1.4 Gene-level Evaluation 26

E.1.5 Perturbation Prediction 28

E.2 Representation Analyses 28

E.3 Direct Cell-type Inference Audits 29

E.4 Ablations, Scaling, and Sensitivity 30

F Extended Case Studies and Diagnostics **31**

F.1 Neocortex Regional-composition Case 31

F.2 Low-accuracy OOD Annotation Case 33

A Semantic Supervision Details

A.1 Ontology-aware cell supervision

Cell identity uses the classification head and asymmetric ontology soft-label loss defined in the main text. The OWL parser retains Cell Ontology classes and `rdfs:subClassOf` links whose endpoints are Cell Ontology terms, with edges directed from child to parent. Alternative CL identifiers, including CL-valued database cross-references, are mapped to canonical CL terms; redundant transitive edges are removed, and acyclicity is checked. Intermediate ontology nodes remain in the graph even though the prediction head is restricted to the observed output classes.

For each output class, shortest upward distances to all reachable ancestors are computed by breadth-first traversal, including the class itself at distance zero. The implementation then evaluates all shared ancestors, as required for multiple inheritance. Missing output terms or class pairs without a shared ancestor raise an error rather than receiving an imputed cost. The offline cost and row-normalized soft-target matrices use the classifier class order: rows index true classes and columns index candidate classes.

Figure A1 summarizes the three canonical behaviors induced by its asymmetric path cost. The construction follows the broader hierarchy-aware principle that semantically related mistakes should receive different costs, while adapting it to the Cell Ontology DAG and the asymmetric specificity preference used by SPORE (Bertinetto et al., 2020; Cultrera di Montesano et al., 2026; Tan et al., 2026).

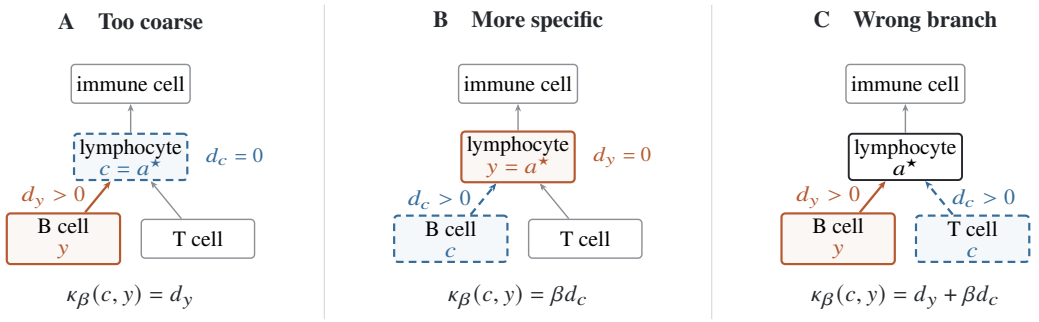

Figure A1: Three cases induced by the asymmetric ontology cost. Orange solid paths mark the reference-side distance $d_y = d_\uparrow(y, a^\star)$; blue dashed paths mark the candidate-side distance $d_c = d_\uparrow(c, a^\star)$. A coarse ancestor prediction pays d_y, a more-specific descendant prediction pays the discounted cost βd_c, and a sibling error pays both. For clarity, the schematic uses tree-shaped examples; the implemented loss minimizes over all shared ancestors in the Cell Ontology DAG.

Implementation checks verify exact, ancestor, descendant, sibling, and multiple-inheritance cases, together with target normalization.

A.2 Biological-context alignment

Both context objectives use the normalized cell projection and frozen BioLORD targets defined in the main text (Remy et al., 2022; 2024). Cosine regression aligns each cell only with its positive target. The primary fixed-bank objective instead scores that target against one entry for every unique canonical context represented in the post-training training split. Repeated cells therefore do not duplicate candidates, and the candidate set is independent of microbatch size and gradient accumulation. The bank is sorted by stable context identifier and materialized as an L2-normalized matrix. Appendix B.2 specifies the fine-tissue, broad-tissue, and disease-or-state strings from which these targets are constructed.

A.3 GENE-SEMANTIC ALIGNMENT

Gene semantics use one fixed, L2-normalized GenePT registry and a trainable projector; sciLaMA provides a related precedent for introducing language-model gene knowledge into single-cell representations (Chen & Zou, 2025; Hu et al., 2025). The registry is mapped to model-vocabulary rows before training. Uncovered rows retain a false coverage mask and never contribute to the loss. At each optimization step, the implementation samples covered model-level gene rows uniformly without replacement, independently of the genes expressed by cells in the current mini-batch. Thus the supervision unit is a vocabulary row, not a cell or contextualized token occurrence. The reported configurations with active gene-semantic supervision sample up to 8,192 covered rows per optimization step. GenePT targets remain fixed; the projector is trainable, and the underlying model-level gene rows are updated only when the backbone stage exposes them as trainable parameters. Model-specific vocabulary mappings and trainable interfaces are reported in Appendix C.2.

B POST-TRAINING DATA AND SEMANTIC RESOURCES

B.1 CELLXGENE POST-TRAINING CORPUS

The shared asset is derived from the Homo sapiens measurement in CELLxGENE Census LTS `2025-11-08` (CZI Cell Science Program et al., 2025). Candidate observations are primary data and belong to one of seven 10x assay categories (3′ v1–v3, 5′ v1–v2, or the 3′/5′ transcription-profiling categories). Rows with unknown donors or missing/unknown `cell_type_ontology_term_id` are removed. Eligible labels have at least 5,000 cells and 30 donors in the filtered Census. Within each eligible label, cells are ordered by a deterministic BLAKE2b priority of seed 13 and `soma_joinid`, and at most 5,000 cells are retained. The capped candidates are then ordered globally by the same priority, and the first 100,000 cells are selected. This produces 375 observed Cell Ontology labels and 20,469 gene features.

Complete donors are assigned to train, validation, or test with seed 13 and nominal 70/15/15 proportions. The realized split contains 69,809/15,001/15,190 cells and 6,061/1,333/1,153 donors, respectively, with no donor overlap. All 375 output classes are represented in the training partition. Only the 69,809 training cells participate in post-training; the full 100,000-cell asset supplies the shared donor-held-out cell-type protocol. The manifest's `soma_joinid`, ontology label, and split are preserved across backbones. Model-native tokenization may remove unavailable genes but may not resample cells or alter split membership.

B.2 CONTEXT-TARGET CONSTRUCTION

The context manifest maps CELLxGENE `tissue` to fine tissue, `tissue_general` to broad or organ-level anatomical context, and `disease` to disease or physiological state. Values are stripped of surrounding whitespace. The case-insensitive placeholders empty string, `na`, `n/a`, `nan`, `none`, `not applicable`, `not available`, `unknown`, and `unspecified` are masked. Disease strings are split only on the literal delimiter `||` ; terms are stripped, deduplicated, and sorted case-insensitively. The explicit value `normal` is retained as a biological state. No ontology-name substitution or synonym expansion is applied.

For the experiments, the deduplicated joint records are serialized exactly as

```
Cellular context: {tissue}, within the broader anatomical
context of {tissue_general}; disease or physiological state:
{terms}.
```

Multiple disease terms are joined by a semicolon followed by one space; an absent disease list is rendered as `not specified`. The sentence builder requires nonempty fine and broad tissue values; all 100,000 cells in the final asset satisfy this requirement. For example:

```
Cellular context: ascending aorta, within the broader
anatomical context of vasculature; disease or physiological
state: normal.
```

Each normalized payload receives a deterministic stable identifier, yielding 712 unique joint contexts. Cells join targets explicitly through `soma_joinid` and this context identifier; JSON or H5AD row order is never used as an implicit join. Dataset, donor, cell identifier, and cell-type label remain available for audits but never enter the semantic text.

B.3 Context-target input-data audit

All three context fields are populated in the final 100,000-cell asset. Their distributions are nevertheless uneven: fine tissue is long-tailed, broad tissue is concentrated, and disease/state is dominated by the explicit "normal" state. Table A1 records the compact audit that governed target construction.

Table A1: Metadata audit for the shared 100k donor-split CELLxGENE asset. "Top-5" is the fraction of cells covered by the five most frequent values.

Field	Unique	Observed cells	Top value	Top-5	Multi-label cells
Fine tissue	362	100,000	blood (31.41%)	50.01%	0
Broad tissue	65	100,000	blood (31.43%)	76.86%	0
Disease/state	225	100,000	normal (61.83%)	79.11%	3,621 (3.62%)

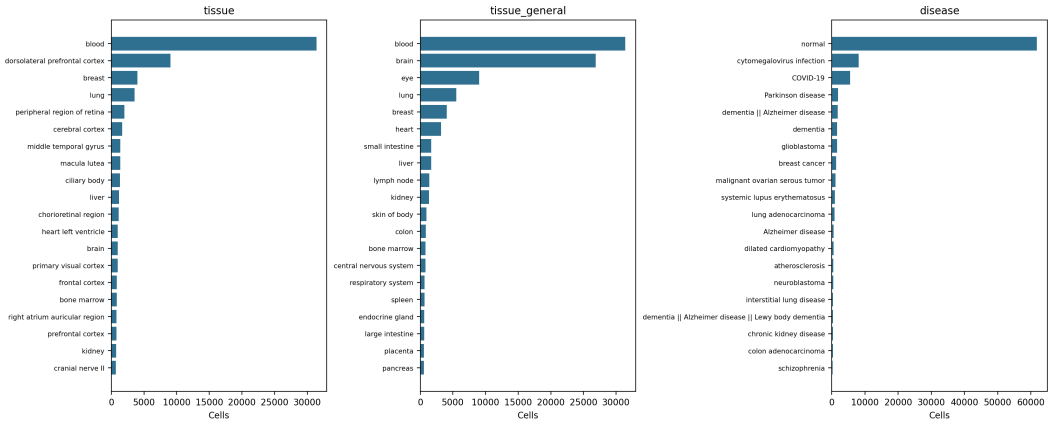

Figure A2: Marginal distributions of the three context fields in the shared 100k donor-split CELLxGENE asset. The figure complements Table A1 by showing the long tail beyond its top-value summaries: fine tissue has many low-frequency values, broad tissue is more concentrated, and `normal` is an explicit majority disease/state value rather than a missing-value code.

The target-construction protocol retains `normal` as an explicit state and preserves all normalized terms for multi-label disease/state entries. Dataset and donor identifiers are excluded from semantic targets. Every observed fine tissue maps to exactly one broad tissue value in this asset. The machine-readable audit retains the complete value counts, split presence, fine-to-broad pairs, and multi-label terms; none is treated as model performance.

B.4 External semantic resources and provenance

Cell Ontology. We use the official versioned `cl-basic.owl` release of `2026-06-08` for AOSL supervision (Tan et al., 2026). Ontologies used only for evaluation are documented with the corresponding protocol in Appendix D.6, not as supervision resources here (Wu et al., 2025).

BioLORD context targets. The 712 canonical strings defined in Appendix B.2 are encoded offline with `FremyCompany/BioLORD-2023`, revision `167aab527b238a50ca65224e6319215d2ff4fc9f`, producing row-normalized float32 vectors of dimension 768. BioLORD-2023 is the biomedical encoder used for these context targets (Remy et al., 2022; 2024).

GenePT targets. The fixed registry is derived from NCBI gene summaries embedded by `text-embedding-ada-002` (Chen & Zou, 2025; Hu et al., 2025). After trimming and uppercasing symbols, removing duplicates, rejecting invalid or zero-norm vectors, and L2 normalizing, the registry contains 93,428 unique symbols with 1,536-dimensional float32 targets. Model-specific mappings belong to Appendix C.2. Likewise, the fixed ESM-2 gene/protein inputs used by UCE and TF-Sapiens are model resources documented there, not GenePT supervision targets (Lin et al., 2023).

C Models, Baselines, and Implementation

C.1 Baselines

We evaluate SPORE against representative single-cell foundation models and non-foundation baselines. Table A2 summarizes the architectural characteristics and roles of all evaluated models.

Single-cell foundation models. **Geneformer** is a rank-based scFM pretrained with masked gene modeling, serving as a representative self-supervised RNA foundation model (Theodoris et al., 2023). **scGPT** represents expression-token models that incorporate expression values through value binning and supports multiple single-cell modalities (Cui et al., 2024). **scFoundation** uses continuous value projection and a large-scale RNA pretraining corpus, providing another representative expression-based foundation model (Hao et al., 2024).

scCello and **TEDDY-G** represent supervised or knowledge-guided scFMs, where cell-type labels, ontology information, or metadata supervision are incorporated during model construction (Yuan et al., 2024; Chevalier et al., 2025). They provide comparisons with models that already exploit biological annotations.

UCE and **TF-Sapiens** are included as gene-prior-aware models that leverage protein-language-model-derived information for gene representation (Rosen et al., 2026; Pearce et al., 2026). They provide complementary pretrained interfaces in which protein- or evolution-derived gene priors are incorporated during model construction.

Supervised and conventional baselines. **CellTypist** provides a regularized linear cell-annotation reference (Domínguez Conde et al., 2022). **scTab** provides atlas-scale supervised classifiers, while **HCE-scTab** further incorporates Cell Ontology structure through hierarchical cross-entropy and provides an ontology-aware supervised annotation reference (Fischer et al., 2024; Cultrera di Montesano et al., 2026).

We further include conventional representation-learning baselines (HVG/PCA, scVI, Harmony) and expression-based or task-specific predictors as non-foundation references where their evaluation interfaces are well defined (Lopez et al., 2018; Korsunsky et al., 2019).

C.2 Backbone-specific SPORE integration

Backbone selection and scope. The post-training backbone sweep was designed to cover distinct pretraining designs. Geneformer represents generic self-supervised pretraining; scCello and TEDDY-G represent models incorporating explicit biological annotations or ontology structure during pretraining; and UCE and TF-Sapiens represent protein-informed architectures using ESM-derived gene representations. scGPT and scFoundation are included as additional frozen scFM references but are not part of the paired post-training sweep.

Geneformer, scCello, and TEDDY-G. All three backbones use the full set of semantic supervision branches and expose trainable model-level gene-embedding tables for GenePT alignment. Covered or sampled embedding rows are projected to the GenePT space through a trainable projector and aligned to fixed GenePT targets. These gene embeddings are updated together with the declared backbone parameters and semantic heads during Stage 2. For scCello and TEDDY-G, the gene branch operates directly on the input gene-embedding table rather than the native prediction or decoder head. Geneformer retains its native rank-tokenization and representation interface; all

Table A2: Summary of evaluated single-cell models and non-foundation baselines.

Model	Params[†]	Pretraining scale	Input genes	Output dim	Gene interface / prior	Supervised
Expression-pretrained foundation models						
Geneformer-6L	10M	30M cells	2,048 ranked genes	256	Lookup	×
scGPT	53M	33M cells	1,200 HVGs	512	Lookup	×
scFoundation	100M	50M cells	19,264 genes	3,072	Value projection	×
Supervised or ontology-aware foundation models						
scCello	10M	22M cells	2,048 ranked genes	200	Lookup	✓
TEDDY-G 70M	71M	116M cells	2,048 ranked genes	512	Lookup	✓
Gene/protein-prior foundation models						
UCE	650M	36M cells	1,024 sampled genes	1,280	ESM-2 15B	×
TF-Sapiens	368M	57M cells	2,047 genes with counts	2,048	ESM-2 3B	×
Supervised classifier baseline						
scTab	10M	22M cells	19,331 protein-coding genes	–	–	✓
Conventional baselines						
HVG/PCA	–	–	HVGs	–	–	×
scVI	–	–	All genes	–	–	×
Harmony	–	–	PCA	–	–	×
Expression MLP	–	–	All genes	–	–	✓

[†] Parameter counts denote trainable parameters of the evaluated model and exclude frozen external encoders where applicable.

Geneformer results reported in this paper use the V1-10M 6-layer backbone with a 256-dimensional hidden representation. TEDDY-G refers to the TEDDY-G 70M variant throughout our experiments.

UCE and TF-Sapiens. Both backbones retain their native ESM-derived gene interfaces during SPORE post-training. Because these interfaces are fixed rather than trainable model-level gene vocabularies, the GenePT branch is not activated ($\lambda_{\text{gene}} = 0$); SPORE therefore applies ontology-aware cell supervision and biological-context supervision only. Their native gene interfaces remain unchanged for gene-level evaluation.

Each backbone otherwise retains its model-specific preprocessing and representation interface.

C.3 STAGE-WISE OPTIMIZATION DETAILS

Stage 1 trains the semantic interfaces with the backbone frozen. Stage 2 is initialized from the selected Stage 1 checkpoint and jointly optimizes the declared backbone parameters and semantic interfaces with separate learning rates.

Stage 2 adaptation is backbone-specific. Geneformer, scCello, and TEDDY-G use full-backbone adaptation. UCE updates only its top eight transformer layers while keeping the native gene-embedding table frozen, whereas TF-Sapiens updates the full backbone at the backbone learning rate while keeping its native gene-embedding table frozen.

For configurations with gene-semantic supervision, the first 5% of Stage 2 serves as a projector warm-up period, during which the backbone remains frozen. The GenePT loss weight is linearly ramped from zero to its maximum value over the first 10% of optimization steps. All reported configurations use AdamW with constant learning rates and no learning-rate scheduler.

C.4 CHECKPOINT SELECTION AND REUSE

Checkpoint selection uses only the post-training validation split, with hMacro-F1 as the selection metric; downstream test performance is not used for checkpoint selection. One selected checkpoint

is retained for each post-trained backbone and reused across all applicable downstream tasks without task-specific backbone selection or adaptation.

C.5 HYPERPARAMETERS

Table A3: Core post-training hyperparameters for the primary configurations reported in the paper. Microbatch is the number of cells per forward pass; effective batch is microbatch multiplied by gradient accumulation. N/A denotes an inactive objective or an inapplicable parameter.

Backbone	Cell (AOSL)	Context	Gene	Micro batch	Accum.	Eff. batch	Epochs	Head LR	Backbone LR	WD	λ_{ctx}	λ_{gene}
Geneformer	✓	✓	✓	32	1	32	3	10^{-3}	2×10^{-5}	0.01	0.1	0.05
scCello	✓	✓	✓	2	1	2	3	10^{-4}	10^{-6}	0.01	0.1	0.05
TEDDY-G 70M	✓	✓	✓	2	1	2	3	10^{-5}	10^{-5}	0.01	0.1	0.05
UCE	✓	✓		1	16	16	3	10^{-4}	10^{-6}	10^{-4}	0.1	N/A
TF-Sapiens	✓	✓		2	8	16	3	10^{-4}	10^{-6}	10^{-4}	0.1	N/A

For all configurations, $\lambda_{\mathrm{cell}} = 1$. For gene-semantic configurations, the value of λ_{gene} shown in the table is its maximum value after the linear ramp described above.

Unless otherwise specified, the AOSL soft-target construction uses $\beta_{\mathrm{AOSL}} = 0.5$ and $\tau_{\mathrm{AOSL}} = 0.5$. UCE and TF-Sapiens use $\beta_{\mathrm{AOSL}} = 0.5$ and $\tau_{\mathrm{AOSL}} = 1.0$.

Across all backbones, the context projector is a two-layer MLP with hidden width 512 and output dimension 768, followed by L2 normalization to match the BioLORD-2023 target space. Geneformer, scCello, TEDDY-G, and TF-Sapiens use the contrastive context objective with temperature 0.07 and $\lambda_{\mathrm{ctx}} = 0.1$. UCE instead uses cosine context regression with $\lambda_{\mathrm{ctx}} = 0.1$, for which the temperature parameter is not used.

The three full three-branch configurations—Geneformer, scCello, and TEDDY-G—use a GenePT row sample size of $M = 8192$ and a $10\times$ GenePT projector learning-rate multiplier.

D EVALUATION PROTOCOLS AND METRICS

Unless otherwise noted, paired frozen and post-trained models use the same downstream data, splits, readouts, and evaluation metrics. Model-specific preprocessing or gene mapping follows each backbone's native input interface without changing the evaluation target. Where the executed protocol differs from a referenced benchmark, we describe the executed setting below.

D.1 CELL-TYPE ANNOTATION TRANSFER

Data and split. We evaluate cell-type transfer on the fixed 100,000-cell human subset constructed from CELLxGENE Census LTS `2025-11-08` (CZI Cell Science Program et al., 2025); Appendix B.1 describes its construction. The task uses a 375-class `cell_type_ontology_term_id` label universe and one shared donor-disjoint train/validation/test manifest. Each model applies its native input preprocessing to the same cells, labels, and split assignments.

Readout. Cell-type evaluation uses the shared donor-held-out split and the same 375-class label universe for all models. Backbones are frozen and the same MLP readout, with hidden dimensions 512 and 375, ReLU activation, and dropout 0.1, is fitted on the training partition, selected on validation data, and evaluated on held-out donors. The readout is trained with AdamW for at most 20 epochs, with validation-based checkpoint selection. Post-training optimization and checkpoint selection are described in Appendices C.3 and C.4.

Metrics. We report hierarchy-compatible accuracy (hAcc) and class-unweighted macro-F1 (hMacro-F1). Let $c_i = 1$ when the prediction $\hat{y}_i$ matches the reference y_i or is an `is_a` descen-

dant of y_i in Cell Ontology, and $c_i = 0$ otherwise. Then

$$\mathrm{hAcc} = \frac{1}{n} \sum_{i=1}^{n} c_i. \tag{11}$$

hMacro-F1 is computed over the 375 classes after compatible descendant predictions are assigned to their reference labels. The additional ontology metrics for direct semantic inference are defined in Appendix D.7.

D.2 CANCER-FINDER

Data, split, and metric. We follow the Cancer-Finder transfer setting on TISCH (Sun et al., 2021; Wu et al., 2025). The downstream task predicts malignant versus non-malignant cell identity and evaluates four held-out tissue domains: Blood, Bone, Brain, and Eye. For each domain, the classifier is fitted using the registered training domains and evaluated on the held-out domain. Performance is reported as the unweighted mean AUROC across the four domains, with per-domain results reported in Table A6.

Readout and model selection. Each backbone provides its frozen cell representation after model-native preprocessing. We use the SequencingCancerFinder predictor trained with the VREx objective (Zhong et al., 2024; Krueger et al., 2021), applied consistently to the frozen and post-trained representations. Downstream training and model selection follow the fixed Cancer-Finder/scFM-Bench evaluation configuration used in our implementation (Wu et al., 2025).

D.3 DRUG SENSITIVITY

Data, split, and metric. We follow the SCAD drug-response transfer setting (Zheng et al., 2023) for Sorafenib, NVP-TAE684, PLX4720_451Lu, and Etoposide. For each compound, all methods use the same registered stratified five-fold split over cell lines. Performance is measured by held-out ROC-AUC for each drug, and the primary score is the unweighted mean across the four drugs. Per-drug results are reported in Table A7.

Readout and model selection. Frozen cell representations are supplied to the SCAD downstream pipeline, with the backbone fixed throughout downstream fitting. We use the released SCAD feature extractor, predictor, and adversarial discriminator (Zheng et al., 2023; Wu et al., 2025). To preserve comparability with the released benchmark, downstream model selection follows the SCAD/scFM-Bench evaluation protocol.

D.4 GENE-LEVEL EVALUATION

These evaluations operate on frozen model-level gene representations rather than cell embeddings. Tissue specificity evaluates single-gene tissue prediction, whereas UGRN evaluates pairwise regulatory-link transfer. For UCE and TF-Sapiens, both tasks use the native fixed ESM-derived gene vectors; these interfaces are unchanged between the frozen and SPORE PT checkpoints and are not affected by GenePT alignment.

D.4.1 TISSUE SPECIFICITY TRANSFER

Data, split, and metric. Gene–Tissue uses the FRoGS tissue-specific gene labels (Chen et al., 2024; Wu et al., 2025), excluding genes assigned to multiple tissue labels. Gene identifiers are mapped to each model's validated gene-vector vocabulary, and genes without a corresponding vector are excluded. We use five fixed 80/20 train/test gene splits with seeds 0–4. Performance is reported as accuracy averaged over the five splits, with macro-F1 as a secondary metric. The evaluated vocabulary contains 291–298 genes depending on the model; coverage is reported with the results.

Readout. For each split, a random forest with 100 trees and maximum depth 5 is fitted to the frozen training-gene vectors and evaluated on the held-out genes. The same readout protocol is used for all representations.

D.4.2 Gene Regulatory Network Inference

Data, split, and metric. UGRN evaluates TF–target links across the registered STRING, NonSpecific, LOFGOF, and Specific network settings (Qi et al., 2026). The Baseline-Emb protocol trains on STRING:hESC and evaluates the registered non-hESC network–data endpoints. Gene identifiers are mapped to each model's validated gene-vector vocabulary before constructing edge features. Performance is measured by AUPRC for each held-out endpoint, and the primary score is the unweighted mean across endpoints. Endpoint-level results are reported in Table A9.

Readout. Baseline-Emb represents each candidate edge using the sum of the frozen TF and target gene vectors and fits the same transfer MLP, $d \to 128 \to 64 \to 1$, with ReLU hidden activations and a sigmoid output, on the STRING:hESC training edges for every backbone. The classifier and negative-sampling protocol are held fixed across representations.

Label-availability diagnostic. For the Geneformer diagnostic, we additionally vary the fraction of STRING:hESC training edges over 1, 5, 10, 25, and 100% while keeping the representations, held-out endpoints, and edge-classifier protocol fixed. All points are evaluated on the same coverage-matched set of 18 non-hESC endpoints (150,207 evaluation edges).

D.5 Perturbation Prediction

Data and metric. We evaluate the Norman_1 panel of PertEval-scFM (Wenteler et al., 2025; Norman et al., 2019) using the shared precomputed SPECTRA train/test artifacts. The SPECTRA sparsification levels are $s \in \{0.0, 0.1, \ldots, 0.7\}$, with larger s corresponding to increasingly disconnected train/test perturbation relationships. Prediction error is measured by test MSE at each split. We summarize performance by trapezoidal integration of MSE over $s = 0.0$–0.7 (AUSPC); lower values indicate lower error across the shift range. ΔAUSPC is the difference between the parameter-free mean baseline and the method AUSPC, so positive values indicate improvement. Table A10 displays the $s = 0.1$–0.7 split results; the $s = 0$ endpoint is included in the AUSPC calculation.

Readout. Frozen and post-trained representations are evaluated with the PertEval-scFM Norman MLP readout using the same SPECTRA sparsification splits and a fixed downstream configuration across all paired comparisons (Wenteler et al., 2025). For each backbone, the post-trained row uses the selected global checkpoint reported in Table 1, with the backbone fixed during downstream fitting.

D.6 Representation Analyses

Cancer label efficiency. The Geneformer backbone remains frozen while a linear Cancer-Finder readout is trained using 1, 5, 10, 25, or 100% of the available downstream training labels. Each fraction is evaluated on the same four held-out domains, and AUROC is macro-averaged across domains. The 100% point uses all available training labels under this linear-probe protocol and is therefore distinct from the task-specific Cancer-Finder result in Table 1.

Ontology geometry. Within each of 39 tissue strata, we correlate mean cross-dataset cosine similarity for each cell-type pair with the negative Cell Ontology shortest-path distance using Spearman ρ, and macro-average the resulting correlations across tissues. Frozen and post-trained representations use the same extraction interface.

Biology-versus-assay accessibility. Cells are exactly matched by Cell Ontology label, tissue, and disease/state before probe fitting. Dataset-disjoint five-fold probes separately predict the biological target and the assay/platform target. Performance is reported as chance-adjusted balanced accuracy, $(\mathrm{BA} - 1/K)/(1 - 1/K)$ for an endpoint with K classes, averaged over the five held-out-dataset folds. Biology and assay are treated as separate endpoints, with frozen–post-training differences computed within each endpoint. This analysis is a matched shortcut-control diagnostic and does not establish causal removal of technical variation.

D.7 Direct Cell-type Inference

ID evaluation. The donor-held-out ID evaluation uses a fixed 164-class candidate bank and 10,494 matched cells from the shared CELLxGENE split.

OOD cohort construction. The OOD cohort is constructed from human primary-cell studies published after 2023, providing a temporal screen that places the cohort after the original pretraining or release window of several earlier backbones. Candidate studies are additionally excluded if their dataset ID, collection ID, or normalized DOI overlaps the frozen post-training train/validation inventory. Cells are retained only when their reference label maps to the fixed 164-class candidate bank. Each retained study contains at least five comparable labels and three donors. We then apply deterministic sampling caps of 100 cells per study–donor–label group, 150 cells per study–label group, and 1,000 cells per study. The resulting manifest contains 10,500 cells from 11 studies, 14 datasets, 302 donors, and 56 observed reference labels.

Inference protocol and metrics. The cell-identity head from the selected global checkpoint is applied directly, without fitting a target-study classifier or updating the backbone. CellTypist, scTab, and HCE-scTab are included as supervised annotation baselines (Domínguez Conde et al., 2022; Fischer et al., 2024; Cultrera di Montesano et al., 2026). The main comparison reports hAcc, hMacro-F1, hierarchical F1 (hF), and LCAE. hAcc treats an exact prediction or a valid `is_a` descendant of the reference as correct; hMacro-F1 uses the same hierarchy-compatible relation before class-unweighted averaging. ID hMacro-F1 uses the fixed 164-class universe, whereas OOD hMacro-F1 is averaged over the 56 reference classes observed in the OOD cohort.

For hF, let $A(y)$ denote the full Cell Ontology `is_a` ancestor closure of class y, including y itself. Following Kiritchenko et al. (2006); Silla & Freitas (2011), we compute

$$hP = \frac{\sum_i |A(\hat{y}_i) \cap A(y_i)|}{\sum_i |A(\hat{y}_i)|}, \qquad hR = \frac{\sum_i |A(\hat{y}_i) \cap A(y_i)|}{\sum_i |A(y_i)|}, \qquad hF = \frac{2hPhR}{hP + hR}.$$

For the strict-error set $\mathcal{E} = \{i : \hat{y}_i \neq y_i\}$, let $d_{\downarrow}(a, y)$ denote the minimum directed hop count from ancestor a to descendant y. We define the symmetric LCA path distance as

$$d_i^{\mathrm{LCA}} = \min_{a \in A(y_i) \cap A(\hat{y}_i)} \left[d_{\downarrow}(a, y_i) + d_{\downarrow}(a, \hat{y}_i) \right], \qquad \mathrm{LCAE} = \frac{1}{|\mathcal{E}|} \sum_{i \in \mathcal{E}} d_i^{\mathrm{LCA}}. \tag{12}$$

The minimum over shared ancestors accommodates multiple inheritance in the ontology DAG. Setting $d_i^{\mathrm{LCA}} = 0$ for exact matches gives

$$\mathrm{LCAA} = \frac{1}{n} \sum_{i=1}^{n} d_i^{\mathrm{LCA}} = \frac{|\mathcal{E}|}{n} \mathrm{LCAE}. \tag{13}$$

The extended metric audit additionally reports reference specificity loss (RSL),

$$d_i^{\mathrm{RSL}} = \min_{a \in A(y_i) \cap A(\hat{y}_i)} d_{\downarrow}(a, y_i), \qquad \mathrm{RSL} = \frac{1}{|\mathcal{E}|} \sum_{i \in \mathcal{E}} d_i^{\mathrm{RSL}}. \tag{14}$$

RSL is asymmetric: a strict-error prediction that is a proper descendant of the reference has zero reference-side distance, whereas LCAE also counts the prediction-side displacement. The OOD cohort is not used for model fitting, threshold selection, or checkpoint selection.

D.8 Ablation and Data Scaling

Component ablations. Using Geneformer as the anchor backbone, we compare Frozen, Cell (flat), Cell (ontology), Context, Gene semantic, and the joint SPORE configuration. Cell (flat) uses ordinary cell-type supervision, Cell (ontology) uses ontology-aware cell supervision, Context isolates biological-context supervision, and Gene semantic trains only the gene-semantic projector. All configurations are evaluated using the same downstream task protocols as the primary frozen-transfer experiments.

Fixed-budget data scaling. We construct nested, cell-type-stratified subsets containing 1, 3, 10, 30, and 100% of the 69,809-cell post-training pool while keeping the validation and test partitions fixed. All 375 training labels are represented at every scale. Each subset is trained under the same 2,000-step optimizer budget, and the reported checkpoint is selected on the fixed validation split. The 0% row denotes the frozen pretrained model and lies outside this post-training grid. The 100% scaling run is therefore distinct from the three-epoch checkpoint used in Tables 1 and 3.

E Extended Results and Analyses

E.1 Extended Frozen Downstream Transfer

Non-foundation decomposition. Table A4 resolves the task-specific constituents of the composite non-foundation reference used in the main downstream-transfer comparison.

Table A4: **Conventional and task-specific baseline references.** The upper block decomposes the simple non-FM baseline in Table 1; the lower block reports additional task-specific references where applicable.

Representation	Task role	Cell type	Cancer	Drug sens.	Tissue spec.	Perturb.
Simple non-FM baseline components						
Expression / Raw-X	cell/cancer/drug	0.778	0.735	0.631	N/A	N/A
One-hot	gene	N/A	N/A	N/A	0.353	N/A
PertEval non-FM	perturb.	N/A	N/A	N/A	N/A	0.0923
Additional task-specific references						
scVI normalized	Cancer	N/A	0.765	N/A	N/A	N/A
FRoGS-ArchS4	gene	N/A	N/A	N/A	0.859	N/A
FRoGS-GO	gene	N/A	N/A	N/A	0.458	N/A

E.1.1 Cell-type Annotation Transfer

Table A5: Cell-type transfer performance under the Cell Ontology child-compatible evaluation. hAcc counts a prediction as correct when it matches the reference label or a legal *is_a* descendant; hMacro-F1 is the macro average over the evaluated cell-type labels.

Method	Setting	hAcc	hMacro-F1
Non-FM baseline	task-specific	0.778	0.507
scGPT	frozen	0.772	0.411
scFoundation	frozen	0.654	0.298
Geneformer	frozen	0.649	0.271
	SPORE PT	0.822	0.497
scCello	frozen	0.747	0.386
	SPORE PT	0.833	0.432
TEDDY-G 70M	frozen	0.665	0.365
	SPORE PT	0.824	0.520
UCE	frozen	0.839	0.578
	SPORE PT	0.874	0.657
TF-Sapiens	frozen	0.871	0.681
	SPORE PT	0.894	0.657

E.1.2 Cancer-Finder

Table A6: Cancer-Finder AUROC across the four held-out tissue domains.

Method	Setting	Blood	Bone	Brain	Eye	Mean
Non-FM baseline	task-specific	0.596	0.542	0.848	0.952	0.735
scGPT	frozen	0.474	0.615	0.506	0.624	0.555
scFoundation	frozen	0.664	0.675	0.906	0.979	0.806
Geneformer	frozen	0.580	0.570	0.658	0.927	0.684
	SPORE PT	0.617	0.639	0.855	0.861	**0.743**
scCello	frozen	0.694	0.611	0.875	0.912	0.773
	SPORE PT	0.767	0.573	0.871	0.938	**0.787**
TEDDY-G 70M	frozen	0.670	0.602	0.741	0.925	0.734
	SPORE PT	0.654	0.608	0.805	0.925	**0.748**
UCE	frozen	0.756	0.564	0.921	0.948	0.797
	SPORE PT	0.820	0.649	0.768	0.961	**0.799**
TF-Sapiens	frozen	0.501	0.572	0.879	0.951	0.726
	SPORE PT	0.678	0.601	0.901	0.929	**0.777**

E.1.3 DRUG SENSITIVITY

Table A7: Drug Sensitivity AUC for four compounds. Macro average is reported across compounds.

Method	Setting	Sorafenib	NVP-TAE684	PLX4720_451Lu	Etoposide	Macro avg.
Non-FM baseline	task-specific	0.597	0.807	0.455	0.665	0.631
scGPT	frozen	0.975	0.900	0.693	0.603	0.793
scFoundation	frozen	0.858	0.589	0.790	0.665	0.726
Geneformer	frozen	0.593	0.575	0.507	0.692	0.592
	SPORE PT	0.819	0.756	0.535	0.659	**0.692**
scCello	frozen	0.676	0.728	0.582	0.528	0.629
	SPORE PT	0.829	0.742	0.493	0.540	**0.651**
TEDDY-G 70M	frozen	0.743	0.889	0.544	0.423	0.650
	SPORE PT	0.860	0.918	0.616	0.607	**0.750**
UCE	frozen	0.461	0.700	0.560	0.467	0.547
	SPORE PT	0.488	0.565	0.507	0.532	0.523
TF-Sapiens	frozen	0.567	0.435	0.499	0.641	0.535
	SPORE PT	0.564	0.481	0.566	0.596	**0.552**

E.1.4 GENE-LEVEL EVALUATION

Both tables follow the shared model-level gene-interface protocol in Appendix D.4. UCE and TF-Sapiens retain the same native fixed ESM-derived vectors in their frozen and SPORE PT rows and do not use GenePT alignment.

Table A8: Tissue-specificity transfer performance, with task-vocabulary sizes shown for reference.

Method	Setting	Tissue acc.	Tissue macro-F1	Tissue genes
Non-FM baseline	task-specific	0.353	0.309	292
scGPT	frozen	0.769	0.770	294
scFoundation	frozen	0.319	0.310	295
Geneformer	frozen	0.803	0.799	292
	SPORE PT	**0.814**	**0.813**	292
scCello	frozen	0.613	0.614	298
	SPORE PT	**0.633**	**0.632**	298
TEDDY-G 70M	frozen	0.576	0.577	292
	SPORE PT	**0.627**	**0.632**	291
UCE	frozen	0.563	0.554	294
	SPORE PT	0.563	0.554	294
TF-Sapiens	frozen	0.525	0.527	293
	SPORE PT	0.525	0.527	293

Table A9: UGRN AUPRC for the complete 19 non-hESC Net×Data endpoint matrix across all displayed frozen and SPORE PT conditions. The AVG row reports the unweighted macro across endpoints.

		scGPT	scFoundation	Geneformer		scCello		TEDDY-G 70M		UCE		TF-Sapiens	
Net	**Data**	Frozen	Frozen	Frozen	SPORE PT	Frozen	SPORE PT	Frozen	SPORE PT	Frozen	SPORE PT	Frozen	SPORE PT
STR	hHEP	0.739	0.657	0.742	0.743	0.719	0.721	0.740	0.732	0.792	0.792	0.794	0.794
	mDC	0.635	0.593	0.655	0.650	0.661	0.660	0.635	0.650	0.753	0.753	0.760	0.760
	mESC	0.697	0.632	0.673	0.676	0.699	0.698	0.692	0.689	0.751	0.751	0.776	0.776
	mH-E	0.764	0.636	0.758	0.762	0.716	0.726	0.775	0.778	0.809	0.809	0.795	0.795
	mH-G	0.798	0.644	0.768	0.775	0.751	0.741	0.790	0.792	0.822	0.822	0.811	0.811
	mH-L	0.755	0.652	0.775	0.775	0.758	0.753	0.795	0.815	0.871	0.871	0.861	0.861
NSP	hHEP	0.596	0.543	0.596	0.590	0.569	0.572	0.590	0.580	0.633	0.633	0.633	0.633
	mDC	0.600	0.549	0.593	0.595	0.582	0.583	0.594	0.604	0.682	0.682	0.693	0.693
	mESC	0.641	0.585	0.629	0.628	0.619	0.619	0.635	0.639	0.666	0.666	0.675	0.675
	mH-E	0.675	0.572	0.675	0.672	0.618	0.619	0.666	0.677	0.706	0.706	0.700	0.700
	mH-G	0.689	0.591	0.682	0.674	0.624	0.620	0.676	0.700	0.718	0.718	0.717	0.717
	mH-L	0.667	0.575	0.683	0.661	0.583	0.583	0.639	0.647	0.649	0.649	0.696	0.696
L/G	mESC	0.602	0.551	0.582	0.575	0.559	0.560	0.597	0.581	0.590	0.590	0.581	0.581
SPC	hHEP	0.616	0.606	0.613	0.624	0.594	0.596	0.599	0.609	0.577	0.577	0.606	0.606
	mDC	0.545	0.538	0.510	0.505	0.508	0.510	0.526	0.527	0.560	0.560	0.542	0.542
	mESC	0.689	0.695	0.691	0.695	0.697	0.696	0.688	0.686	0.697	0.697	0.697	0.697
	mH-E	0.726	0.754	0.777	0.767	0.797	0.790	0.771	0.763	0.721	0.721	0.740	0.740
	mH-G	0.746	0.770	0.796	0.790	0.816	0.816	0.763	0.755	0.718	0.718	0.740	0.740
	mH-L	0.676	0.670	0.661	0.648	0.696	0.689	0.683	0.687	0.649	0.649	0.677	0.677
AVG		**0.677**	**0.622**	**0.677**	**0.675**	**0.661**	**0.661**	**0.676**	**0.680**	**0.703**	**0.703**	**0.710**	**0.710**

UGRN label-availability diagnostic. Figure A3 evaluates the matched Geneformer representations across the full label-efficiency curve. SPORE improves macro AUPRC by $+0.004$, $+0.013$, and $+0.014$ at 1, 5, and 10% labeled edges, respectively, and is near parity or slightly lower at 25–100% (-0.003 and -0.004). Within this coverage-matched diagnostic, the representation difference is therefore most favorable in the low-label regime and narrows as more training edges are observed.

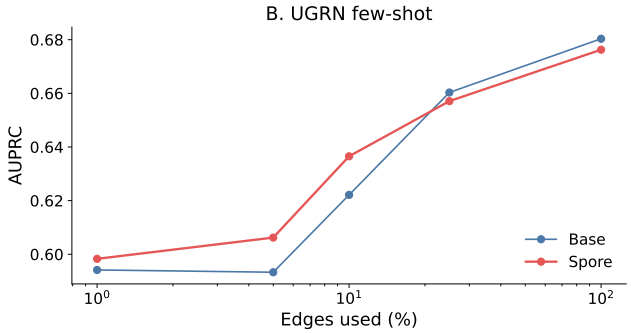

Figure A3: **UGRN label-availability diagnostic for matched Geneformer representations.** Macro AUPRC is averaged over 18 coverage-matched non_hESC endpoints (150,207 evaluation edges, seed 42). SPORE improves transfer at 1–10% labeled edges and is near parity or slightly lower at 25–100%.

E.1.5 PERTURBATION PREDICTION

Table A10: PertEval-scFM Norman_1 results. Entries are mean test MSE at seven displayed SPECTRA split fractions. AUSPC and ΔAUSPC are reported in 10^{-2}; ΔAUSPC is relative to the parameter-free mean baseline.

		Test MSE ↓								
Method	**Setting**	S **0.1**	S **0.2**	S **0.3**	S **0.4**	S **0.5**	S **0.6**	S **0.7**	**AUSPC** $(10^{-2})\downarrow$	**ΔAUSPC** $(10^{-2})\uparrow$
Mean baseline	reference	0.06050	0.05809	0.06904	0.07248	0.06353	0.07974	0.09125	4.7850	0.0000
Non-FM baseline	task-specific	0.05966	0.05793	0.06714	0.06967	0.06275	0.07727	0.09116	4.6927	0.0923
scGPT	frozen	0.05953	0.05794	0.06785	0.07086	0.06322	0.07997	0.09199	4.7462	0.0388
scFoundation	frozen	0.05955	0.05811	0.06784	0.07023	0.06265	0.07743	0.09126	4.7073	0.0777
Geneformer	frozen	0.05963	0.05818	0.06890	0.07351	0.06542	0.08213	0.09450	4.8429	-0.0579
	SPORE PT	0.05959	0.05798	0.06842	0.07256	0.06498	0.08173	0.09364	4.8134	-0.0284
scCello	frozen	0.05956	0.05793	0.06806	0.07202	0.06467	0.08143	0.09363	4.7975	-0.0125
	SPORE PT	0.05955	0.05792	0.06805	0.07192	0.06460	0.08138	0.09360	4.7947	-0.0097
TEDDY-G 70M	frozen	0.05968	0.05830	0.06904	0.07308	0.06473	0.08152	0.09649	4.8377	-0.0527
	SPORE PT	0.05934	0.05781	0.06814	0.07215	0.06448	0.08111	0.09332	4.7879	-0.0029
UCE	frozen	0.05951	0.05795	0.06779	0.07019	0.06267	0.07792	0.09101	4.7080	0.0770
	SPORE PT	0.05930	0.05779	0.06757	0.07006	0.06262	0.07783	0.09108	4.6985	0.0865
TF-Sapiens	frozen	0.05831	0.05724	0.06572	0.06830	0.06258	0.07672	0.09123	4.6326	0.1524
	SPORE PT	0.05829	0.05722	0.06554	0.06818	0.06252	0.07684	0.09118	4.6292	0.1558

E.2 REPRESENTATION ANALYSES

This section provides the exact numerical ledger for the two panels of Figure 3, followed by one auxiliary control for biology-versus-assay accessibility under a matched interface.

E.2.1 NUMERICAL RESULTS FOR FIGURE 3

Held-out AUROC is non-monotonic with label fraction for both Frozen and SPORE, with lower performance in the full-label setting. Domain-wise inspection shows that this behavior is shared across the two representations rather than being specific to post-training, consistent with variation in cross-domain generalization as the source training set expands.

Table A11: **Exact numerical results for the matched Geneformer representation analyses in Figure 3.** All Cancer entries use the frozen linear-probe protocol of Figure 3B and report mean ± SD across four held-out domains. The 100% point denotes using all available downstream training labels under this linear-probe protocol and is therefore not the same endpoint as the task-specific Cancer-Finder readout in Table 1. Ontology geometry is the macro Spearman ρ across 39 tissue strata under the indicated matched readout interface.

A. Label-efficient Cancer linear probe (AUROC)

Labels	Frozen	SPORE	Δ
1%	0.691 ± 0.136	0.746 ± 0.138	+0.055
5%	0.712 ± 0.159	0.755 ± 0.146	+0.043
10%	0.734 ± 0.156	0.754 ± 0.153	+0.020
25%	0.731 ± 0.164	0.744 ± 0.166	+0.013
100%	0.711 ± 0.169	0.725 ± 0.165	+0.014

B. Cross-dataset ontology geometry

Readout interface	Frozen ρ	SPORE ρ	Δ
First token	0.2552	0.3857	+0.1305
Mean pool	0.0754	0.0965	+0.0211

E.2.2 BIOLOGY-VERSUS-ASSAY ACCESSIBILITY CONTROL

Table A12: **Matched biology-versus-assay accessibility control.** Adjusted balanced accuracy is macro-averaged over five dataset-disjoint folds after exact matching by Cell Ontology label, tissue, and disease/state. The 95% confidence intervals are dataset-cluster bootstrap intervals for the paired SPORE–Frozen difference; both include zero. This control does not establish causal removal of technical variation.

Endpoint	Frozen	SPORE	Δ (95% CI)	Datasets
Biology	0.550	0.561	$+0.012$ $[-0.026, +0.070]$	123
Assay	0.426	0.497	$+0.071$ $[-0.017, +0.130]$	162

E.3 DIRECT CELL-TYPE INFERENCE AUDITS

Table A13 reports the extended ID/OOD ontology-distance diagnostics for the direct cell-type inference comparison.

Table A13: Ontology-distance audit on the Direct Inference methods. Error rate is strict-label error frequency; RSL is the asymmetric reference-side diagnostic, LCAE is symmetric distance conditional on a strict error, and LCAA is symmetric distance over all samples. Lower is better. ID contains 10,494 cells (10,493 available UCE predictions); OOD contains 10,500 cells.

	Donor-held-out ID				Study-held-out OOD			
Method	**Error**	**RSL**	**LCAE**	**LCAA**	**Error**	**RSL**	**LCAE**	**LCAA**
CellTypist	0.3629	1.8288	2.7920	1.0132	0.5500	2.0710	3.7314	2.0523
scTab–Linear	0.3107	1.7359	2.8564	0.8874	0.5431	1.9455	3.3042	1.7947
scTab–MLP	0.2982	1.5714	2.7766	0.8279	0.5158	1.7661	3.3674	1.7370
scTab–TabNet	0.2952	1.7995	2.9303	0.8651	0.5202	2.0564	3.8482	2.0018
HCE-scTab TabNet	0.5495	**0.4672**	2.8874	1.5865	0.6817	**1.1942**	3.9444	2.6890
SPORE–Geneformer	0.3611	0.8155	2.9274	1.0570	0.5708	1.5949	3.4205	1.9523
SPORE–UCE	**0.1452**	1.5814	3.1903	**0.4634**	**0.4902**	1.7853	3.1240	**1.5313**
SPORE–TF-Sapiens	0.1921	0.7981	**2.7515**	0.5286	0.5695	1.2426	**2.9577**	1.6845

Metric sensitivity to descendant-compatible errors. RSL and LCAE encode different error semantics. RSL assigns zero reference-side distance when a strict-error prediction is a proper descendant of the reference, whereas LCAE measures symmetric ontology displacement. HCE-scTab has

a higher fraction of descendant-of-reference strict errors than SPORE–TF-Sapiens in both ID and OOD evaluation (Table A14). Because these errors receive zero RSL by definition, the differences between RSL and LCAE reflect their different treatment of descendant-compatible predictions.

Table A14: Descendant-of-reference strict errors for HCE-scTab and SPORE–TF-Sapiens. Percentages use strict errors or all evaluated samples as the denominator, respectively.

Split	Method	Strict errors	Descendant errors	Error share	All-sample share
ID	HCE-scTab	5,766	4,368	75.75%	41.62%
ID	SPORE–TF-Sapiens	2,016	1,058	52.48%	10.08%
OOD	HCE-scTab	7,158	3,516	49.12%	33.49%
OOD	SPORE–TF-Sapiens	5,980	2,434	40.70%	23.18%

E.4 Ablations, Scaling, and Sensitivity

E.4.1 Fixed-compute Data Scaling

Table A15: Fixed-compute post-training data efficiency on Geneformer. The 0% row is the frozen reference; the nested post-training subsets use 2,000 optimizer steps with fixed validation and test partitions.

Fraction	Cell (hAcc) ↑	Cancer (AUROC) ↑	Drug sens. (AUC) ↑
0% (Frozen)	0.649	0.684	0.592
1%	0.470	0.674	0.536
3%	0.564	0.701	0.572
10%	0.679	0.702	0.598
30%	0.737	0.706	0.541
100%	0.778	0.716	0.574

E.4.2 Post-training Hyperparameter Sensitivity

Table A16: Geneformer post-training hyperparameter sensitivity (single seed, seed 37): one hyperparameter of the canonical SPORE configuration is varied at a time and the checkpoint is re-evaluated on the cell-type gate and the Cancer-Finder, Gene-Tissue, and UGRN tasks of the Experiment 1 downstream bundle. Cell-type reports hierarchy-compatible hAcc and hMacro-F1; Cancer-Finder reports the four-domain macro AUROC; tissue specificity reports the FRoGS Gene-Tissue macro-F1; UGRN reports the test-average mean AUPRC.

Configuration	λ_{ctx}	λ_{gene} max	**Backbone LR**	**Cell** hAcc	**Cell** hMacro-F1	**Cancer** AUROC	**Tissue spec.** F1	**UGRN** AUPRC
canonical	0.1	0.05	2×10^{-5}	0.8215	0.4648	0.7429	0.8133	0.6750
λ_{ctx} low	0.05	0.05	2×10^{-5}	0.8221	0.4626	0.7415	0.8616	0.6742
λ_{ctx} high	0.2	0.05	2×10^{-5}	0.8221	0.4623	0.7426	0.8967	0.6762
λ_{gene} max low	0.1	0.025	2×10^{-5}	0.8219	0.4643	0.7428	0.8269	0.6753
λ_{gene} max high	0.1	0.1	2×10^{-5}	0.8225	0.4620	0.7428	0.8130	0.6754
LR low	0.1	0.05	1×10^{-5}	0.8093	0.4502	0.7237	0.8443	0.6780
LR high	0.1	0.05	4×10^{-5}	0.8313	0.4826	0.7496	0.8285	0.6786

The reported SPORE configuration uses $\lambda_{\text{ctx}} = 0.1$, λ_{gene} max 0.05, and a backbone LR of 2×10^{-5} (Appendix C.5). Table A16 varies one of these quantities at a time around the Geneformer configuration using a single post-training seed, and re-evaluates cell-type transfer, Cancer-Finder, Gene–Tissue, and UGRN under their standard downstream protocols (Appendices D.1, D.2, D.4.1, and D.4.2).

Across the seven variants, cell-type hAcc remains within ± 0.012 of the canonical 0.8215, Cancer-Finder AUROC within ± 0.013 of 0.743, and UGRN AUPRC within ± 0.003 of 0.675. Gene–Tissue varies more broadly (0.813–0.897), with the largest value occurring at the higher context weight. Thus, the cell-type, Cancer-Finder, and UGRN results are relatively stable across these local one-at-a-time variations, while Gene–Tissue shows greater sensitivity to the tested loss weights.

F Extended Case Studies and Diagnostics

F.1 Neocortex Regional-composition Case

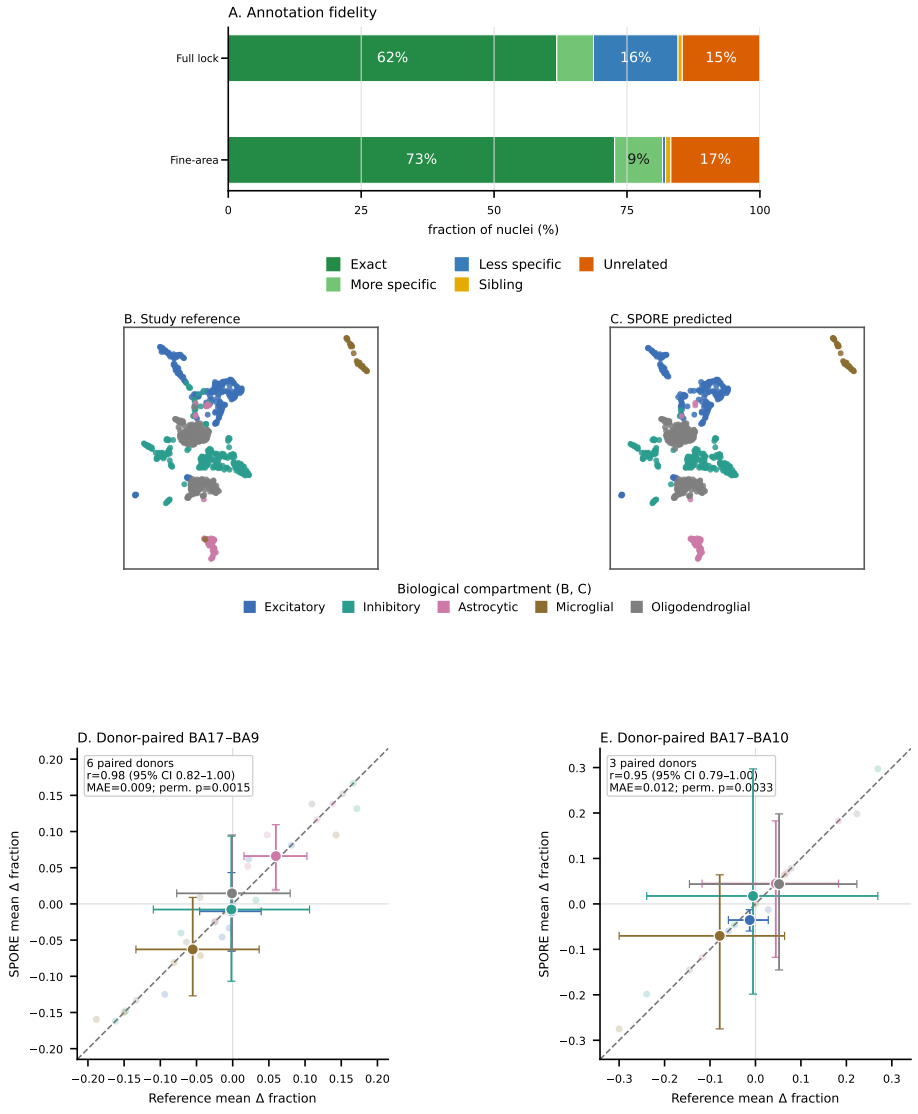

Figure A4: **Complete audit of the held-out neocortex regional case.** (A) Five-way ontology relationship between direct predictions and study labels in the full evaluation set and fine-area subset. (B–C) The same post-hoc UMAP layout colored by the five prespecified biological compartments derived from study reference labels (B) and SPORE predictions (C). These projections are illustrative and are not used for quantitative inference. (D–E) Donor-paired BA17–BA9 and BA17–BA10 five-compartment contrasts with donor-bootstrap 95% CIs. Permutation p values shown in the figure use the within-donor randomization procedure described above.

The main-text case study uses the held-out developing-human-neocortex study of Wang et al. (2025). The evaluation contains 1,000 nuclei from 27 donors. Across the full set, 618 predictions are exact, 69 more specific, 159 less specific, 8 sibling, and 146 unrelated to the reference label, corresponding to strict accuracy 0.618 and descendant-compatible accuracy 0.687.

The regional analysis uses 699 nuclei from 13 donors with fine BA17/V1, BA9, or BA10 labels. Within this subset, 508 predictions are exact, 63 more specific, 4 less specific, 7 sibling, and 117 unrelated (strict accuracy 0.727; descendant-compatible accuracy 0.817). Cell Ontology labels are collapsed into five prespecified compartments: excitatory, inhibitory, astrocytic, microglial, and oligodendroglial, with the last combining the oligodendrocyte and oligodendrocyte-precursor branches.

Regional composition is compared within paired donors. For each donor and compartment, we compute the BA17 fraction minus the comparison-region fraction independently from the study annotations and the direct SPORE annotations. Pearson r and MAE summarize agreement between the resulting paired contrasts, with 95% confidence intervals obtained by resampling paired donors.

As a complementary randomization diagnostic, direct-label compartment assignments are permuted between the two regions within each donor while preserving donor-level compartment totals and regional sample sizes.

Table A17: **Paired regional-composition statistics for the held-out neocortex case.** Composition contrasts are five-compartment donor-paired means; r and MAE compare study-reference and SPORE contrasts, with donor-bootstrap 95% CIs. The reported p value is obtained from within-donor permutation of direct-label compartment assignments between regions while preserving donor totals and region sizes.

Region pair	Donors	Cells	r (95% CI)	MAE (95% CI)	p
BA17–BA9	6	405	0.980 [0.823, 0.998]	0.009 [0.004, 0.020]	0.0015
BA17–BA10	3	193	0.945 [0.795, 0.999]	0.012 [0.010, 0.016]	0.0033

The six BA17–BA9 paired donors are NIH-4341, NIH-4373, NIH-4458, NIH-5162, NIH-5376, and NIH-5554. The three BA17–BA10 paired donors are NIH-1325, NIH-1671, and NIH-4392.

Table A18 reports the compartment-level donor-paired contrasts and donor-bootstrap intervals used in the two scatter panels.

Table A18: **Compartment-level donor-paired contrasts for the unseen-neocortex case.** Each entry is the mean BA17-minus-comparison-area fraction difference with a donor-bootstrap 95% CI. Reference and SPORE values are computed from the same paired nuclei summarized in Table A17.

Region pair	Compartment	Reference Δ (95% CI)	SPORE Δ (95% CI)
BA17–BA9	Excitatory	-0.002 [-0.046, 0.039]	-0.010 [-0.065, 0.043]
	Inhibitory	-0.002 [-0.110, 0.106]	-0.008 [-0.107, 0.094]
	Astrocytic	0.060 [0.016, 0.103]	0.066 [0.019, 0.109]
	Microglial	-0.055 [-0.134, 0.037]	-0.063 [-0.127, 0.009]
	Oligodendroglial	-0.001 [-0.077, 0.079]	0.015 [-0.065, 0.095]
BA17–BA10	Excitatory	-0.013 [-0.060, 0.028]	-0.035 [-0.060, -0.013]
	Inhibitory	-0.005 [-0.239, 0.269]	0.017 [-0.198, 0.297]
	Astrocytic	0.045 [-0.118, 0.183]	0.045 [-0.118, 0.183]
	Microglial	-0.079 [-0.300, 0.064]	-0.070 [-0.275, 0.064]
	Oligodendroglial	0.052 [-0.145, 0.223]	0.044 [-0.145, 0.198]

At the compartment level, the reference and direct-annotation contrasts are closely aligned for BA17–BA9, including the directions of the astrocytic and microglial differences. BA17–BA10 shows a similar qualitative pattern but has wider uncertainty because only three donors contribute paired observations.

F.2 Low-accuracy OOD Annotation Case

We inspect the large B-cell lymphoma study with the lowest strict agreement in the final study-disjoint OOD cohort (Li et al., 2025). It contains 1,000 lymph-node nuclei profiled by 10x multiome with 14 Cell Ontology reference labels. The frozen direct-inference interface annotates all nuclei without target-specific fitting.

Figure A5 decomposes the predictions into exact (25.3%), more-specific descendant (22.0%), less-specific ancestor (42.1%), sibling (1.8%), and unrelated (8.8%) relations. Much of the strict error is therefore hierarchical rather than unrelated: immune references are frequently mapped to broader T-cell categories, whereas some endothelial references are mapped to more specific vessel subtypes. A separate name-based audit flags 51 of 220 descendant predictions with an explicit organ-restricted modifier inconsistent with the lymph-node source. This diagnostic resolves how strict annotation errors are distributed across ontology relations.

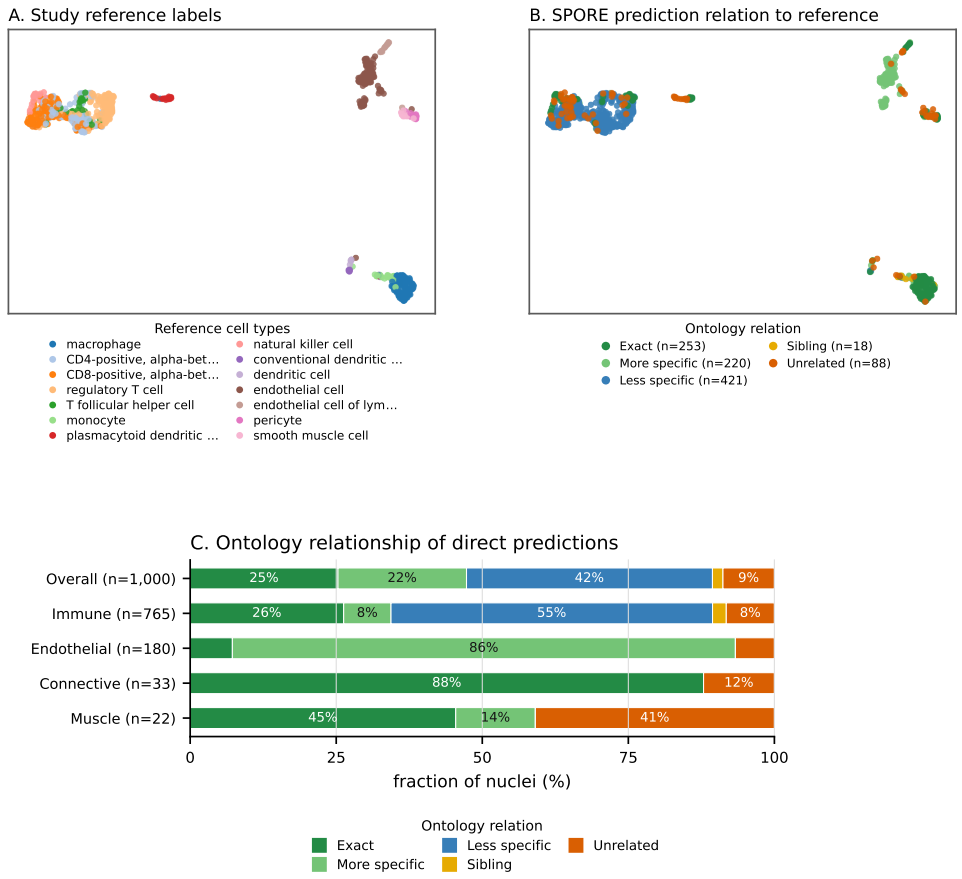

Figure A5: **Qualitative diagnostic of ontology-structured prediction errors in the large B-cell lymphoma study.** (A) The 1,000 nuclei colored by 14 reference labels. (B) The same UMAP colored by exact, more-specific, less-specific, sibling, or unrelated prediction–reference relations. (C) Their distribution overall and within four reference-lineage branches. The UMAP is display-only; reported shares are computed directly from the predictions.