TOTICELL: STRUCTURED MASKING OVER FULL-COVERAGE GENE BLOCKS FOR SINGLE-CELL REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

A masking ratio specifies how much expression is hidden during pretraining, but not which genes are hidden together. We introduce TotiCell, a compact single-cell representation model whose block tokenizer represents the full set of 21,912 genes—including genes with zero expression in a given cell—as a fixed sequence of 48 tokens, roughly 1/67 of the input length of the audited gene-level baselines. TotiCell is trained with GroupMask, a structured masking policy that groups the 48 blocks into three disjoint, exhaustive target sets and masks one entire set per cell. GroupMask preserves RandomMask's per-block masking probability and visible-token budget while replacing the combinatorial mask distribution with three recurring target sets. Trained on 50.5 million cells, GroupMask leads five of seven task means over RandomMask, most clearly in perturbation retrieval, where all three datasets favor it (mean MRR gain 0.0293). Scaling the recipe yields the highest annotation, perturbation-retrieval, and mapping means among the evaluated methods with complete task coverage, and exceeds CellVQ on five of seven task means with 55.5% fewer total parameters. The scaled model also achieves the best mean rank percentile among the nine complete-coverage methods. In a controlled single-RTX-3090 benchmark, the smaller architecture processes over 1,600 cells per second while using under 0.5 GiB of memory. Three-seed probes favor recurring mask sets and reduced zero-target loss weight, and characterize partition and block-count sensitivity.

1 INTRODUCTION

Single-cell transcriptomics measures cellular identity and state across tissues, donors, and experimental conditions. Foundation models seek to turn these sparse, high-dimensional measurements into reusable representations (Theodoris et al., 2023; Cui et al., 2024; Hao et al., 2024). Recent systems have expanded both data and model scale: scFoundation uses about 100 million parameters and over 50 million cells, CellFM scales to 800 million parameters, and CellVQ reports 511 million parameters pretrained on more than 68 million cells (Hao et al., 2024; Zeng et al., 2025; Wang et al., 2026). This progress raises a complementary question: how far can representation quality be driven by the pretraining design—by how a model represents its input and defines its reconstruction objective—rather than by scale alone?

We introduce TotiCell, a compact single-cell representation model built on a block tokenizer over a fixed gene set. TotiCell assigns all 21,912 genes to 48 gene blocks and encodes each block as a single token, so every cell is represented by a constant-length sequence of 48 block tokens. Unlike the audited CellVQ and scFoundation input sequences, TotiCell retains the full fixed gene set, including genes with zero measured expression. Those zero-expression genes are not discarded: they remain in their block vectors and, when their block is masked, enter the reconstruction objective with a reduced weight. The audited gene-level sequences grow with the expressed-gene count and average roughly 3,200 tokens per cell, versus 48 for full-input TotiCell evaluation. The block tokenizer is fixed and identical across our masking comparisons.

A masking ratio fixes how much information is hidden during training, but not the structure of the mask sets. **RandomMask** over gene blocks draws any 16 of 48 blocks, so every one of the $\binom{48}{16}$

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.

subsets is possible. We train TotiCell with **GroupMask**, a structured masking policy that constrains these sets. The 48 gene blocks—themselves a fixed, balanced, random partition of the genes—are randomly divided into three disjoint, exhaustive target sets of 16. For each cell, GroupMask masks one entire target set and keeps the remaining 32 blocks visible. Every block is masked with probability $1/3$, exactly as under RandomMask's 16-of-48 draw, and the visible-token budget is unchanged. GroupMask is therefore strictly more constrained than RandomMask: it restricts masking to three recurring mask patterns rather than all $\binom{48}{16}$ subsets, while matching the single-block marginals. It requires only a fixed partition and one categorical draw per cell, with no learned mask generator.

We study TotiCell in a gene-block masked autoencoder pretrained on 50.5 million cells. GroupMask and RandomMask share the same architecture, training budget, and visible-token count. Under validation-based checkpoint selection, GroupMask leads five of the seven task means over its matched RandomMask checkpoint, most clearly in perturbation retrieval, where all three sources favor it. Scaling the same recipe, TotiCell leads three complete-coverage task columns and exceeds CellVQ on five of seven task means, using 44.5% of its reported parameter count. Its constant 48-token input also yields a large measured throughput and memory advantage in a controlled single-GPU inference benchmark.

Our contributions are fourfold. (1) We introduce TotiCell, a compact model whose block tokenizer represents all 21,912 genes, including zero-expression genes, in a fixed sequence of 48 tokens—roughly 1/67 of the audited gene-level input length, reducing sequence-processing cost. (2) We formulate GroupMask and characterize how it constrains the mask-set distribution while preserving single-block marginals and the visible-token budget. (3) We evaluate representations on 22 settings from 11 datasets, showing that GroupMask improves over RandomMask in most task families—most pronounced in perturbation retrieval—and that the scaled model transfers strongly on donor-disjoint human evaluation cohorts. (4) We analyze the design through paired comparisons, three-seed mask-support and objective probes, block-count sensitivity, and a controlled inference benchmark. The combined evidence characterizes the effectiveness of GroupMask, how the recipe scales, and which aspects merit further study.

2 Related work

Single-cell foundation models. scBERT, Geneformer, and scGPT adapt Transformer-based pretraining to gene expression (Yang et al., 2022; Cui et al., 2024; Theodoris et al., 2023). Their input representations differ: scBERT adds gene2vec identity embeddings to binned expression tokens, scGPT sums separate gene and expression-value embeddings, and Geneformer uses a rank value encoding that orders genes by normalized expression. scFoundation uses an asymmetric encoder–decoder, CellFM a retention-based architecture, and CellVQ a discrete cell codebook (Hao et al., 2024; Zeng et al., 2025; Wang et al., 2026). CellPLM instead treats cells as tokens to model inter-cellular context (Wen et al., 2023). RegFormer introduces regulatory priors into representation learning (Hu et al., 2026). These are distinct architectural and biological-prior choices. TotiCell compresses the full fixed gene set into 48 block tokens; our input-length comparison concerns the audited CellVQ and scFoundation encoders, not every single-cell model.

Reconstruction objectives and masking policies. Masked autoencoders separate a visible-input encoder from a reconstruction decoder (He et al., 2022). TxFM applies masked autoencoding to transcriptomes using individual gene tokens, visible genes sampled from the measured gene set, and reconstruction through a cell-level bottleneck (Kenyon-Dean et al., 2026). CorrMask addresses a different design axis: it uses an expression-derived dependency graph to mask correlated genes jointly (Hacohen et al., 2026). A masking policy is not itself an encoder architecture. GroupMask requires neither a dependency graph nor a learned mask generator: it chooses among three fixed, exhaustive target sets over randomly formed gene blocks. The masking comparison matches architecture, training budget, single-block masking probabilities, and visible-token count, with checkpoint selection defined in Section 4.1.

Evaluating reusable representations. Benchmarks find substantial variation across single-cell tasks and competitive performance from classical representations (Kedzierska et al., 2025; Wu et al., 2025). We therefore report task-specific scores and retain PCA-HVG alongside pretrained models.

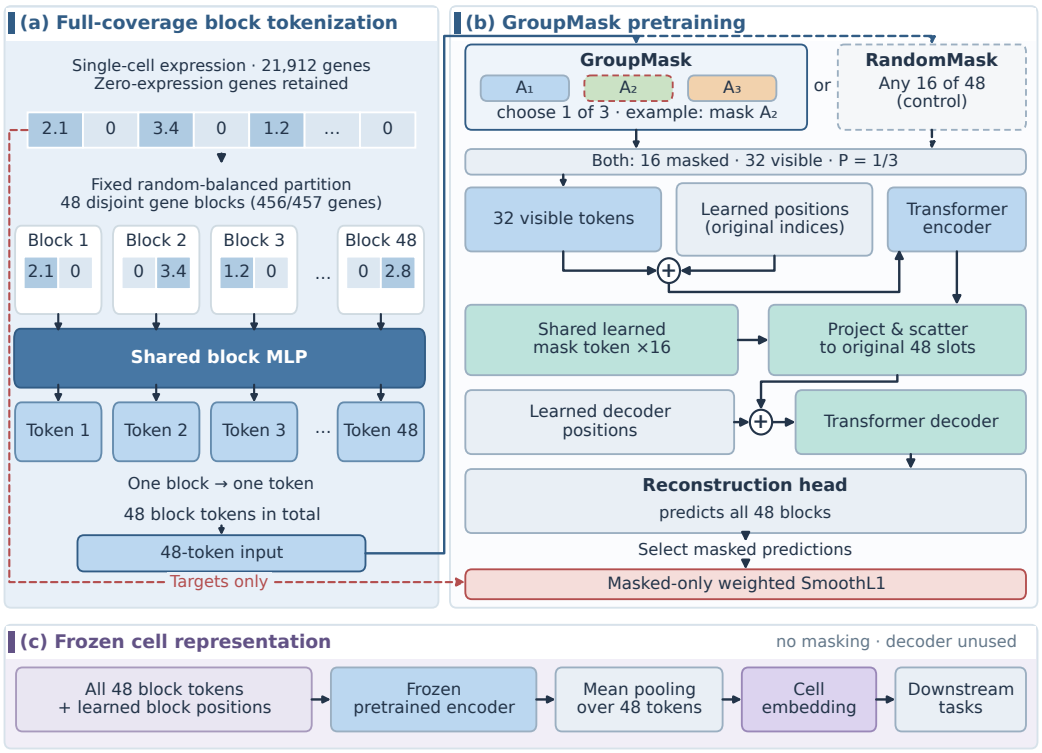


Figure 1: TotiCell architecture. (a) Full-coverage tokenization retains 21,912 genes, including per-cell zeros; four blocks illustrate the shared MLP mapping. (b) GroupMask chooses one of three fixed 16-block target sets; RandomMask samples any 16 blocks. Encoder and decoder positions are added separately; masked expression supplies loss targets only. (c) Frozen evaluation mean pools all 48 encoder outputs, without the decoder.

The matched masking experiment tests a pretraining design choice, while the 22-setting evaluation suite characterizes the resulting representations across tasks.

3 METHOD

3.1 TOTICELL: FULL-COVERAGE GENE-BLOCK TOKENIZATION

Let $x \in \mathbb{R}^G_{\geq 0}$ denote the aligned expression profile for a cell. The fixed gene set contains $G = 21{,}912$ genes. These genes are divided at random into $B = 48$ non-overlapping, balanced blocks of 456 or 457 genes each; smaller blocks receive one zero padding slot, excluded from the reconstruction loss.

For count inputs, the pipeline uses library-size normalization to 10^4 followed by $\log(1 + x)$; benchmark inputs already supplied in normalized form use the recorded dataset-specific transform. A shared MLP maps each visible block to one token (Figure 1(a)). Learned position embeddings indexed by the original block identities are added before the encoder. During pretraining, only 32 visible tokens are encoded; their projections are scattered into the original 48-position order, with a shared learned mask token filling the 16 masked positions. Separate decoder-position embeddings are then added before the reconstruction decoder (Figure 1(b)).

Design rationale. Three properties motivate the design. (i) Every gene in the fixed set belongs to one block; per-cell zero-expression targets receive weight λ_0 when masked. (ii) Block compression fixes the full input at 48 tokens, with 32 visible during pretraining. (iii) GroupMask repeatedly reconstructs each block from the same co-visible blocks. Full gene coverage and compression come

from tokenization; zero-target weighting controls the loss, and the third property is the masking intervention examined in Section 5.3 and the budget-matched proxy experiments.

3.2 GROUPMASK: CONSTRAINING THE MASK SUPPORT TO THREE TARGET SETS

Masks are defined over the block indices $\mathcal{B} = \{1, \dots, B\}$, and $m = 16$ blocks are masked per cell. **RandomMask** draws M uniformly from the $\binom{B}{m}$ subsets of size m, so every 16-block subset is possible. **GroupMask** fixes a partition $\mathcal{B} = A_1 \dot\cup A_2 \dot\cup A_3$ with $|A_j| = m$, formed by randomly dividing the 48 blocks; the three sets carry no biological meaning. It then draws J uniformly from $\{1, 2, 3\}$ and sets $M = A_J$. In both cases the visible set is $V = \mathcal{B} \setminus M$.

For each block b,

$$\Pr_{\text{RandomMask}}(b \in M) = m/B = 1/3 = \Pr_{\text{GroupMask}}(b \in M). \tag{1}$$

The equality holds for every gene assigned to a block and, by linearity, for the expected total masked expression positions. It describes the policy distribution; realized per-minibatch group frequencies need not be exactly balanced.

Let $Z_b = \mathbf{1}\{b \in M\}$. For distinct blocks,

$$\mathbb{E}_{\text{RandomMask}}[Z_b Z_c] = \frac{m(m-1)}{B(B-1)}, \qquad \mathbb{E}_{\text{GroupMask}}[Z_b Z_c] = \begin{cases} 1/3, & b, c \in A_j \text{ for some } j, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Thus the masking ratio alone does not specify which blocks are hidden together, and therefore does not determine the reconstruction task. GroupMask reduces the mask support from $\binom{48}{16} = 2{,}254{,}848{,}913{,}647$ patterns to three and is strictly more constrained than RandomMask, while matching the single-block marginals and the visible-token count. The mask-set entropies are 28.444 nats for RandomMask and 1.099 nats for GroupMask (natural logarithms). The resulting recurring context–target assignments are the design rationale for GroupMask, and Appendix I visualizes the exact co-masking probabilities.

3.3 OBJECTIVE AND FROZEN READOUT

Let $\mathcal{T}$ be all valid masked gene positions across a minibatch. With $w(y) = 1$ for positive targets and $w(0) = \lambda_0$, the implemented objective is

$$\mathcal{L} = \frac{\sum_{(i,g)\in\mathcal{T}} w(x_{ig})\ \text{SmoothL1}(\hat{x}_{ig} - x_{ig})}{\sum_{(i,g)\in\mathcal{T}} w(x_{ig})}, \quad \lambda_0 = 0.2, \tag{3}$$

where λ_0 is the loss weight for zero-expression targets. SmoothL1 uses transition parameter one. Reduced weighting retains zero-target supervision while limiting its aggregate contribution in sparse profiles; $\lambda_0 = 0$ removes these targets from the loss, not from the input.

At evaluation, the frozen encoder receives all 48 block tokens with their learned block-position embeddings (Figure 1(c)). Let $H \in \mathbb{R}^{48\times d}$ denote its final tokens. We use

$$z(x) = \frac{1}{48}\sum_{b=1}^{48} H_b \in \mathbb{R}^d. \tag{4}$$

The decoder is unused at inference.

4 EXPERIMENTAL PROTOCOL

4.1 MODELS AND TRAINING

The method study uses two 96M TotiCell models that differ only in their masking policy, denoted **GroupMask** and **RandomMask** throughout. Both use encoder width 768, depth 12, and 12 heads, with a width-384, depth-4 decoder, totaling 96,450,121 parameters (88,861,440 in the encoder). They share the same training corpus of 50,466,718 cells, block partition, training seed (42), global

Table 1: Frozen-representation benchmark scores (higher is better), averaged equally over available settings within each task. Params (M) gives complete-model parameter counts; PCA's – denotes not applicable. Bold marks the best mean among methods with complete task coverage; † marks partial means (Table 6); – in score columns means no result. Partial means are excluded from best-score highlighting and overall ranks. GroupMask and RandomMask denote the masking comparison; TotiCell denotes the scaled model, as defined in Section 4.1. TotiCell models use mean readouts; CellFM uses nonzero-token mean. PCA is fitted within training folds for supervised tasks and on target cells for integration. Task procedures and coverage are detailed in the appendix.

Method	Params (M)	Ann.	Int.	Few-shot	Disease	Perturb.	Mapping	OOD
Settings / metric		5 / F1	6 / scIB	2 / AULC	2 / AUC	3 / MRR	2 / F1	2 / AUC
GroupMask	96.45	0.8851	0.5948	0.8006	0.8550	0.7909	0.8160	**0.9348**
RandomMask	96.45	0.8847	0.5891	**0.8066**	0.8592	0.7616	0.8137	0.9348
TotiCell	227.36	**0.8892**	0.5836	0.8057	0.8731	**0.8101**	**0.8307**	0.9289
CellVQ	511.00	0.8673	0.5853	0.7892	**0.8889**	0.7284	0.6848	0.8802
scFoundation	100.00	0.8822	**0.5965**	0.7764	0.8521	0.7491	0.7341	0.9048
scGPT	51.33	0.8273	0.5776	0.7137	0.8603	0.7256	0.7140	0.7691
Geneformer	104.00	0.8543	0.5581	0.7287	0.8659	0.7239	0.7889	0.7713
CellFM	80.00	0.7858	0.4875	0.5806	0.8066	0.7109	0.7564	0.4951
CellPLM	85.00	0.8854†	0.6128†	0.6823	0.9276†	0.6338†	0.6048	0.8165
PCA-HVG	–	0.8434	0.5468	0.5559	0.8004	0.8047	0.7645	0.5770

batch size 768, learning rate 4×10^{-4}, weight decay 0.05, and warmup of 0.1 corpus pass. Within a common four-pass training budget, both use their best validation-proxy checkpoint—GroupMask at three passes and RandomMask at four—with GroupMask's earlier optimum indicating faster convergence on this criterion.

The scaled model, denoted **TotiCell** in comparisons, uses width 1024, depth 16, and 16 heads with a width-512, depth-6 decoder, totaling 227,363,785 parameters. It shares the corpus, block partition, GroupMask target sets, learning rate, and four-pass training endpoint; a physical batch of 384 with two accumulation steps matches the effective batch of 768. Frozen mean readouts have dimension 768 for the masking comparison and 1024 for the scaled model. No task selects its own checkpoint or pooling.

4.2 Evaluation suite

Frozen representations are compared on 22 settings drawn from 11 datasets and spanning seven task families: five annotation settings, six integration settings, two few-shot curves (summarized by normalized log-shot AULC), two disease-state settings, three perturbation-retrieval settings, two cross-atlas mappings, and two OOD settings. Annotation and mapping train supervised probes on frozen embeddings, few-shot varies the labeled cells per class over 1, 5, 10, 20, 50, and 100, integration uses the composite $0.6\,\mathrm{avg}_{\mathrm{bio}} + 0.4\,\mathrm{avg}_{\mathrm{batch}}$, perturbation evaluates pseudo-replicate retrieval, and OOD uses distances to known-class centroids. Splits, coverage, and uncertainty are detailed in Appendix B.

The human-donor cohorts in the primary evaluation have no donor overlap with TotiCell's pretraining corpus.

CellVQ, scFoundation, scGPT, Geneformer, CellFM, and PCA-HVG cover all 22 settings; CellPLM covers 18. Task means weight available settings equally within each family and never combine different metrics. Incomplete means are marked †; coverage counts appear in Table 6.

5 Frozen representation results

5.1 A compact model with strong multi-task performance

Table 1 reports frozen-representation scores together with complete-model parameter counts. **TotiCell** achieves the highest complete-coverage task means in annotation (0.8892), perturbation retrieval (0.8101), and cross-atlas mapping (0.8307). These leading results span supervised transfer, perturbation-state identification, and atlas-to-atlas transfer.

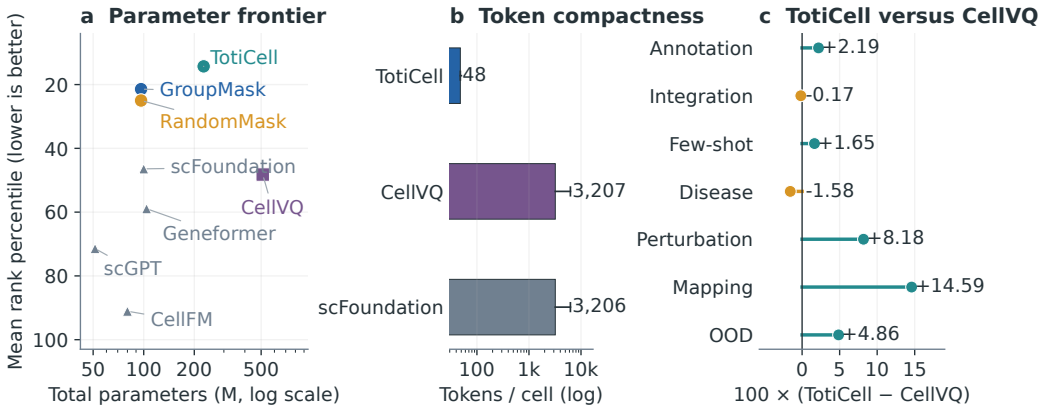


Figure 2: Compact model, frontier performance. (a) Total parameters (log scale) versus mean rank percentile over seven task families (lower is better), using the nine methods with all 22 scores. CellPLM is excluded for incomplete coverage; PCA-HVG enters the ranking but has no plotted pretrained-model parameter count. (b) Encoder input tokens per cell on the same 16,382 Pancreas cells: bars show means and ticks the 90th percentile; Table 2 reports medians. Both TotiCell architectures use 48 tokens, while the audited gene-level tokenizers average about 3,200. (c) TotiCell minus CellVQ for all seven task means, in units of $100\times$ the metric; positive values favor TotiCell. Rank is descriptive and roster-dependent, and parameter counts are not FLOPs or runtime.

Against CellVQ, a strong comparator at more than twice the size, TotiCell exceeds it on five of seven task means: annotation by 2.19 score points, few-shot AULC by 1.65, perturbation retrieval by 8.18, mapping by 14.59, and OOD AUROC by 4.86, where a score point denotes 0.01 of the corresponding metric (Figure 2(c)). CellVQ remains stronger by 0.17 score points on integration and 1.58 on disease-state classification. TotiCell has 55.5% fewer total parameters and is trained on 50.5M cells versus CellVQ's reported more than 68M (Wang et al., 2026).

The perturbation-retrieval results illustrate a consistent source-level advantage. TotiCell exceeds CellVQ on Dixit, Norman, and Adamson by 9.71, 4.77, and 10.05 score points, respectively (Figure 3). These are retrieval tests of observed perturbation representations, rather than predictions of unseen responses.

Figure 2(a) places the comparison on a parameter–performance frontier. Across the seven task families, TotiCell attains the best mean rank percentile (14.3) among the nine methods with complete coverage, followed by GroupMask (21.4) and RandomMask (25.0), versus 48.2 for CellVQ. Among methods with complete coverage of each task, scFoundation leads integration (0.5965), CellVQ disease classification (0.8889), RandomMask few-shot transfer (0.8066), and GroupMask OOD detection (0.9348). PCA-HVG remains competitive on perturbation retrieval (0.8047).

5.2 Efficiency by design

Figure 2(b) and Table 2 isolate the cost of the representation. On the shared pancreas cells, the block encoder receives 48 tokens per cell regardless of how many genes that cell expresses, whereas the gene-level CellVQ and scFoundation tokenizers average roughly 3,200 tokens and reach a 90th percentile near 6,300. The reduction comes from block compression, not gene removal: all 21,912 genes, including zero-expression genes, remain in the block vectors and the reconstruction objective.

Table 2 reports single-GPU inference efficiency on 16,382 Pancreas cells. At batch 1, per-cell latency is 3.1 and 4.1 ms for the two TotiCell models, against 71.7 and 72.0 ms for CellVQ and scFoundation. At its best batch, the smaller model processes about 1,605 cells per second (batch 64) and TotiCell about 744 cells per second (batch 32), versus about 6.6 cells per second for CellVQ and scFoundation, which are limited to batch 1 under a 22 GiB cap; the speedups are about 240-fold and 113-fold. Peak allocated memory stays below 0.9 GiB for both TotiCell models, against about 22 GiB for CellVQ and 16 GiB for scFoundation. These measurements are specific to the frozen checkpoints, the pancreas data, and one RTX 3090.

Table 2: Inference efficiency on one RTX 3090 (24 GiB) on 16,382 Pancreas cells. B=1 is batch-1 latency; Batch gives the batch for throughput and peak allocated memory. Cells/s excludes preprocessing; E2E includes it. Model loading and disk I/O are excluded. CellVQ and scFoundation fit only batch 1 under the 22 GiB cap. Values are three-repeat medians; speedup is relative to CellVQ. Tokens/cell are medians. Full timing and variability details appear in Appendix A.

Model	Params (M)	Tokens	B=1 (ms)	Batch	Cells/s	GiB	E2E cells/s	Speedup
GroupMask	96.45	48	3.12	64	1,605.4	0.43	1052.7	244.5×
TotiCell	227.36	48	4.15	32	744.0	0.84	515.8	113.3×
CellVQ	511.00	2,504.5	71.70	1	6.6	21.89	5.4	1.0×
scFoundation	100.00	2,504.5	72.03	1	6.6	15.78	5.5	1.0×

Table 3: Two central comparisons at the task-mean level. Left: GroupMask versus RandomMask. Right: TotiCell versus CellVQ. Differences are $100\times$(GroupMask minus RandomMask) and $100\times$(TotiCell minus CellVQ), respectively. Positive differences are bold. Checkpoint selection is defined in Section 4.1; all task metrics and settings match Table 1.

	Masking comparison			Model comparison		
Task	RandomMask	GroupMask	100Δ	CellVQ	TotiCell	100Δ
Annotation	0.8847	0.8851	**+0.049**	0.8673	0.8892	**+2.190**
Integration	0.5891	0.5948	**+0.571**	0.5853	0.5836	-0.168
Few-shot	0.8066	0.8006	-0.603	0.7892	0.8057	**+1.651**
Disease	0.8592	0.8550	-0.423	0.8889	0.8731	-1.576
Perturbation	0.7616	0.7909	**+2.927**	0.7284	0.8101	**+8.179**
Mapping	0.8137	0.8160	**+0.239**	0.6848	0.8307	**+14.591**
OOD	0.9348	0.9348	**+0.0002**	0.8802	0.9289	**+4.862**

5.3 GroupMask improves perturbation retrieval

The same-size comparison examines the masking policy under a shared training and checkpoint-selection protocol. GroupMask improves over RandomMask in five of the seven task families, with the clearest gain in perturbation retrieval. Perturbation-retrieval MRR increases from 0.7616 to 0.7909, a gain of 2.927 score points (Table 3, left). Integration, mapping, and annotation improve by 0.571, 0.239, and 0.049 score points, respectively; few-shot and disease transfer favor RandomMask. OOD is effectively tied, with a mean AUROC difference of only 1.70×10^{-6}.

All three perturbation sources favor GroupMask: Dixit improves by 5.960 score points, Norman by 2.407, and Adamson by 0.413 (Figure 3). All 22 setting-level differences, including those favoring RandomMask, appear in Appendix Figure 5.

5.4 Complementary strengths across settings

The two comparisons agree on the perturbation-retrieval direction, but their other strengths differ. TotiCell exceeds CellVQ in both mapping directions; GroupMask gains over RandomMask on AIDA-to-HIHA but loses on HIHA-to-AIDA. Few-shot transfer on HYPOMAP and AIDA favors RandomMask over GroupMask. Appendix J shows all four central models at every label budget. These dataset-level contrasts complement the task means rather than implying universal superiority.

6 Smaller-scale mask-design and robustness probes

We use a separate 30,653,641-parameter model, five million training cells, two passes (9,766 steps), physical batch 512 with accumulation two, and training seeds 42–44. The fixed endpoint proxy is the mean of linear-probe and kNN macro-F1 on a fixed 20,000/10,000-cell training/validation split, with evaluation seed 42 and all block tokens visible. It is distinct from the 22-setting downstream suite. Error bars use sample SD across the three training seeds; shared baseline runs are not additional independent replicates.

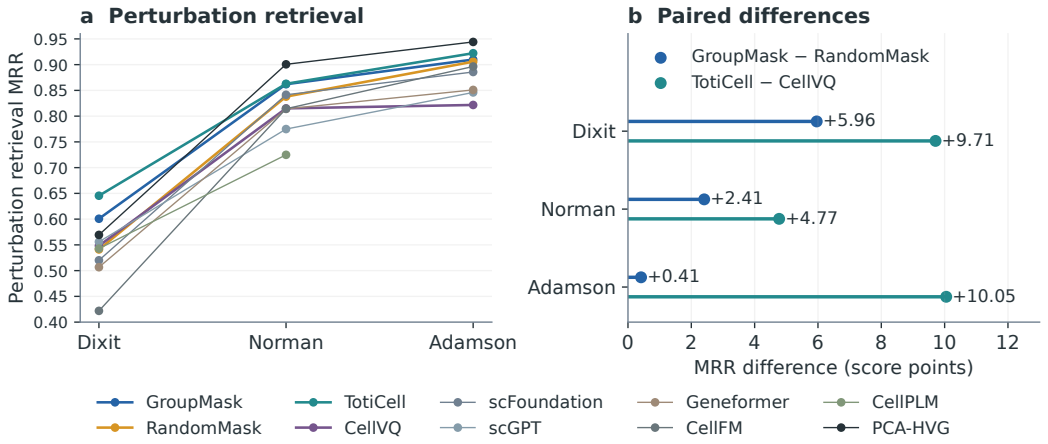


Figure 3: Perturbation retrieval as a representative transfer case. (a) All three datasets and all available comparator results, including PCA-HVG; CellPLM–Adamson is unavailable. Lines connect dataset means for readability. (b) GroupMask versus RandomMask and TotiCell versus CellVQ improve on all three sources. MRR averages three artificial query/gallery splits of observed perturbations and does not measure response generation or independent biological replication. Full setting results, including negative differences in other tasks, are retained in Appendix Figure 5.

Matched marginals and expanded mask support. GroupMask with one three-set partition scores 0.471152 ± 0.003042, versus 0.435037 ± 0.005619 for RandomMask. Paired gains for seeds 42–44 are $+0.0313$, $+0.0324$, and $+0.0446$. To test support size, we add complete partitions of the same 48 blocks. **Bank-6** (coverage depth $k = 2$) draws one of six size-16 target sets from two partitions; **Bank-9** ($k = 3$) draws one of nine sets from three partitions. Every block still has masking probability $1/3$, and each cell uses one draw, not multiple masking operations.

Bank-6 scores 0.456603 ± 0.002991 and Bank-9 0.450127 ± 0.001272. All three seeds follow the order **GroupMask** $>$ **Bank-6** $>$ **Bank-9** $>$ **RandomMask** (Figure 4(a)). Thus, expanding support lowers the proxy within these tested banks while remaining above RandomMask. This supports recurring mask structure in this setting; it does not establish a universal entropy law, since support size, co-masking relationships, and context recurrence change together. Twelve- and twenty-four-pattern banks remain untested.

Gene partitions, block count, and zero targets. A training-expression dependency graph provides a complementary test of the random tokenizer. Under RandomMask, grouping positively related genes into the same block scores 0.441859 ± 0.002943, dispersing them scores 0.436683 ± 0.000614, and random blocks score 0.435037 ± 0.005619. Grouped exceeds random on every seed, so the evidence does not show that biological structure is useless. Instead, random blocks remain a simple, prior-free choice with a larger observed gain from GroupMask in this proxy; the interaction between graph-based partitions and GroupMask is untested (Appendix E).

With GroupMask and a fixed masking fraction of $1/3$, B36, B48, and B72 score 0.473832, 0.471152, and 0.472337, respectively. The small differences support local robustness, not the optimality of B48; parameter counts and token costs also change (Appendix F). Reduced zero-target weighting gives the highest observed proxy (Figure 4(b)): $\lambda_0 = 0.2$ scores 0.4712, versus 0.4035 at 0 and 0.4617 at 1 (gains 0.0677 and 0.0095). This supports retaining downweighted zero-target supervision. Random-partition and target-grouping sensitivity are retained in Appendix H.

7 Discussion and Conclusion

TotiCell makes the joint reconstruction task an explicit design choice for single-cell pretraining. Three exhaustive target sets produce competitive representations from random gene blocks, without a dependency graph or a learned masking module, and the full-coverage tokenizer keeps all 21,912

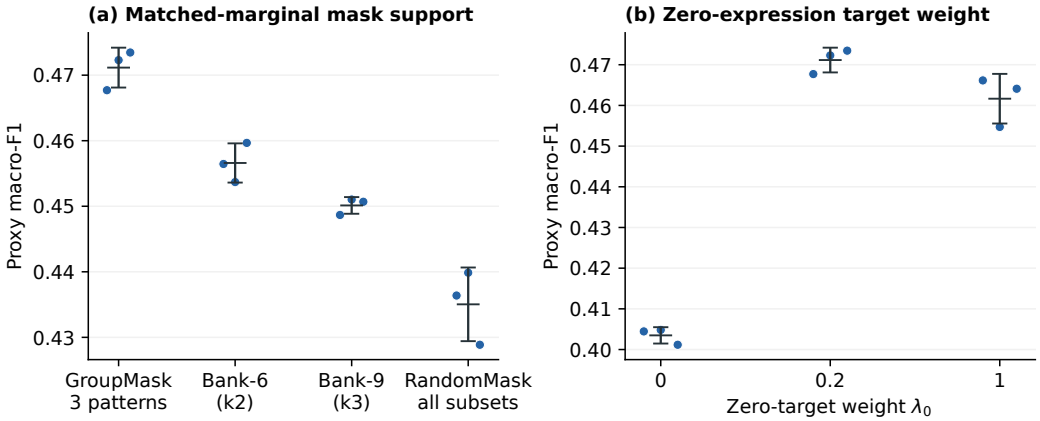


Figure 4: Three-seed proxy experiments at a fixed two-pass endpoint. Dots are training seeds; black horizontal markers and bars are mean ± sample SD. (a) Matched per-block marginal $1/3$ and B48/M16 across three-, six-, and nine-pattern banks and RandomMask. RandomMask has $\binom{48}{16}$ possible subsets; category spacing does not represent log-support distance. (b) Zero-target weights, with GroupMask fixed; weight 0.2 is highest among the three values.

genes—including zero-expression genes—inside the reconstruction objective at a constant 48-token input. GroupMask leads five of the seven task means and improves perturbation retrieval on all three sources; the scaled model leads three complete-coverage task columns and exceeds CellVQ on five with substantially fewer parameters. Its compact input also yields a large throughput and memory advantage in our single-GPU benchmark. A compact block autoencoder can therefore be a strong multi-task representation model when its pretraining task is designed deliberately.

Two design elements plausibly contribute. Recurring target sets improve perturbation retrieval over RandomMask, and the three-seed support experiment favors the smallest tested bank. Reduced zero-target weighting yields a higher observed proxy than either excluding zero targets from the loss or weighting them equally with positive targets. Expression-informed grouping gives a modest gain under RandomMask, suggesting that block partitions informed by biological priors may be a promising direction for future work, while block-count probes suggest a broad local operating range.

Each large model uses one pretraining seed, and the validation-selected masking checkpoints differ in training exposure despite a shared total budget. External-model donor and cell overlap remain unknown. Finally, perturbation retrieval evaluates observed representations, not unseen-response prediction. Extending reconstruction pretraining toward generative objectives through decoder-side architectural changes remains a promising direction. Within these limits, TotiCell offers a simple, scalable masking design with strong transfer performance and compact inference cost.

Reproducibility Statement

The appendix specifies model dimensions, masking and seed construction, downstream splits, aggregation, and the inference-benchmark protocol, and the main text specifies input preprocessing. All available primary setting-level scores and coverage are reported in the appendix; exact method identifiers for the frozen checkpoints appear in Appendix A.

Ethics Statement

This study analyzes existing transcriptomic datasets and introduces no prospective recruitment or clinical intervention. The representations and disease-state benchmarks are research tools, not clinically validated diagnostic systems. Their use for patient-level decisions would require independent validation, assessment of population-specific performance, and appropriate oversight. Access to

and reuse of the underlying datasets remain subject to their original terms; derived representations should be handled with the privacy protections appropriate to human molecular data.

Statement on the use of AI

AI tools, including OpenAI Codex and ZCode, were used for the following. Writing: drafting and revising manuscript text, including restructuring sections, editing captions, and preparing the LaTeX and plotting code that renders results computed by the authors. Retrieval: literature lookup and citation checking. The authors directed all of this work, independently verified every reported number against the recorded results, and take full responsibility for the scientific content. AI tools are not authors. No synthetic biological data or mathematical proofs were generated by AI.

References

Britt Adamson, Thomas M. Norman, Marco Jost, Min Y. Cho, et al. A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell*, 167:1867–1882.e21, 2016. doi: 10.1016/j.cell.2016.11.048.

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*, 21:1470–1480, 2024. doi: 10.1038/s41592-024-02201-0.

Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, et al. Perturb-Seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell*, 167:1853–1866.e17, 2016. doi: 10.1016/j.cell.2016.11.038.

Alon Hacohen, Joseph Bingham, Binyamin Perets, and Dvir Aran. Genes are not words: Dependency-aware masking for single-cell foundation models. In *Gen2 Workshop at ICLR*, 2026. URL https://openreview.net/pdf?id=QxAfyuHQJ6.

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*, 21:1481–1491, 2024. doi: 10.1038/s41592-024-02305-7.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/He_Masked_Autoencoders_Are_Scalable_Vision_Learners_CVPR_2022_paper.html.

Luni Hu, Hua Qin, Yilin Zhang, Yi Lu, et al. RegFormer: a single-cell foundation model powered by gene regulatory hierarchies. *Nature Communications*, 17:6432, 2026. doi: 10.1038/s41467-026-72198-x.

Kasia Z. Kedzierska, Lorin Crawford, Ava P. Amini, and Alex X. Lu. Zero-shot evaluation reveals limitations of single-cell foundation models. *Genome Biology*, 26:101, 2025. doi: 10.1186/s13059-025-03574-x.

Kian Kenyon-Dean, Alina Selega, Ihab Bendidi, Jordan M. Sorokin, Luca Bertinetto, David Errington, Hayley Donnella, and Oren Kraus. Effective biological representation learning by masking gene expression. *arXiv preprint arXiv:2605.31562*, 2026. doi: 10.48550/arXiv.2605.31562. URL https://arxiv.org/abs/2605.31562.

Kian Hong Kock, Le Min Tan, Kyung Yeon Han, Yoshinari Ando, Damita Jevapatarakul, Ankita Chatterjee, Quy Xiao Xuan Lin, Eliora Violain Buyamin, Radhika Sonthalia, Deepa Rajagopalan, et al. Asian diversity in human immune cells. *Cell*, 188(8): 2288–2306.e24, 2025. doi: 10.1016/j.cell.2025.02.017. AIDA (Asian Immune Diversity Atlas) Phase 1 Data Freeze v2; the 200,000-cell subset uses CELLxGENE dataset version f89a12c2-7a3b-415b-ab87-bbc550fe17f4 from collection ced320a1-29f3-47c1-a735-513c7084d508.

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

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*, 365:786–793, 2019. doi: 10.1126/science.aax4438.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nature Biotechnology*, 42(6):927–935, 2024. doi: 10.1038/s41587-023-01905-6. Published online 17 August 2023.

Lucas Schirmer, Dmitry Velmeshev, Staffan Holmqvist, Max Kaufmann, Sebastian Werneburg, Diane Jung, Stephanie Vistnes, John H. Stockley, Adam Young, Maike Steindel, et al. Neuronal vulnerability and multilineage diversity in multiple sclerosis. *Nature*, 573(7772):75–82, 2019. doi: 10.1038/s41586-019-1404-z. Source of the scGPT multiple-sclerosis benchmark, accessed there through EMBL-EBI Single Cell Expression Atlas experiment E-HCAD-35.

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

Jue Wang, Cheng Tan, Zhangyang Gao, Sida Shao, Shiping Liu, and Stan Z. Li. Illuminating cell states by a comprehensive and interpretable single cell foundation model. *Nature Communications*, 17:4037, 2026. doi: 10.1038/s41467-026-70071-5.

Hongzhi Wen, Wenzhuo Tang, Xinnan Dai, Jiayuan Ding, Wei Jin, Yuying Xie, and Jiliang Tang. CellPLM: Pre-training of cell language model beyond single cells. *bioRxiv*, 2023. doi: 10.1101/2023.10.03.560734.

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*, 26:334, 2025. doi: 10.1186/s13059-025-03781-6.

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

Yuansong Zeng, Jiancong Xie, Ningyuan Shangguan, Zhuoyi Wei, et al. CellFM: a large-scale foundation model pre-trained on transcriptomics of 100 million human cells. *Nature Communications*, 16:4679, 2025. doi: 10.1038/s41467-025-59926-5.

A IMPLEMENTATION AND MODEL PROVENANCE

The primary representations are identified by the frozen checkpoints listed below. Each label is the internal run name recorded during training, followed by what it denotes.

- `ours_plan45_fullgroup_b48_m16_ep3` — GroupMask: 96M parameters, validation-selected at three passes.
- `ours_plan45_uniform_b48_m16_ep4` — RandomMask: 96M parameters, validation-selected at four passes.
- `ours_plan47_v1_pass_4_proxybest` — TotiCell: 227M parameters, final four-pass endpoint.

The `proxybest` suffix marks the endpoint selected by the validation proxy; it does not denote task-specific selection, and all dataset exports use the same representation. The GroupMask policy uses three target sets unless identified as Bank-6 or Bank-9.

The seed-4102 random-balanced B48 artifact assigns each of 21,912 genes to exactly one block: 24 blocks of 457 genes and 24 of 456. One padding slot in smaller blocks is excluded from the loss. The seed-4301 target-set manifest contains three disjoint size-16 groups covering all 48 blocks. Loader validation rejects repeated block IDs, incomplete coverage, and duplicate sets. Both evaluated GroupMask checkpoints use grouped masking throughout the four-pass budget; the masking schedule never switches policy within the training horizon. Masking used for reconstruction validation is separate from the training policy and does not affect the downstream readout, which is computed with all tokens visible.

The per-microbatch base mask seed is $s + ta + a_i$, where s is the training seed, t the optimizer step, a the accumulation factor, and a_i the accumulation index. Cell row r uses the base seed plus r. Adjacent microbatches consequently reuse seed values for shifted rows. The formal marginal equalities describe the policy distribution, while the realized temporal sequence is not IID. The same seed construction applies to the masking comparisons; the supplementary experiments do not silently change it.

Table 4: Model and training configurations. GroupMask and RandomMask share architecture and a four-pass training budget; their validation-selected checkpoints differ in exposure. Small-model B36/B72 variants change tokenizer and positional parameter counts as listed in Appendix F.

	GroupMask	RandomMask	TotiCell	Small probe
Encoder width/depth	768/12	768/12	1024/16	512/8
Total parameters (M)	96.45	96.45	227.36	30.65
Encoder parameters (M)	88.86	88.86	207.66	27.23
Training cells (M)	50.47	50.47	50.47	5.00
Budget / evaluated pass	4/3	4/4	4/4	2/2
Physical batch / accumulation	768/1	768/1	384/2	512/2
Training seed(s)	42	42	42	42,43,44

Inference audit and parameter provenance. The efficiency table reports serial measurements on one physical RTX 3090 (24 GiB), using each model's documented tokenization. Every model ran after the GPU cooled to at most 50°C. Concurrent measurements, including a run overlapping an external process, were excluded. Timings use CUDA synchronization and the median of three repeats. Pure-model latency excludes model loading and disk I/O; end-to-end throughput includes preprocessing. The best batch succeeded in all repeats under a 22 GiB allocated-memory cap. CellVQ and scFoundation were limited to batch one; TotiCell used batches 64 and 32 at 96M and 227M. This is a fixed architecture-cost measurement on the 16,382-cell Pancreas input, independent of the predictive evaluation. scFoundation end-to-end throughput has a coefficient of variation of 23.16%, versus 1.47% for pure-model timing; the prespecified three-repeat median includes the outlying repeat.

TotiCell parameter counts were checked against checkpoint tensors. CellVQ's 511M and scFoundation/Geneformer counts follow their published models; CellPLM uses its 85M checkpoint. The

scGPT whole-human checkpoint contains 51,330,049 parameters. Encoder input lengths were audited on the same Pancreas cells: TotiCell always uses 48 tokens; CellVQ and scFoundation average 3,206.72 and 3,206.17, with median 2,504.5 and 90th percentile 6,257. These counts are not FLOP estimates or evidence of equal pretraining exposure.

B TASK DEFINITIONS AND EVALUATION SCOPE

The primary suite comprises 22 settings from 11 sources. Table 5 specifies the exact allocation. The sources include the scIB multi-study Pancreas benchmark (Luecken et al., 2022), AIDA Phase 1 Data Freeze v2 (Kock et al., 2025), processed MS data (Schirmer et al., 2019; Cui et al., 2024), CELLxGENE atlas subsets, and GEARS-processed Perturb-seq data (Dixit et al., 2016; Adamson et al., 2016; Norman et al., 2019; Roohani et al., 2024).

Table 5: Selected tasks and datasets. An OOD mode or mapping direction is one setting; each few-shot curve is summarized once by AULC.

Task	Settings	Evaluation sources
Annotation	5	HYPOMAP, HIHA, SLE, Pancreas, LBCL
Integration	6	HYPOMAP, SLE, HIHA, AIDA v2, Pancreas, Sound Life
Few-shot	2	HYPOMAP, AIDA v2
Disease state	2	SLE, MS
Perturbation retrieval	3	Dixit, Norman, Adamson
Cross-atlas mapping	2	AIDA v2 to HIHA; HIHA to AIDA v2
OOD detection	2	HIHA Easy; HIHA Hard

Annotation and splits. Logistic regression is fitted to standardized frozen embeddings. Validation macro-F1 selects C from $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$, followed by refitting on training plus validation. Atlas annotation uses five donor-group folds. HIHA uses donor-disjoint train, validation, and test folds. Pancreas uses leave-one-technology-out evaluation over eight platforms: each sorted platform is held out for testing, the next for validation, and the rest for training. Frozen encoders still use downstream labels through the supervised probe.

Integration. At most 20,000 cells are evaluated per dataset using $0.6\,\mathrm{avg}_{\mathrm{bio}} + 0.4\,\mathrm{avg}_{\mathrm{batch}}$ (Luecken et al., 2022). Biological conservation averages NMI, ARI, cell-type silhouette, isolated-label silhouette, and graph connectivity; batch mixing averages batch silhouette, iLISI, and kBET. Unavailable components use a NaN-ignoring mean. The graph uses 30 neighbors and seed 4302. These components define the reported composite; it is not interchangeable with every score described as scIB. Integration is a single target-dataset evaluation without replicate uncertainty.

Few-shot and disease classification. Few-shot uses 1, 5, 10, 20, 50, and 100 labels per class, ten sampling seeds, and five folds. Its AULC is the trapezoidal integral of macro-F1 against natural-log shots, divided by $\log(100) - \log(1)$. Sampling seeds vary labeled cells, not pretrained models. Disease classification first averages embeddings within donor–cell-type–condition groups. Cell types must meet a minimum of eight donors per condition across the recorded splits. SLE compares SLE with normal samples; MS uses the recorded MS/control donor protocol. The primary score is macro-AUROC, aggregated over eligible cell types and folds.

Cross-atlas mapping and OOD. Mapping requires at least 500 reference cells and 100 query cells per shared Cell Ontology label, with at least two donors in each atlas. Query labels define a closed-set benchmark; reference train/validation partitions vary across five folds while the external query is reused. Scores apply to the eligible-label subset, not every query cell. Both mapping directions use HIHA and AIDA. Export coverage is checked independently of ontology eligibility.

OOD standardizes with known-class training cells and scores distance to the nearest known-class centroid. Easy and Hard unknown classes come from the source validation/test ontology branches. HIHA's branch specifications and seeds are preserved, and their minimum unknown-cell and donor support requirements are rechecked. AUROC is the primary metric. No graph threshold, branch choice, or endpoint is revised according to the direction of the results.

Perturbation retrieval. Each condition is split into three pseudo-replicates with seed 4301; perturbations with at least 100 cells are retained. A perturbation vector is its mean embedding minus the corresponding control mean. One pseudo-replicate is queried against the mean of the other two; cosine similarity ranks perturbation identities. MRR averages identities and then the three query splits. These artificial splits share an experimental source and do not provide independent biological replication or evaluate generated responses to unseen perturbations. The supplied Adamson and Pancreas files lack trustworthy raw counts; the documented normalized-expression input is retained where required.

C COMPARATOR COVERAGE AND MISSING RESULTS

All TotiCell primary models use mean pooling. CellFM uses nonzero-gene-token mean; other encoders retain documented readouts. Geneformer uses reconstructed normalized-log1p input where native-count input is unavailable: Pancreas annotation and integration, MS disease classification, and Adamson perturbation retrieval. These are input-adapted comparisons, not evaluations on native counts. PCA-HVG uses 2,000 genes selected by training-expression variance and 256 principal components (seed 4301; at most 50,000 fitting cells). For supervised tasks, gene selection, scaling, and PCA are fitted only on the outer training fold; OOD fits known-class training cells only, and mapping fits reference training cells over shared genes. Integration and perturbation retrieval instead use PCA fitted on the target expression population. All reported PCA results follow these fitting rules; no hyperparameters were selected from the reported scores.

Table 6 reports available/required setting counts. Task means average the available settings equally; incomplete means carry † in Table 1, and a dash denotes no available score. CellVQ, scFoundation, scGPT, Geneformer, CellFM, and PCA-HVG have complete coverage. PCA covers all 22 settings, including both mapping directions and both HIHA OOD modes. CellPLM has four input/coverage gaps: Pancreas annotation and integration, MS disease classification, and Adamson perturbation retrieval. Partial means describe different setting subsets and are not directly comparable with complete means; they enter neither best-score highlighting nor the complete-roster rank.

Table 6: Available settings divided by required settings in the selected suite. Incomplete task means in Table 1 use exactly these available settings.

Model	Ann.	Int.	Few-shot	Disease	Perturb.	Mapping	OOD
GroupMask	5/5	6/6	2/2	2/2	3/3	2/2	2/2
RandomMask	5/5	6/6	2/2	2/2	3/3	2/2	2/2
TotiCell	5/5	6/6	2/2	2/2	3/3	2/2	2/2
CellVQ	5/5	6/6	2/2	2/2	3/3	2/2	2/2
scFoundation	5/5	6/6	2/2	2/2	3/3	2/2	2/2
scGPT	5/5	6/6	2/2	2/2	3/3	2/2	2/2
Geneformer	5/5	6/6	2/2	2/2	3/3	2/2	2/2
CellFM	5/5	6/6	2/2	2/2	3/3	2/2	2/2
CellPLM	4/5	5/6	2/2	1/2	2/3	2/2	2/2
PCA-HVG	5/5	6/6	2/2	2/2	3/3	2/2	2/2

Ranking. Within each task, the nine methods with all 22 scores are ranked by task mean, including PCA-HVG. Rank r becomes $100(r-1)/(N-1)$, with $N = 9$ and averaged ties; the seven task percentiles are then averaged. The means are 14.29 for TotiCell, 21.43 for GroupMask, 25.00 for RandomMask, 46.43 for scFoundation, 48.21 for CellVQ, 58.93 for Geneformer, 71.43 for scGPT, 73.21 for PCA-HVG, and 91.07 for CellFM. These are roster-dependent summaries, not biological effect sizes.

D ALL SELECTED SETTING SCORES

Table 7 retains the four central representations at every selected setting. Tables 8–9 extend the display to the full comparator roster. Figure 5 includes positive and negative differences. Means are conditional on frozen checkpoints; the main-scale models do not have independent pretraining-seed

replication. Mapping folds reuse the query atlas, and few-shot sampling seeds are nested within folds. The tables therefore avoid treating rows or pseudo-replicates as independent biological uncertainty units.

Table 7: All 22 primary settings. Model names follow Section 4.1; Δ is GroupMask minus RandomMask on the original metric scale. A dash is missing, not zero.

Task	Dataset / setting	GroupMask	RandomMask	TotiCell	CellVQ	Δ(GroupMask–RandomMask)
Annotation	HYPOMAP	0.946501	0.942875	0.944874	0.935523	+0.003626
Annotation	HIHA	0.894751	0.894257	0.898582	0.875297	+0.000494
Annotation	SLE	0.970081	0.969864	0.970494	0.971130	+0.000216
Annotation	Pancreas	0.752765	0.752553	0.764128	0.713205	+0.000212
Annotation	LBCL	0.861608	0.863721	0.868058	0.841494	-0.002113
Integration	HYPOMAP	0.593384	0.577486	0.580678	0.501475	+0.015898
Integration	SLE	0.618742	0.612594	0.608892	0.637701	+0.006148
Integration	HIHA	0.664684	0.659744	0.659338	0.588854	+0.004940
Integration	AIDA	0.658711	0.653815	0.642599	0.684910	+0.004896
Integration	Pancreas	0.372197	0.367406	0.355339	0.492041	+0.004791
Integration	Sound Life	0.661108	0.663536	0.654867	0.606822	-0.002428
Few-shot	HYPOMAP	0.861422	0.867100	0.856940	0.775366	-0.005678
Few-shot	AIDA	0.739738	0.746113	0.754469	0.803026	-0.006374
Disease	SLE	0.923357	0.926817	0.927880	0.942742	-0.003460
Disease	MS	0.786667	0.791667	0.818333	0.835000	-0.005000
Perturbation	Dixit	0.600724	0.541124	0.645433	0.548308	+0.059600
Perturbation	Norman	0.861963	0.837889	0.862806	0.815066	+0.024074
Perturbation	Adamson	0.909874	0.905740	0.922202	0.821700	+0.004134
Mapping	AIDA to HIHA	0.745628	0.738010	0.753911	0.485618	+0.007618
Mapping	HIHA to AIDA	0.886469	0.889306	0.907581	0.884049	-0.002837
OOD	HIHA / Hard	0.932856	0.931605	0.923997	0.895002	+0.001251
OOD	HIHA / Easy	0.936791	0.938039	0.933715	0.865461	-0.001248

Table 8: Selected setting scores for TotiCell configurations, CellVQ, and scFoundation. Missing values are retained explicitly.

Setting	GroupMask	RandomMask	TotiCell	CellVQ	scFoundation
Annotation / HYPOMAP	0.9465	0.9429	0.9449	0.9355	0.9341
Annotation / HIHA	0.8948	0.8943	0.8986	0.8753	0.8725
Annotation / SLE	0.9701	0.9699	0.9705	0.9711	0.9690
Annotation / Pancreas	0.7528	0.7526	0.7641	0.7132	0.7927
Annotation / LBCL	0.8616	0.8637	0.8681	0.8415	0.8428
Integration / HYPOMAP	0.5934	0.5775	0.5807	0.5015	0.6088
Integration / SLE	0.6187	0.6126	0.6089	0.6377	0.6277
Integration / HIHA	0.6647	0.6597	0.6593	0.5889	0.6342
Integration / AIDA	0.6587	0.6538	0.6426	0.6849	0.6512
Integration / Pancreas	0.3722	0.3674	0.3553	0.4920	0.4182
Integration / Sound Life	0.6611	0.6635	0.6549	0.6068	0.6391
Few-shot / HYPOMAP	0.8614	0.8671	0.8569	0.7754	0.8503
Few-shot / AIDA	0.7397	0.7461	0.7545	0.8030	0.7024
Disease / SLE	0.9234	0.9268	0.9279	0.9427	0.9258
Disease / MS	0.7867	0.7917	0.8183	0.8350	0.7783
Perturbation / Dixit	0.6007	0.5411	0.6454	0.5483	0.5200
Perturbation / Norman	0.8620	0.8379	0.8628	0.8151	0.8415
Perturbation / Adamson	0.9099	0.9057	0.9222	0.8217	0.8857
Mapping / AIDA to HIHA	0.7456	0.7380	0.7539	0.4856	0.6499
Mapping / HIHA to AIDA	0.8865	0.8893	0.9076	0.8840	0.8182
OOD / HIHA hard	0.9329	0.9316	0.9240	0.8950	0.8959
OOD / HIHA easy	0.9368	0.9380	0.9337	0.8655	0.9137

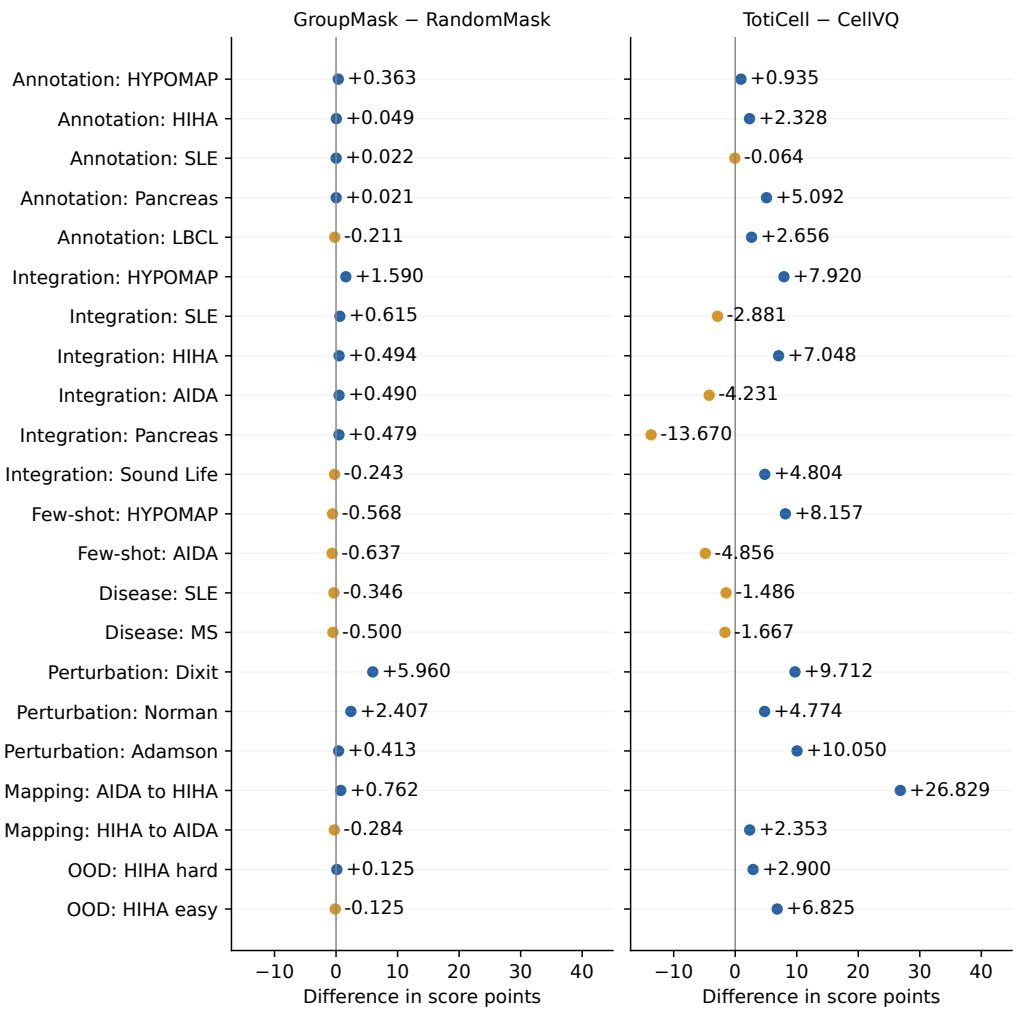


Figure 5: All selected-setting differences in score points. Left: GroupMask minus RandomMask. Right: TotiCell minus CellVQ. Both panels use the same numerical scale. A positive point favors the first model. The tiny mean OOD difference arises from opposite Easy/Hard directions. These descriptive comparisons use different training endpoints or model scales.

Table 9: Selected setting scores for the remaining comparators. PCA-HVG covers all 22 settings; CellPLM retains its coverage gaps.

Setting	scGPT	Geneformer	CellFM	CellPLM	PCA-HVG
Annotation / HYPOMAP	0.9186	0.9066	0.8845	0.9194	0.8887
Annotation / HIHA	0.8087	0.8534	0.7135	0.8392	0.7964
Annotation / SLE	0.9632	0.9639	0.9028	0.9639	0.9487
Annotation / Pancreas	0.6466	0.7130	0.6592	–	0.7429
Annotation / LBCL	0.7991	0.8344	0.7692	0.8189	0.8402
Integration / HYPOMAP	0.5860	0.5399	0.5371	0.5891	0.5655
Integration / SLE	0.6204	0.5657	0.4259	0.6114	0.5697
Integration / HIHA	0.5984	0.6112	0.5094	0.6137	0.5552
Integration / AIDA	0.6115	0.6138	0.5340	0.6263	0.6045
Integration / Pancreas	0.4538	0.4150	0.4080	–	0.4505
Integration / Sound Life	0.5956	0.6032	0.5109	0.6235	0.5355
Few-shot / HYPOMAP	0.8165	0.7763	0.7207	0.7401	0.6126
Few-shot / AIDA	0.6109	0.6810	0.4406	0.6246	0.4992
Disease / SLE	0.9056	0.9186	0.8148	0.9276	0.8691
Disease / MS	0.8150	0.8133	0.7983	–	0.7317
Perturbation / Dixit	0.5559	0.5066	0.4219	0.5428	0.5694
Perturbation / Norman	0.7749	0.8141	0.8143	0.7249	0.9006
Perturbation / Adamson	0.8460	0.8509	0.8964	–	0.9441
Mapping / AIDA to HIHA	0.6117	0.7248	0.7329	0.4450	0.7103
Mapping / HIHA to AIDA	0.8164	0.8529	0.7799	0.7646	0.8187
OOD / HIHA hard	0.6872	0.7507	0.5433	0.7686	0.5519
OOD / HIHA easy	0.8511	0.7918	0.4469	0.8643	0.6020

E EXPRESSION-INFORMED GENE PARTITIONS

Random gene blocks deliberately avoid reliance on a curated ontology or pathway resource. They also avoid estimating a tissue-dependent graph before training. This simplicity is a design choice rather than evidence that biological organization has no value. We test it with a CorrMask-inspired *static partition* experiment: it uses an expression-derived dependence graph, not GO/pathway annotations and not a reproduction of CorrMask's dynamic masking policy (Hacohen et al., 2026).

A fixed sample of 100,000 cells is drawn without replacement, proportionally across shards, from the five-million-cell training split only (seed 5200). A rank-weighted profile of the top 2,048 expressed genes nominates up to 800,000 candidate edges through a sparse sketch. Positive correlations are retained with minimum co-presence 200, minimum gene presence 50, $\rho \geq 0.2$, and at most 50 neighbors per gene. Validation cells and labels are excluded from graph construction. Under the same capacity and detection-balance constraints, *Grouped* maximizes within-block positive edge weight and *Dispersed* minimizes it, using greedy assignment and 30,000 swaps with fixed construction seeds 5201 and 5202. Blocks are frozen before training and cover all 21,912 genes once.

Only the gene-to-block partition changes relative to the reused RandomMask B48 baseline. All three conditions use 30,653,641 parameters, B48/M16, two passes, the same training/evaluation recipe, and seeds 42–44. Grouped improves the proxy by 0.006822 over random on average and by 0.005177 over Dispersed; Grouped exceeds both on every seed (Figure 6). Dispersed is only 0.001645 above random on average, with mixed seed directions. Thus, expression-derived grouping helps modestly under RandomMask, but is not necessary to obtain the stronger GroupMask proxy result on random blocks. Because graph-based blocks were not combined with GroupMask, the experiment does not estimate that interaction or establish random partitions as globally optimal.

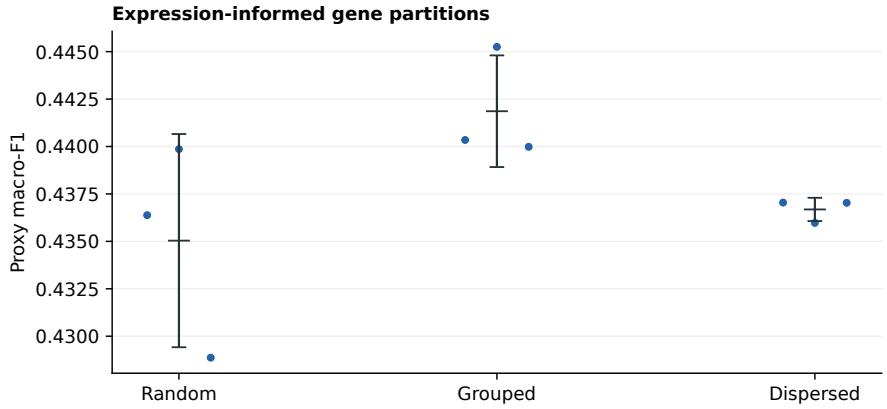


Figure 6: Random versus expression-informed gene-to-block partitions under the same Random-Mask policy. Each small point is one training seed; black horizontal markers and error bars show mean and sample SD. Grouped places positively related genes together; Dispersed separates them. This is a static expression-graph comparison, not a curated biological-prior ablation.

F BLOCK-COUNT SENSITIVITY

We compare B36/M12, B48/M16, and B72/M24 while retaining three exhaustive target sets and marginal masking probability $1/3$. Gene partitions use the same seed-4102 balanced-permutation construction. B36 has 24 blocks of 609 genes and 12 of 608; B72 has 24 blocks of 305 genes and 48 of 304. Target groups use seed 4301. Each condition uses seeds 42–44, five million training cells, and a fixed two-pass endpoint; B48 reuses the baseline.

The means are 0.473832 for B36, 0.471152 for B48, and 0.472337 for B72 (Figure 7). B36 exceeds B48 by 0.002680 and B72 by 0.001495 in mean, but B36–B72 differences change sign across seeds. These differences are small relative to seed variability. The models have 30,994,785, 30,653,641, and 30,321,713 parameters, and 24, 32, and 48 visible training tokens. Thus this is a configuration-level sensitivity analysis, not an equal-parameter or equal-FLOP test of the tokenizer bottleneck. It supports a locally broad operating range and does not identify B48 as the best setting.

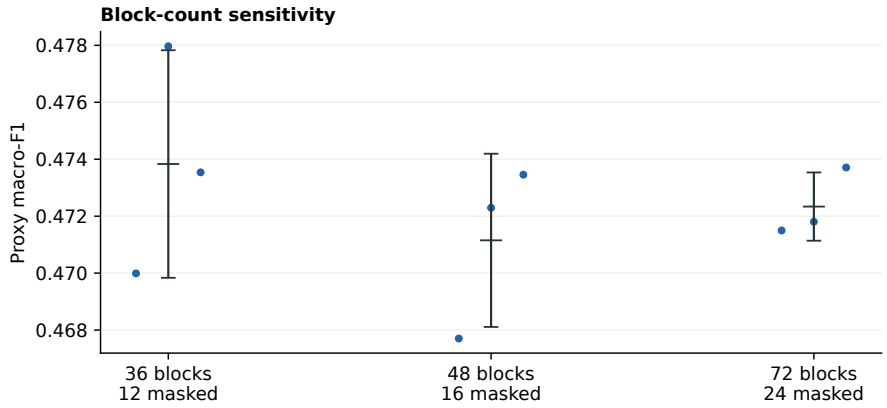


Figure 7: Block-count sensitivity with GroupMask and fixed masked fraction $1/3$. Points are training seeds; black horizontal markers and bars are mean and sample SD. Changing block count changes gene capacity per block, tokenizer parameter count, and token-processing cost.

G MASK-SUPPORT CONSTRUCTION AND PROXY RESULTS

Bank-6 and Bank-9 expand a common master bank by adding complete three-set partitions. The first partition is the GroupMask partition; coverage depth k selects the first k partitions, for $3k$ target sets. Each block occurs in exactly k sets, so a uniform draw gives marginal $k/(3k) = 1/3$. Mask sizes, visible-token budget, training steps, loss, and evaluation seeds are unchanged. Bank-6 and Bank-9 denote six and nine available target sets; each cell still masks exactly 16 blocks.

The GroupMask and RandomMask reference rows are shared baseline runs. Table 10 lists the sample means and SDs. A decline across the three tested support sizes is limited evidence about these constructions, not proof of a universal monotonic entropy-performance relationship.

Table 10: Proxy comparisons at the fixed pass-2 endpoint. Each row contains three training seeds. Sample SD uses denominator $n - 1$; repeated reference rows reuse the same runs.

Axis	Condition	Mean	Sample SD
Mask support	GroupMask 3 patterns	0.471152	0.003042
Mask support	Bank-6 (k2)	0.456603	0.002991
Mask support	Bank-9 (k3)	0.450127	0.001272
Mask support	RandomMask all subsets	0.435037	0.005619
Gene partition	Random	0.435037	0.005619
Gene partition	Grouped	0.441859	0.002943
Gene partition	Dispersed	0.436683	0.000614
Block count	36 blocks 12 masked	0.473832	0.003997
Block count	48 blocks 16 masked	0.471152	0.003042
Block count	72 blocks 24 masked	0.472337	0.001199

H EXISTING ROBUSTNESS PROBES

These probes draw on a proxy study with 33 training runs and four milestones (0.5, 1, 1.5, and 2 passes). All reported comparisons here use the fixed pass-2 endpoint. Random block-partition seeds 4101/4102/4103 yield means 0.473006/0.471152/0.475752; changing the fixed three-set partition P1/P2/P3 yields 0.471152/0.470779/0.469813 (Figure 8). These latter partitions each have three patterns and must not be confused with expanding a bank to six or nine patterns.

For the zero-target comparison, weights 0, 0.2, and 1 yield proxy means 0.403483, 0.471152, and 0.461652, with sample SDs 0.002008, 0.003042, and 0.006099. Thus 0.2 has the highest observed mean among these three values. All three conditions retain zero-expression genes in the input; weight 0 excludes their reconstruction errors only. Close endpoint means are evidence of limited observed sensitivity, not statistical equivalence.

Policies that keep a fixed set of blocks permanently visible or permanently masked would change the single-block marginals. The mask-support experiment instead holds every marginal at $1/3$ while expanding the available target sets.

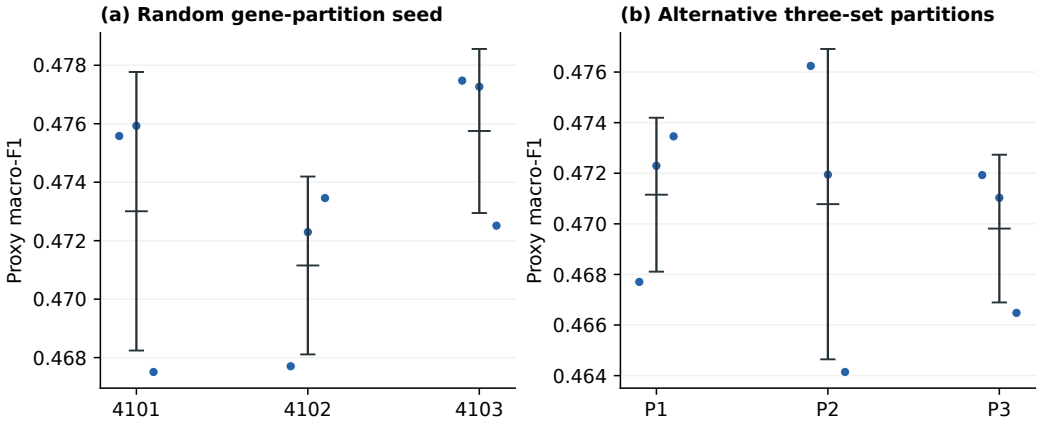


Figure 8: Existing random-partition and target-grouping robustness probes. Each panel changes one fixed construction while preserving GroupMask with three target sets. Dots are training seeds and bars are sample SD. These are not independent repeats of a broad random partition population.

I JOINT MASKING STRUCTURE

For RandomMask, distinct block pairs have co-masking probability $16 \cdot 15/(48 \cdot 47) \approx 0.1064$. For GroupMask, pairs in one target set have probability $1/3$, and pairs in different target sets have probability zero. The diagonal is $1/3$ in both. Figure 9 depicts these analytical probabilities, with block indices reordered only for display; it is not an empirical gene-correlation matrix. RandomMask has finite support $\binom{48}{16}$, rather than mathematically infinite support.

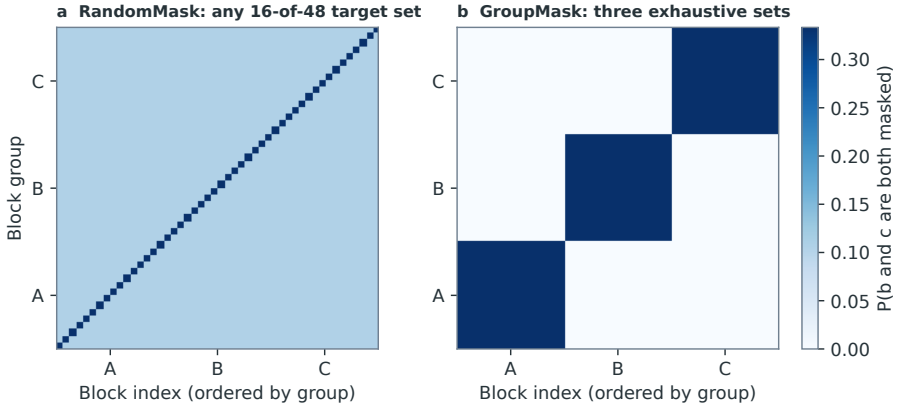


Figure 9: Exact co-masking probabilities for the two B48/M16 policies. Equal diagonals encode matched marginals; off-diagonal structure distinguishes recurring GroupMask targets from RandomMask.

J FEW-SHOT CURVES ON THE SELECTED DATASETS

Figure 10 shows HYPOMAP and AIDA at all six label budgets for GroupMask, RandomMask, TotiCell, and CellVQ. Shading is one sample SD across five fold means, after averaging the ten label-sampling seeds within each fold; it does not capture pretraining-seed uncertainty. Both curves favor RandomMask over GroupMask in AULC. The main table averages these two curves.

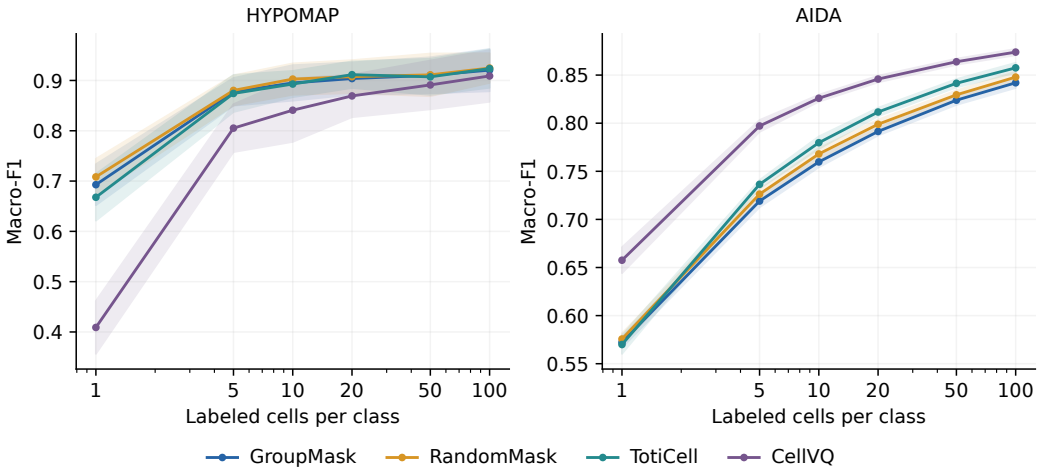


Figure 10: Few-shot transfer on HYPOMAP and AIDA. Every tested label budget is shown; lines are fold means and shading is sample SD across folds after seed averaging. The encoders and readout remain frozen.