SIGCELL: SKETCHED ISOTROPIC GAUSSIANIZATION OF CELL TOKENS FOR TRANSFERABLE CYTOMETRY SET ENCODER

Anonymous authors
Paper under double-blind review

ABSTRACT

Cytometry measures dozens of protein markers on hundreds of thousands of single cells in a blood sample and is a standard tool for profiling the human immune system. However, even with the same marker panel, cytometry data varies greatly across institutions, so a model trained at one institution does not transfer smoothly to another. Existing pretrained cytometry encoders have largely been evaluated within the data they were trained on, or fine-tuned anew with labels for each downstream dataset, so whether a frozen encoder transfers to another institution remains open. We introduce a cell-SIGReg loss term that improves the transferability of set encoders, and present SIGCell (Sketched Isotropic Gaussianized Cell set encoder), a transferable cytometry set encoder trained with it. We identify a sample-context bias in masked set encoders: a cell's token depends strongly on which sample it is encoded with, which is associated with poor transfer of source-fitted cell-type read-outs across samples and institutions. Cell-SIGReg applies the sketched isotropic Gaussian regularizer to the cell tokens pooled over a mini-batch of samples, penalizing the separating structure between samples. We benchmark SIGCell against other pretrained cytometry encoders and classical baseline algorithms. On external datasets, SIGCell improves COVID-19 severity grading (balanced accuracy .501, against at most .441 for the other frozen representations) and transfers a source-fitted COVID-19 detection rule across institutions without target-cohort labels (AUROC .949 and .995 on two external datasets, first among ten models). Ablations and token-level analyses show that the gain is associated with a much weaker sample-context bias: the share of cell-token variance that lies between samples falls from .93–.98 to .12–.24. Cell-level latent regularization is thus a simple way to improve cross-institution transfer in cytometry set encoders. Code available: https://anonymous.4open.science/r/SIGCell-C75B

1 INTRODUCTION

Mass cytometry measures protein expression on each of hundreds of thousands of single cells in a tube of blood, resolving the immune cell types of a donor and the states they are in. Such profiles carry clinical signal: in COVID-19 they define a circulating signature of the disease and track patients' immune recovery in intensive care (Geanon et al., 2021; Penttilä et al., 2021), and pretrained encoders can compress a sample into a representation from which diagnosis can be predicted (Lee et al., 2025; Kim et al., 2024; Qadri et al., 2026; Ding et al., 2025; Zhuang et al., 2026). To be reused, such a representation has to survive the differences between institutions. Even with the same commercial marker panel (Bagwell et al., 2020), our three institutions differ enough that a classifier tells their samples apart perfectly, and the differences remain among healthy donors alone and within single gated cell types, so they are not explained by disease or by which cells were drawn (Appendix C). If a representation does not carry across such batch effects, every new site must collect and label its own cohort, and batch-correction methods need shared controls, batch labels or technical replicates that a new site often lacks (Van Gassen et al., 2020; Pedersen et al., 2022; Trussart et al., 2020). A published encoder is therefore most useful if its frozen representation stays informative, without new labels, on blood drawn and stained at an institution it has never seen.

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.

Whether existing encoders meet this bar is largely untested: they are typically evaluated on their pretraining data or adapted to each downstream dataset using its labels (Section 2). To our knowledge, the only published evaluation applying a cytometry model fitted on one dataset to another without the second dataset's labels is GPCT's, between two mouse datasets on the label they share (Zhuang et al., 2026). However, it was neither validated on human data nor compared with other models. We evaluate cross-institution transfer on human cytometry data in two settings: (i) a linear probe on each frozen encoder, refitted within the target dataset, and (ii) a decision rule fitted at the source institution and transferred with its threshold unchanged, for which SIGCell keeps its encoder frozen and each compared model is adapted in its own regime (Section 5.2).

Under these evaluations, strong in-distribution performance does not necessarily translate to strong external transfer (Section 5.1), and masked set encoders such as MAESTRO (Lee et al., 2025) show a failure mode that follows from their design. Such an encoder computes each cell's token in the context of the other cells of its sample, but its objective reads only the pooled sample embedding, so nothing asks a cell's token to mean the same thing in two samples. From random initialization on, these encoders write the sample into every token: after training, .93–.98 of the cell-token variance lies between samples, against .10–.19 in the raw markers, and re-encoding the same cells among another sample's cells moves their tokens by six to nine times their cell-to-cell spread (the root-mean-square distance of their tokens from their mean). We call this the *sample-context bias* (*sample bias* for short). It differs from the sample-level information a sample embedding is meant to carry: the sample bias sits in the coordinates of the cells themselves, so the same cell receives a different token in each sample it is encoded with. Each sample keeps its cell types apart, but in a frame of its own (its own placement of the token coordinates), so a read-out (a classifier fitted on frozen representations) fitted in one frame is misread in another; a probe refitted on the sample embedding within one dataset does not reveal this.

We propose SIGCell (Sketched Isotropic Gaussianized Cell), a masked set encoder whose cell tokens are trained with a new regularizer, cell-SIGReg. It is inspired by the statistic of SIGReg, the sketched isotropic Gaussian test of LeJEPA (Balestriero & LeCun, 2025), which needs no labels and whose effect depends on the population it is computed over. On an embedding with one point per sample, all the variance it shapes lies between samples, and matching an isotropic Gaussian pushes them apart, which is how it prevents collapse. Cell-SIGReg computes the same statistic over the cell tokens of a mini-batch pooled across its few samples, where it acts the other way: at its isotropic target, a cloud pooled from S samples in d dimensions holds at most $(S-1)/d$ of its variance between samples (Proposition 1), so the test caps how far samples sit apart in the projection it sees without sample labels, while standardized per sample it cannot see the sample bias at all.

Our contributions are as follows; Section 2 places each against prior work.

- **The sample bias, a structural failure of masked set encoders** (Section 5.4): present at initialization, kept by all three objectives we test that train only through the sample embedding, and measurable without labels in any set encoder.
- **Cell-SIGReg: the SIGReg statistic over cell tokens pooled across samples** (Section 3): the same SIGReg statistic pushes samples apart over one embedding per sample and, over cell tokens pooled across a few samples, caps their separation in the tested projection (Proposition 1), while a per-sample version is blind to the sample bias; cell-SIGReg applies it there, without labels, reference samples or changes to the encoder.
- **Frozen transfer with a shared cell frame** (Section 5): only the conditions with the cell term bring the sample bias close to the raw-marker level and let a cell-type read-out fitted at the source work at institutions never seen in training (.13–.19 to .64–.69 balanced accuracy on Geanon). Added to sample-SIGReg, which alone lowers every external read-out below reconstruction alone, the cell term raises external severity grading from .397 to .501; SIGCell is on par with MAESTRO, another masked set encoder trained by distribution reconstruction, in distribution and above it in both transfer settings.

2 Related work

Pretrained encoders for cytometry. Pretrained cytometry encoders differ on two design axes (Table 10). The first is whether a cell is encoded in the context of the other cells of its sample:

MAESTRO, GPCT and EventHorizon let the cells of a sample attend to one another (Lee et al., 2025; Zhuang et al., 2026; Medina Grespan et al., 2026), EventHorizon after first encoding each cell from its own markers, while cyMAE, CytoBERT and ImmuneFM encode each cell from its own markers alone (Kim et al., 2024; Qadri et al., 2026; Ding et al., 2025). The second is whether the sample representation is fixed by pretraining or formed with each task's labels. MAESTRO, EventHorizon and DinoFlow (O'Fallon et al., 2026) pretrain a sample embedding, which MAESTRO probes on held-out samples of its own corpus; cyMAE instead selects a pooling rule per task, GPCT trains a decoder over its frozen encoder, and CytoBERT and ImmuneFM fine-tune the whole encoder. SIGCell sits with MAESTRO on both axes. The second axis sets our roster: frozen linear probing admits only sample representations fixed without labels (Section 5.1), and source-fitted transfer adapts each model in its own regime (Section 5.2). EventHorizon and DinoFlow had released no code at the time of writing, and ImmuneFM's paper links none, so we could not confirm its public repository as the authors' reference implementation; none of the three is compared. We also compare against four training-free cytometry representations, CKME (Shan et al., 2022), k-means, FlowSOM (Van Gassen et al., 2015) and manual gating; generic set architectures, Deep Sets (Zaheer et al., 2017) and Set Transformer (Lee et al., 2019); and supervised cytometry models, CellCnn (Arvaniti & Claassen, 2017), CytoSet (Yi & Stanley, 2021) and CytoDx (Hu et al., 2019). To our knowledge, no prior work compares these approaches on one corpus under one donor-grouped protocol.

Controlling the latent space. Many self-supervised representation objectives require explicit or implicit mechanisms to prevent collapse, and these mechanisms also shape the geometry of the learned embedding (Bardes et al., 2022; Zbontar et al., 2021; Ermolov et al., 2021; Caron et al., 2021; Oquab et al., 2023; Sablayrolles et al., 2019). SIGReg was introduced with LeJEPA (Balestriero & LeCun, 2025) and is used, there and in later applications, primarily as an anti-collapse regularizer (Rusnak et al., 2026; Lee & Chen, 2026; Zhang et al., 2026; Letellier et al., 2026). Cell-SIGReg is inspired by this statistic but computes it over intermediate cell tokens pooled across samples, where it suppresses sample-specific structure. To our knowledge, SIGReg has not previously been studied in cytometry or for cross-institution transfer.

Batch effects and site signatures. Cytometry batch-correction methods typically use information linking or identifying batches, such as shared controls (Van Gassen et al., 2020), batch labels (Pedersen et al., 2022; Korsunsky et al., 2019), or technical replicates (Trussart et al., 2020). Cell-SIGReg uses none of these and does not explicitly align cohorts or samples; it targets the sample-specific structure the encoder introduces and leaves differences present in the measurements (Section 5.4). Site-specific signatures in learned representations are also documented in histopathology (Howard et al., 2021); our target is sample-dependent distortion of cell tokens, not site identity in the data.

3 SIGCELL

3.1 BACKBONE: A MASKED SET ENCODER

A sample is an unordered set of cells $X = \{x_1, \ldots, x_N\}$, where each $x_i \in \mathbb{R}^{30}$ contains the expression levels of 30 protein markers. The encoder lifts each cell to 384 dimensions, passes the set through five Induced Set-Attention Blocks (ISABs) with 16 inducing points, and pools the resulting cell tokens into a single 256-dimensional sample representation z using one-seed Pooling by Multihead Attention (PMA) followed by an affine head (Lee et al., 2019). Encoder attention, including the pooling layer, uses α-entmax (Peters et al., 2019); Appendix A defines the blocks and decoder in full. With a fixed number of inducing points, the encoder scales linearly in N, allowing hundreds of thousands of cells from one sample to be encoded jointly.

The encoder is pretrained by masked-set reconstruction. At each step, we draw a random subset of 10,000 cells and sample a masking fraction r from $\{0, .2, .4, .6, .8\}$. Inspired by MAESTRO's released code (Lee et al., 2025), we order the subset along its first principal component and mask the $10{,}000\,r$ cells at one end, chosen at random, so that a phenotypically coherent block is hidden rather than scattered cells. We encode only the visible cells and ask a decoder with two Self-Attention Blocks (SABs) to reconstruct the entire subset from z alone using a permutation-invariant distance between the generated and observed sets (Appendix A). Because the decoder must also reconstruct cells hidden from the encoder, z must capture information about the sample-level cell distribution

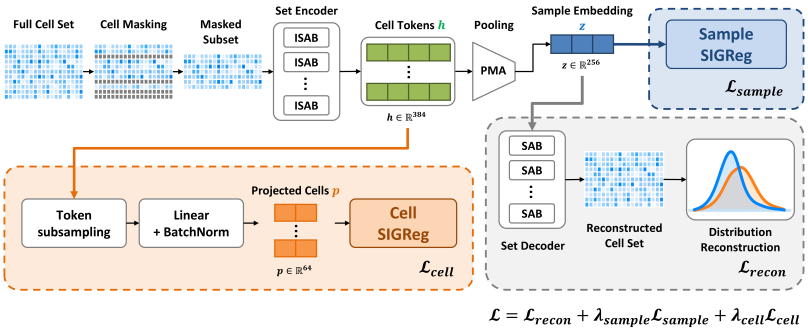

Figure 1: **SIGCell architecture.** A training step draws a subset of the sample's cells, hides a fraction of them, and encodes only the ones left. The encoder emits one token h_i per encoded cell and pools them into the sample embedding z; its blocks are defined in Appendix A. *Reconstruction* (gray) regenerates the *whole* subset, hidden cells included, from z alone. *Sample-SIGReg* (blue) acts on the embedding a probe reads; *cell-SIGReg* (orange), our contribution, acts on a projection of the cell tokens discarded after training, so nothing downstream sees it. λ_{sample}, λ_{cell} are the weights of Equation 1; zeroing either removes its term and its modules, the design space of Section 3.3.

beyond the visible cells alone. Like MAESTRO, which we compare against, SIGCell is therefore a masked set encoder that learns its sample representation by reconstructing a sample's cell distribution at the set level.

The encoder also emits one token $h_i \in \mathbb{R}^{384}$ per encoded cell; these tokens carry the sample-context bias of Section 5.4 and are where our cell-level regularizer acts.

3.2 SIGReg at the sample and cell levels

SIGReg (Balestriero & LeCun, 2025) projects a batch of embeddings onto random unit directions and compares each one-dimensional projection with $\mathcal{N}(0,1)$ using the Epps–Pulley statistic (Epps & Pulley, 1983) (Appendix B). SIGReg requires no labels, teacher network, or batch identifiers, so SIGCell trains 16.2M parameters, about half the 32.5M of MAESTRO, whose self-distillation keeps a full exponential-moving-average (EMA) copy of the network as its teacher (Table 11).

We compute this statistic at two sites, the sample embedding z that a downstream probe reads and the cell tokens:

$$\mathcal{L} = \underbrace{\mathcal{L}_{\text{recon}}}_{\text{masked set}} + \lambda_{\text{sample}} \underbrace{\text{SIGReg}(z)}_{\text{sample-SIGReg}} + \lambda_{\text{cell}} \underbrace{\text{SIGReg}(\text{BN}(WH))}_{\text{cell-SIGReg}}, \tag{1}$$

where H denotes the cell tokens pooled across the effective mini-batch. Each optimizer step takes an effective batch of 12 samples (three on each of four GPUs), each a 10,000-cell subset; sample-SIGReg is computed over their 12 sample embeddings. For cell-SIGReg, we draw up to 2,048 visible-cell tokens per sample and project them from 384 to 64 dimensions with a learned linear map W and batch normalization (BN) over the three samples on each GPU, yielding up to 24,576 projected tokens in total. Each SIGReg statistic is computed once over its full effective batch; because the statistic grows with the number of points, the cell-level one is pinned to $n = 12$ (Appendix B).

The sample site and the cell site. The two terms apply one statistic to two populations: at the sample site each sample is one point, and matching an isotropic Gaussian pushes the samples apart, which prevents collapse; at the cell site, twelve samples contribute up to 2,048 points each.

Proposition 1. *If a cloud in $\mathbb{R}^d$ pooled from the cells of S samples has identity covariance, as at the target of SIGReg, at most a fraction $(S-1)/d$ of its variance, the trace ratio $\operatorname{tr}\Sigma_{\text{B}}/\operatorname{tr}\Sigma$ of between-sample to total covariance, lies between samples; if it has that covariance after standardization within G groups of samples, at most $(S-G)/d$, which is zero when every sample is its own group.*

Proof. Write $I_d = \Sigma_{\text{W}} + \Sigma_{\text{B}}$, the within- and between-sample covariances; both are positive semidefinite, so $\Sigma_{\text{B}} \preceq I_d$. After centering, the sample means of each group, weighted by their cell

counts, average to zero ($G = 1$ without standardization), so Σ_{B} has rank at most $S - G$ and, being $\preceq I_d$, trace at most $S - G$, out of a total of d. $\square$

For cell-SIGReg, which tests a 64-dimensional projection batch-normalized within $G = 4$ groups of the $S = 12$ samples, the bound is $.125$, and the trained model's share is $.030$. Beyond second moments, the test also rejects a batch of samples that are shifted copies of one cloud. At the target, the statistic thus bounds how far samples may sit apart in the tested head, without receiving sample identities or aligning any sample to another, and standardized per sample it is blind: scaling every sample's offset by five leaves it unchanged but raises the training statistic twelve-fold (Appendix B). The bound concerns the 64 directions the statistic sees, not the 384-dimensional tokens (Section 5.4).

3.3 Design space

Setting either regularization weight to zero removes the corresponding term and its associated modules, yielding a 2×2 design in which the encoder, reconstruction objective, corpus, and training budget are fixed: reconstruction only, sample-SIGReg only, cell-SIGReg only, and both. We also retrain MAESTRO, a masked set encoder that adds self-distillation to distribution reconstruction, on the same corpus and training budget at matched width and depth, and report it beside the four conditions as a reference.

SIGCell uses $\lambda_{\mathrm{sample}} = .05$, the published SIGReg default, and $\lambda_{\mathrm{cell}} = 1$, set once from the relative training-loss scale when cell-SIGReg was introduced; neither value was selected on external-cohort performance, so the sweeps of Section 5.5 are sensitivity analyses.

4 Experimental setup

4.1 Datasets and tasks

All three datasets are whole-blood mass-cytometry datasets acquired with the same commercial 30-marker panel. Pretraining uses $1{,}872$ samples from one institution without using their labels. Of these, $1{,}658$ have one of thirteen diagnosis labels; after excluding 126 repeated acquisitions of technical-control donors, $1{,}532$ samples form the internal classification task. The two external datasets are Geanon (Geanon et al., 2021), with 92 samples assigned to four COVID-19 severity classes, and Penttilä (Penttilä et al., 2021), with 124 samples spanning healthy controls, ward COVID-19, and multiple stages of the intensive care unit (ICU) course. Neither external dataset contributes any cells to pretraining (Appendix C).

Geanon's four classes define an ordered COVID-19 severity scale and serve as the external target for frozen linear probing. Penttilä's deposited labels combine disease status with position along a patient's ICU course and do not define a single severity scale, so we fit no severity probe there and use it only for the two evaluations that need no such labels: source-fitted COVID-19 detection (acute COVID-19 against healthy donors) and the cell-type read-out.

The internal classification task is transductive because its samples were seen, without labels, during pretraining. In contrast, both external datasets are entirely unseen during pretraining, so our cross-institution transfer claims are based on inductive evaluation.

4.2 Evaluation axes

Frozen linear probing (Section 5.1) fits a linear classifier on a fixed representation within the dataset being evaluated, measuring the linear accessibility of task-relevant information rather than the best performance attainable after adaptation. *Source-fitted transfer* (Section 5.2) asks whether a diagnostic rule learned at one institution works at another without any labels there. Each model is trained for COVID-19 detection on the 322 internal samples, adapted the way its authors adapt it to a new task (from a linear classifier on a frozen encoder to end-to-end fine-tuning), and the learned rule, decision threshold included, is applied unchanged to both external datasets. Appendix H states what each model freezes and fits, and Appendix I lists the deviations from the published recipes. *Ablation* (Section 5.3) changes only the latent-space control, holding the encoder, reconstruction objective, corpus and schedule fixed. *Sample-context bias analysis* (Section 5.4) measures how cell-SIGReg changes cell-token geometry without fitting downstream labels.

Table 1: **Frozen linear probing.** Balanced accuracy (BA) and macro one-vs-rest AUROC, mean ± SD over ten donor-grouped split seeds; **bold** best in a column, underline second. BA chance is .077 and .250 for the two tasks; AUROC chance is .5. The internal task is transductive for both set encoders and excludes the 126 repeated acquisitions of the technical-control donors, leaving 1,532 samples; the severity task is inductive. Manual gating is the composition of the fourteen gated cell types both datasets share.

	Internal diagnosis, 13 classes		COVID-19 severity grading, 4 classes	
model	BA	AUROC	BA	AUROC
SIGCell	.789±.011	.979±.002	**.501**±.032	**.680**±.028
MAESTRO	**.791**±.010	**.984**±.003	.439±.036	.666±.029
CKME	.719±.011	.964±.002	.441±.046	.660±.035
k-means	.655±.012	.940±.003	.404±.053	.610±.032
FlowSOM	.591±.015	.904±.003	.398±.046	.618±.029
manual gating	.488±.009	.853±.002	.426±.036	.628±.022

4.3 Protocol

Frozen linear probing uses five-fold cross-validation grouped by donor, repeated with ten seeds and identical folds for every model; the probe is a standard scaler and an ℓ_2-regularized logistic regression with balanced class weights and $C = 1$, both fit on training folds only. In source-fitted transfer the frozen encoders' probe is fitted on the training part of each of five donor-grouped folds of the 322 COVID-19 detection samples per seed, averaging the five rules' metrics (Appendix G); the threshold is fixed at .5 before either external dataset is scored, and no external label is used. Results are mean ± SD over ten seeds. Every pretrained or supervised comparison model is retrained from its released code on our corpus and panel, the pretrained ones for 500 epochs at an effective batch of 12 samples; Appendix G gives the full protocol and Appendix I each model's deviations.

5 Results

5.1 Frozen linear probing

Frozen linear probing fixes the representation and fits only a linear classifier on it, so it admits only sample representations fixed without labels: the two set encoders, whose learned pooling is part of the representation, and the four training-free baselines (Appendix G). The other pretrained models form theirs with each task's labels (Section 2), so they appear in Section 5.2 instead.

SIGCell and MAESTRO are both masked set encoders trained by distribution reconstruction, here on the same corpus under the same training schedule; the controlled comparison is the ablation of Section 5.3. In Table 1, MAESTRO is slightly ahead on the internal thirteen-diagnosis task, .791 against .789 balanced accuracy and .984 against .979 AUROC; on the external severity task the order reverses, .501 and .680 for SIGCell against .439 and .666.

5.2 Source-fitted transfer, each model in its own adaptation regime

We next compare each model in its own downstream adaptation regime. All methods are fitted on the same internal COVID-19 detection task and applied to both external datasets without recalibration; SIGCell keeps its encoder frozen and fits only a linear classifier. Everything outside each regime is shared, from the source roster to the test cell budget (Appendix I).

SIGCell is first in every column while fitting only a linear probe on a frozen encoder. The rows also differ in architecture, pretraining and size (Tables 11 and 13), so the table compares whole models; the controlled evidence for what cell-SIGReg adds is Table 3. Two observations qualify this comparison. First, models that retrain a network vary far more across seeds than models with frozen encoders, because a seed re-partitions the source for a frozen encoder's probe but repeats the training itself for a retrained network (Appendix G).

Table 2: **Source-fitted transfer, every model in its own adaptation regime.** Fitted on the 322 internal samples of the COVID-19 detection task and applied to the whole external dataset: frozen rows fit one rule on each training part of five donor-grouped folds and average the five rules' metrics; rows that train a network are refit on all 322 samples. Appendix H states what each row keeps frozen and what it fits on the source labels. **Bold** best in a column, underline second. Mean ± SD over ten split seeds.

	Geanon		Penttilä	
model	AUROC	BA	AUROC	BA
SIGCell	**.949**±.002	**.871**±.016	**.995**±.002	**.965**±.006
MAESTRO	.832±.002	.712±.012	.977±.003	.843±.006
cyMAE	.908±.002	.836±.013	.989±.002	.930±.003
GPCT	.909±.019	.705±.064	.975±.008	.804±.053
CytoBERT	.886±.165	.797±.155	.953±.026	.859±.053
CellCnn	.930±.006	.603±.094	.989±.008	.881±.070
CytoSet	.847±.134	.559±.050	.924±.129	.656±.115
CytoDx	.892±.002	.487±.000	.946±.001	.935±.005
Deep Sets	.920±.003	.684±.129	.981±.006	.735±.158
Set Transformer	.911±.024	.861±.023	.955±.047	.900±.060

Table 3: **Latent-space control.** The four SIGCell conditions share encoder, pretext, corpus, budget and seed and differ only in the objective; the self-distillation row is MAESTRO, retrained on the same corpus and budget. Mean ± SD over ten split seeds, each axis re-partitioned as in Section 4. Fourteen gated cell types, fitted on internal cells and applied directly to external cells (chance .071).

	frozen linear probing (BA)		source-fitted (AUROC)		cell-type read-out	
latent-space control	internal	severity	Geanon	Penttilä	Geanon	Penttilä
reconstruction only	.792±.006	.459±.039	.871±.003	.968±.005	.164±.008	.260±.018
self-distillation (MAESTRO)	.791±.010	.439±.036	.832±.002	.977±.003	.191±.026	.573±.016
sample-SIGReg only	**.807**±.008	.397±.030	.864±.006	.939±.004	.133±.011	.191±.014
cell-SIGReg only	.785±.013	.422±.041	.881±.003	.973±.002	**.691**±.010	**.787**±.007
both (SIGCell)	.789±.011	**.501**±.032	**.949**±.002	**.995**±.002	.635±.016	.757±.006

Second, much of the degradation appears at the fixed decision threshold rather than in ranking alone. On Geanon, all methods retain AUROC between .83 and .95, whereas balanced accuracy varies much more widely. For example, CytoDx reaches AUROC .892 but classifies none of the fifteen healthy Geanon samples correctly at the transferred threshold. This calibration shift coincides with a large change in case mix between the source and external datasets, although the prevalence difference alone does not establish its cause.

5.3 Ablation: contributions of sample-SIGReg and cell-SIGReg

Cell-SIGReg has its largest effect on cross-sample cell-level transfer (Figure 2). With sample-SIGReg present, adding the cell term raises the Geanon and Penttilä cell-type read-outs from .133 to .635 and from .191 to .757. The effect is similar without sample-SIGReg: reconstruction alone reaches .164 and .260, whereas cell-SIGReg alone reaches .691 and .787.

At the sample level, however, the effect of cell-SIGReg depends strongly on the presence of sample-SIGReg. Relative to sample-SIGReg alone, adding the cell term improves external severity balanced accuracy by .105 and source-fitted AUROC by .085 on Geanon and .056 on Penttilä. In contrast, relative to reconstruction alone, cell-SIGReg alone reduces severity balanced accuracy by .037. The interaction on severity is therefore +.141: the external sample-level gain is not additive. Sample-SIGReg alone improves the internal probe but reduces every external read-out, and the full configuration costs .019 balanced accuracy in distribution against it (Appendix E). A large part of SIGCell's

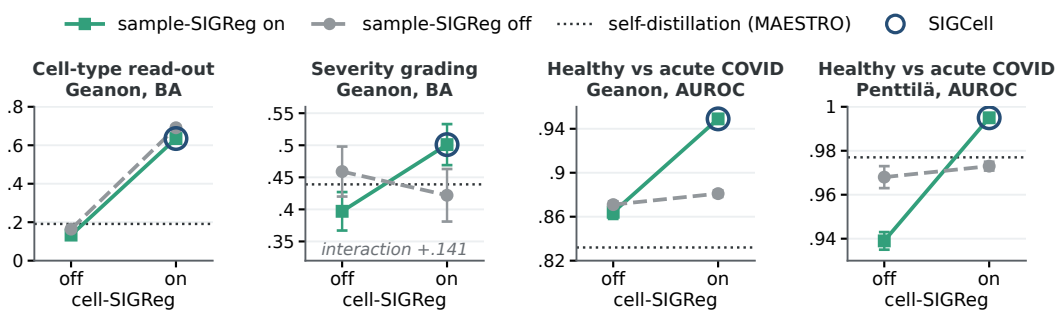

Figure 2: **The ablation of Table 3 as a 2×2 design.** Each read-out against cell-SIGReg off or on, with sample-SIGReg on (green) or off (gray); the ring marks SIGCell and the dotted line self-distillation (MAESTRO). Parallel lines mean the two terms add, crossing lines that they interact. Mean $\pm$ SD over ten split seeds; y-ranges differ by panel.

benchmark performance derives from the masked set-encoder backbone and reconstruction objective rather than from SIGReg alone.

5.4 What cell-SIGReg changes: the sample-context bias

Two label-free measurements of the cell tokens quantify the sample bias (Table 6; full definitions in Appendix D). The *between-sample share of token variance* is the fraction of the total cell-token variance of a dataset explained by which sample a token comes from, the between-group term of a one-way variance decomposition with the sample as the group, corrected for the sampling noise of the sample means; it is near 0 when all samples share one token cloud and near 1 when the sample alone fixes a cell's token. Without cell-SIGReg, .93–.98 of cell-token variance lies between samples on all three datasets, against .10–.19 for the raw markers; with cell-SIGReg the share falls to .12–.24. A large share could also reflect genuine biological differences between samples, so we also test context dependence directly, holding the cells fixed and changing only their sample context.

The *context-swap displacement* re-encodes the same 1,000 cells inside another sample's cells and measures how far their tokens move, in units of their own cell-to-cell spread ρ, the root-mean-square distance of the tokens from their mean (Equation 9); a representation computed cell by cell, such as the raw markers, cannot move at all. Without cell-SIGReg the tokens move by six to nine times their spread, toward the token cloud of the new context, whereas re-encoding them with *other cells of the same sample* moves them by under a third of one spread. With cell-SIGReg the displacement falls to about one spread. The sample bias precedes training, with .95–.97 of token variance between samples at random initialization (Appendix D), and the sample-level objectives strengthen it: 6.4 spreads under reconstruction alone, 8.9 with sample-SIGReg, and 7.5 in MAESTRO.

Cell-type information nevertheless remains decodable within a sample, at .77–.82 in every condition, with or without the cell term (Table 6); what fails without cell-SIGReg is the transfer of a read-out across samples. Together with the context swap, this suggests that cell-SIGReg improves transfer by making cell-token coordinates comparable across samples, not by separating cell types better within a sample, where its read-out is slightly lower (Appendix D, Table 7).

Unlike batch correction (Section 2), cell-SIGReg uses no batch labels or reference samples, at the default weight the between-sample share stays above its raw-marker level, and cohort identity stays linearly decodable from every representation; it removes structure the set encoder adds, so the two could be combined.

5.5 Sensitivity to the loss weights

The cell weight. Increasing λ_{cell} changes the cell-token geometry monotonically: the cell-type read-out rises from .18 to .83 internally and from .13 to .72 on Geanon, while both the between-sample share of token variance and the context-swap displacement fall (Appendix E); at $\lambda_{\text{cell}} = 2$ the share, .15, falls below the raw markers' .19, removing structure the measurements themselves carry.

The sample-level read-outs, however, are not monotonic in λ_{cell}. Source-fitted transfer improves up to the default setting and then changes little, with a small decrease in AUROC at twice the default. At $\lambda_{\text{cell}} = .25$, Geanon balanced accuracy drops to $.553$ even though AUROC remains $.834$, reproducing within the sweep the threshold sensitivity observed in Section 5.2. The default was fixed before external evaluation (Section 3.3); it is the best of the five on the severity probe, while $\lambda_{\text{cell}} = 2$ gives stronger cell-type read-outs at a small cost there.

The sample weight. With $\lambda_{\text{cell}} = 1$, the default $\lambda_{\text{sample}} = .05$ gives the strongest severity and source-fitted transfer results among the four values tested, while increasing the weight to $.2$ degrades every external sample-level read-out. Increasing λ_{sample} also reduces cell-level performance: the cell-type read-out falls from $.80$ to $.61$ internally and from $.69$ to $.52$ on Geanon across the sweep.

6 Discussion

Why the sample term needs the cell term. Isotropy of the embedding a probe reads is motivated by the bias and variance of the probes fitted on it, in an analysis where test points follow the training distribution (Balestriero & LeCun, 2025); in distribution our results agree, as sample-SIGReg alone gives the best internal probe of Table 3. Transfer adds a step this argument does not cover: in a sample-specific frame, the samples of a new institution are encoded among their own cells and their tokens move with that context (Section 5.4). The sample term alone does worse than reconstruction alone on every external read-out, but once the cell term has made the frame shared, the same term raises external severity grading from $.422$ to $.501$, an interaction of $+.141$. We read the two terms as doing different work, the cell term making samples comparable and the sample term shaping the embedding the probe reads; neither alone improves external severity grading over reconstruction.

6.1 Limitations

One marker panel. All models use the same 30-marker panel (Appendix C) and SIGCell is trained for that input space: a sample measured with another panel cannot be encoded without modifying the model, and our experiments do not establish that cell-SIGReg behaves the same in a variable-panel encoder. The transfer studied here is cross-institution transfer within a fixed panel, not panel-independent transfer.

Small external datasets. The only external severity task contains 92 samples across four classes, with 14 samples in the smallest class, so its cross-validated estimates remain sensitive to a small number of cases. The source-fitted binary evaluations are similarly constrained by the number of healthy controls: Geanon contains 15 and Penttilä only 8, so a single healthy error changes balanced accuracy by $.033$ and $.063$, respectively. Larger external datasets are needed to determine the stability of the sample-level margins reported here.

7 Conclusion

In a masked set encoder, a cell's token can depend strongly on its sample context. We show that, without cell-level regularization, $.93$–$.98$ of cell-token variance lies between samples, and changing only the sample context moves the same cells by six to nine times their cell-to-cell spread; this distortion impairs the transfer of source-fitted cell-type read-outs across datasets. Cell-SIGReg applies an isotropic Gaussian constraint to cell tokens pooled across samples, using no sample or batch labels. We demonstrate that it lowers the between-sample share to $.12$–$.24$ and raises a source-fitted cell-type read-out on an unseen dataset from $.13$–$.19$ to $.64$–$.69$ balanced accuracy. Together with sample-SIGReg, it yields a frozen representation that grades external COVID-19 severity at $.501$ balanced accuracy, against at most $.441$ for the other frozen representations, and whose source-fitted COVID-19 detection rule transfers to two external datasets without recalibration at AUROC $.949$ and $.995$, first among the ten models compared. The sample-context bias is thus an important failure mode of masked cytometry set encoders, and cell-level latent regularization one way to mitigate it.

AI USE STATEMENT

In this work, we used generative AI tools to polish writing and to find related works. Additionally, we used AI tools for proving mathematical claim in Proposition 1. We take responsibility for the final content of this work.

ETHICS STATEMENT

This study analyzes de-identified human peripheral blood cytometry data. The internal cohort was obtained under a data use agreement from the authors of (Lee et al., 2025). The two external cohorts are previously published (Geanon et al., 2021; Penttilä et al., 2021). We have carefully considered potential ethical issues related to data privacy, fairness, and misuse. We believe our findings pose no foreseeable risks of harm, and all contributions were made in compliance with the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

We have included key methodological and experimental details in the main paper to support reproducibility. Details on training data distributions and hyperparameter settings for pre-training are provided in Appendix. We will also release all codes to facilitate faithful reproduction of our findings. The two external cohorts are publicly available from their original studies (Geanon et al., 2021; Penttilä et al., 2021). And the internal cohort can be obtained under a data use agreement from the authors of (Lee et al., 2025).

REFERENCES

Eirini Arvaniti and Manfred Claassen. Sensitive detection of rare disease-associated cell subsets via representation learning. *Nature Communications*, 8:14825, 2017. doi: 10.1038/ncomms14825.

Charles Bruce Bagwell, Benjamin Hunsberger, Beth Hill, Donald Herbert, Christopher Bray, Thirumahal Selvanantham, Stephen Li, Jose C. Villasboas, Kevin Pavelko, Michael Strausbauch, Adeeb Rahman, Gregory Kelly, Shahab Asgharzadeh, Azucena Gomez-Cabrero, Gregory Behbehani, Hsiaochi Chang, Justin Lyberger, Ruth Montgomery, Yujiao Zhao, Margaret Inokuma, Ofir Goldberger, and Greg Stelzer. Multi-site reproducibility of a human immunophenotyping assay in whole blood and peripheral blood mononuclear cells preparations using CyTOF technology coupled with Maxpar Pathsetter, an automated data analysis system. *Cytometry Part B: Clinical Cytometry*, 98(2):146–160, 2020. doi: 10.1002/cyto.b.21858.

Randall Balestriero and Yann LeCun. LeJEPA: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*, 2021.

Sirui Ding, Sanchita Bhattacharya, and Atul J. Butte. ImmuneFM: Pre-training foundation model from cytometry data for immunology research. *bioRxiv*, 2025. doi: 10.1101/2025.07.09.664020.

T. W. Epps and Lawrence B. Pulley. A test for normality based on the empirical characteristic function. *Biometrika*, 70(3):723–726, 1983. doi: 10.1093/biomet/70.3.723.

Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. Whitening for self-supervised representation learning. In *International Conference on Machine Learning (ICML)*, 2021.

Daniel Geanon, Brian Lee, Edgar Gonzalez-Kozlova, Geoffrey Kelly, Diana Handler, Bhaskar Upadhyaya, John Leech, Ronaldo M. De Real, Manon Herbinet, Assaf Magen, Diane Del Valle, Alexander Charney, Seunghee Kim-Schulze, Sacha Gnjatic, Miriam Merad, and Adeeb H. Rahman. A streamlined whole blood CyTOF workflow defines a circulating immune cell signature of COVID-19. *Cytometry Part A*, 99(5):446–461, 2021. doi: 10.1002/cyto.a.24317.

F. M. Howard, J. Dolezal, S. Kochanny, J. Schulte, H. Chen, L. Heij, D. Huo, R. Nanda, O. I. Olopade, J. N. Kather, N. Cipriani, R. L. Grossman, and A. T. Pearson. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nature Communications*, 12:4423, 2021. doi: 10.1038/s41467-021-24698-1.

Zicheng Hu, Benjamin S. Glicksberg, and Atul J. Butte. Robust prediction of clinical outcomes using cytometry data. *Bioinformatics*, 35(7):1197–1203, 2019. doi: 10.1093/bioinformatics/bty768.

J. Kim, M. Ionita, M. Lee, M. L. McKeague, A. Pattekar, M. M. Painter, J. Wagenaar, V. Truong, D. T. Norton, D. Mathew, Y. Nam, S. A. Apostolidis, C. Clendenin, P. Orzechowski, S. H. Jung, J. Woerner, C. A. G. Ittner, A. P. Turner, M. Esperanza, T. G. Dunn, N. S. Mangalmurti, J. P. Reilly, N. J. Meyer, C. S. Calfee, K. D. Liu, M. A. Matthy, L. B. Swigart, E. L. Burnham, J. McKeehan, S. Gandotra, D. W. Russel, K. W. Gibbs, K. W. Thomas, H. Barot, A. R. Greenplate, E. J. Wherry, and D. Kim. Cytometry masked autoencoder: An accurate and interpretable automated immunophenotyper. *Cell Reports Medicine*, 5(11):101808, 2024. doi: 10.1016/j.xcrm.2024.101808.

I. Korsunsky, N. Millard, J. Fan, K. Slowikowski, F. Zhang, F. Zhang, K. Wei, Y. Baglaenko, M. Brenner, P.-R. Loh, and S. Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods*, 16(12):1289–1296, 2019. doi: 10.1038/s41592-019-0619-0.

Jheng-Ling Lee and Shang-Tse Chen. Joint-embedding prediction of masked point tubes for self-supervised learning on 4D point cloud videos. *arXiv preprint arXiv:2608.24093*, 2026.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International Conference on Machine Learning (ICML)*, 2019.

Matthew Eric Lee, Jaesik Kim, Matei Ionita, Jonghyun Lee, Michelle L. McKeague, Yonghyun Nam, Irene Khavin, Yidi Huang, Victoria Fang, Sokratis Apostolidis, Divij Mathew, Shwetank, Ajinkya Pattekar, Zahabia Rangwala, Amit Bar-Or, Benjamin A. Fensterheim, Benjamin A. Abramoff, Rennie L. Rhee, Damian Maseda, Allison R. Greenplate, E. John Wherry, and Dokyoon Kim. MAESTRO: Masked encoding set transformer with self-distillation. In *International Conference on Learning Representations (ICLR)*, 2025.

Guillaume Letellier, Antonin Vacheret, and Frederic Jurie. JetParticle-JEPA: An efficient self-supervised representation learning method for jet tagging in high-energy physics. *arXiv preprint arXiv:2606.14813*, 2026.

Mattia Medina Grespan, Muir J. Morrison, Brendan O'Fallon, Ryan Shean, Nick C. Spies, and David P. Ng. EventHorizon: A foundation model for clinical flow cytometry. *bioRxiv*, 2026. doi: 10.64898/2026.06.18.733197.

B. O'Fallon, M. Morrison, M. Medina Grespan, N. C. Spies, and D. P. Ng. DinoFlow: Self-supervised pretraining in flow cytometry enables accurate detection of common hematopathological disorders. *Cytometry Part B: Clinical Cytometry*, 110(5):347–358, 2026. doi: 10.1002/cyto.b.70043.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

C. B. Pedersen, S. H. Dam, M. B. Barnkob, M. D. Leipold, N. Purroy, L. Z. Rassenti, T. J. Kipps, J. Nguyen, J. A. Lederer, S. H. Gohil, C. J. Wu, and L. R. Olsen. cyCombine allows for robust integration of single-cell cytometry datasets within and across technologies. *Nature Communications*, 13:1698, 2022. doi: 10.1038/s41467-022-29383-5.

P. A. Penttilä, S. Van Gassen, D. Panovska, L. Vanderbeke, Y. Van Herck, K. Quintelier, A. Emmaneel, J. Filtjens, B. Malengier-Devlies, K. Ahmadzadeh, P. Van Mol, D. M. Borràs, A. Antoranz, F. M. Bosisio, E. Wauters, K. Martinod, P. Matthys, Y. Saeys, A. D. Garg, J. Wauters, and F. De Smet. High dimensional profiling identifies specific immune types along the recovery trajectories of critically ill COVID19 patients. *Cellular and Molecular Life Sciences*, 78(8): 3987–4002, 2021. doi: 10.1007/s00018-021-03808-8.

Ben Peters, Vlad Niculae, and André F. T. Martins. Sparse sequence-to-sequence models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019. doi: 10.18653/v1/P19-1146.

Syed Abdul Haseeb Qadri, Bjarne C. Hiller, Felix Blanke, Vanja Sophie Cangalovic, Kutalmış Coşkun, Amin Mirzaei, Tom Siegl, Sebastian Bader, Thomas Kirste, and Martin Becker. CytoBERT: A foundation model for cytometry data. *arXiv preprint arXiv:2608.14414*, 2026.

Alexander M. Rusnak, Sophia Kovalenko, Jingru Wang, Ismail Moudden, Xiru Wang, and Frédéric Kaplan. Polis: 3D self-supervision at city scale. *arXiv preprint arXiv:2608.29426*, 2026.

Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, and Hervé Jégou. Spreading vectors for similarity search. In *International Conference on Learning Representations (ICLR)*, 2019.

Siyuan Shan, Vishal Athreya Baskaran, Haidong Yi, Jolene Ranek, Natalie Stanley, and Junier B. Oliva. Transparent single-cell set classification with kernel mean embeddings. In *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB)*, 2022. doi: 10.1145/3535508.3545538.

M. Trussart, C. E. Teh, T. Tan, L. Leong, D. H. Gray, and T. P. Speed. Removing unwanted variation with CytofRUV to integrate multiple CyTOF datasets. *eLife*, 9:e59630, 2020. doi: 10.7554/eLife.59630.

Sofie Van Gassen, Britt Callebaut, Mary J. Van Helden, Bart N. Lambrecht, Piet Demeester, Tom Dhaene, and Yvan Saeys. FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry Part A*, 87(7):636–645, 2015. doi: 10.1002/cyto.a.22625.

Sofie Van Gassen, Brice Gaudilliere, Martin S. Angst, Yvan Saeys, and Nima Aghaeepour. CytoNorm: A normalization algorithm for cytometry data. *Cytometry Part A*, 97(3):268–278, 2020. doi: 10.1002/cyto.a.23904.

Haidong Yi and Natalie Stanley. CytoSet: Predicting clinical outcomes via set-modeling of cytometry data. In *Proceedings of the 12th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB)*, 2021. doi: 10.1145/3459930.3469529.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Russ R. Salakhutdinov, and Alexander Smola. Deep sets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning (ICML)*, 2021.

Tinghe Zhang, Jian Xu, Jiaheng Chen, Jiaxing Li, Yucheng Xiao, and Qiang Wang. NodeJEPA: Structure-conditioned latent prediction for node-level graph self-supervised learning. *arXiv preprint arXiv:2608.04381*, 2026.

Zixin Zhuang, Benjamin S. Mashford, Liang Zheng, and T. Daniel Andrews. Towards a cytometry foundation model: Interpretable sample-level predictive modelling via pretrained transformers. *bioRxiv*, 2026. doi: 10.64898/2026.03.31.712806.

A THE SET-ATTENTION BLOCKS

Figure 1 names three blocks. All three come from Set Transformer (Lee et al., 2019), and all three are built from the same multihead attention block (MAB), so we state that first. Writing X for the queries and Y for the keys and values,

$$\mathrm{MAB}(X, Y) = \mathrm{LayerNorm}(A + \mathrm{FF}(A)), \qquad A = \mathrm{LayerNorm}(X + \mathrm{MHA}(X, Y, Y)), \quad (2)$$

with MHA the usual multihead attention and FF a cell-wise feed-forward map, a gated linear unit in our implementation. The three blocks differ only in what they attend to.

Induced Set-Attention Block (ISAB). Attention of a set to itself costs $O(N^2)$ in the number of cells, which cytometry cannot afford. ISAB routes it through $m \ll N$ learnable inducing points $I \in \mathbb{R}^{m \times d}$ in two steps:

$$\mathrm{ISAB}_m(X) = \mathrm{MAB}(X, \tilde{I}) \in \mathbb{R}^{N \times d}, \qquad \tilde{I} = \mathrm{MAB}(I, X) \in \mathbb{R}^{m \times d}. \quad (3)$$

The inducing points first read the set, then the set reads them back, so information passes between cells only through a bottleneck of size m and the cost falls to $O(Nm)$. The block is permutation *equivariant*: permute the cells and its outputs permute with them, which is what lets it produce one token per cell. SIGCell stacks five of these with $m = 16$.

Pooling by Multihead Attention (PMA). The same block with a small set of learnable seed vectors Q as the queries,

$$\mathrm{PMA}(X) = \mathrm{MAB}(Q, X), \quad (4)$$

which turns a set of any size into as many outputs as there are seeds. With one seed it is permutation *invariant* and collapses a sample's cells into a single vector: that is the step from h to z in Figure 1, and it is what makes the sample embedding independent of the order the cells arrive in. The decoder uses the same block in the opposite direction, with one seed per output cell, to unpool z back to a set.

Self-Attention Block (SAB). The set attends to itself,

$$\mathrm{SAB}(X) = \mathrm{MAB}(X, X), \quad (5)$$

which captures pairwise interactions among the elements and returns a set of the same size; stacking them encodes higher-order interactions. It is permutation equivariant and it is quadratic in N, which is why it appears only in the decoder, where the set size is the fixed number of reconstructed cells.

How SIGCell instantiates them. The encoder is a linear lift to 384 dimensions, five ISAB_{16} blocks, and a one-seed PMA followed by an affine head to $z \in \mathbb{R}^{256}$; its cell tokens h are the output of the last ISAB. The decoder is a PMA with one seed per reconstructed cell, two SABs, and a gated linear unit back to the 30 marker dimensions. Attention in the encoder and in its pooling is α-entmax at $\alpha = 1.15$ (Peters et al., 2019), so those weights are sparse and cells can be attended to exactly zero; the decoder uses ordinary softmax. The MAESTRO checkpoint we compare against was retrained on this corpus at matched width and depth; Appendix I lists how each implementation departs from its paper.

The reconstruction loss is a permutation-invariant distance between the generated and the observed set, so the target is the sample's cell distribution rather than any particular cell: the energy distance for the first fifteen epochs, entropic optimal transport thereafter.

B HOW SIGREG IS COMPUTED

Section 3 and Equation 1 state where the two SIGReg terms act. This section states what the statistic is, so that the weights swept in Appendix E have a definite meaning.

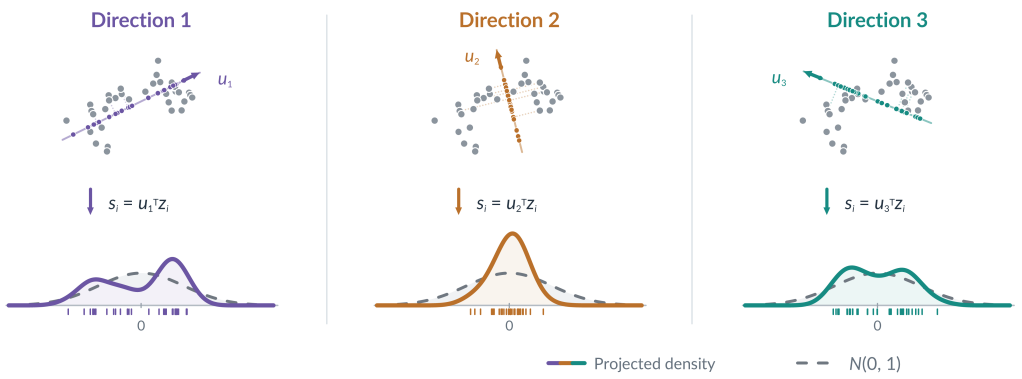

Figure 3: **What the sketch tests.** A batch of embeddings (gray, drawn here in two dimensions for legibility) is projected onto a random unit direction u_j; the colored dots are the projected values $s_i = u_j^\top z_i$, and the curve below each panel is their density against the $\mathcal{N}(0, 1)$ target (dashed). No single direction determines the joint distribution: direction 1 is bimodal, direction 2 is too concentrated, direction 3 is too flat, and a distribution can be Gaussian along one direction while failing along another. Match each 1D sketch to the same Gaussian target, and the only joint law that satisfies every direction at once is the isotropic Gaussian, which is why the statistic is a mean over many directions redrawn each step rather than a single projection.

The statistic. Let $Z \in \mathbb{R}^{n \times d}$ be the embeddings SIGReg is applied to. The term draws M directions $u_1, \dots, u_M$ uniformly on the unit sphere of $\mathbb{R}^d$ (Gaussian vectors normalized to unit length), and projects the batch onto each (Figure 3), $s_i^{(j)} = u_j^\top z_i$. Each projection is compared with $\mathcal{N}(0, 1)$ through its empirical characteristic function

$$\hat{\varphi}_j(t) \;=\; \frac{1}{n}\sum_{i=1}^{n} e^{\mathrm{i}ts_i^{(j)}} \;=\; \frac{1}{n}\sum_{i=1}^{n} \cos\!\left(ts_i^{(j)}\right) \;+\; \frac{\mathrm{i}}{n}\sum_{i=1}^{n} \sin\!\left(ts_i^{(j)}\right), \tag{6}$$

and the Epps–Pulley statistic (Epps & Pulley, 1983) is the squared distance between that function and the Gaussian one, $\varphi(t) = e^{-t^2/2}$, integrated against the Gaussian weight:

$$T_j \;=\; n\int_{-5}^{5} \left|\hat{\varphi}_j(t) - e^{-t^2/2}\right|^2 e^{-t^2/2}\,\mathrm{d}t, \qquad \mathrm{SIGReg}(Z) \;=\; \frac{1}{M}\sum_{j=1}^{M} T_j. \tag{7}$$

The integral is evaluated by the trapezoidal rule on 17 knots spread over $t \in [-5, 5]$, so one evaluation costs $n \times M \times 17$ cosines and sines and nothing else: there is no pairwise distance, no covariance matrix and no eigendecomposition, and the cost is linear in the number of points. The directions are redrawn at every training step from a generator seeded by the step counter and shared by all ranks, so one step tests M directions and successive steps cover the sphere. The batch is never standardized before the test, because a shifted mean, a wrong scale and a deficient rank are exactly the deviations the statistic is there to feel.

One statistic over the pooled batch. Under data parallelism each rank computes the two sums of Equation 6 over its own share of the batch, and the resulting $2 \times M \times 17$ table (real and imaginary parts, per direction, per knot) is averaged across ranks before Equation 7 is formed. The average is taken through an autograd-aware collective, so every rank scores, and receives gradients from, the pooled effective batch rather than its own slice. This is why the effective batch is part of the objective and not only of the optimizer: at the sample site the test sees the 12 embeddings of the effective batch through 1,024 directions in $\mathbb{R}^{256}$, and at the cell site it sees up to 24,576 projected tokens of those same twelve samples through 512 directions in $\mathbb{R}^{64}$. Only the batch normalization of the cell head is not pooled: it standardizes the tokens of each GPU's three samples separately, so the tested cloud is four standardized groups of three samples each, the setting $G = 4$ of Proposition 1. Tokens exist only for the cells the encoder sees, because masked cells are never encoded. Each sample's 10,000-cell subset is drawn without replacement (with replacement only for a sample with fewer cells),

the mask removes $10{,}000\,r$ of them, and the $2{,}048$ tokens are drawn without replacement from the visible rest. At $r = .8$ only $2{,}000$ cells are visible, so all of them are used and the cloud holds $24{,}000$ tokens; since n is pinned to 12 at the cell site (below), this changes the number of points, not the scale of the statistic.

The scale factor, and why the cell term overrides it. The factor n in Equation 7 is what makes T_j a calibrated goodness-of-fit statistic: it grows with the sample size, so a fixed discrepancy is judged more harshly when more points attest to it. SIGReg sets it to the number of points entering the test (Balestriero & LeCun, 2025), and we keep that at the sample site, where it is the effective batch, $n = 12$. At the cell site the number of points is up to $24{,}576$, and keeping the calibrated factor would multiply the term by two thousand relative to the sample term: at initialization it dominates the objective and drives the latent onto the degenerate plateau described below. We therefore pin the cell statistic to the batch scale, $n = 3 \times 4 = 12$ (three samples on each of four GPUs), so both terms enter the loss at one scale and λ_{cell} means the same thing as λ_{sample}. This is a deviation from SIGReg's calibration and is the reason the cell term's weight had to be set from the first-epoch loss scale rather than inherited; Appendix I lists it with the others.

Two consequences of the form. Because $\hat{\varphi}$ is built from cosines and sines, the gradient of T_j with respect to an embedding is bounded however far that embedding is from the target, which is what makes the term safe to add to a reconstruction loss at a large weight. The same form has a failure mode at the opposite end: as a batch contracts toward a single point the projections concentrate and $\partial T_j / \partial z_i \to 0$, so a representation that has already collapsed feels no pressure to recover. SIGReg therefore only constrains shape. Here it always accompanies the reconstruction term, which keeps information in the embedding. Neither term reads a label, a teacher, a batch identifier or a reference sample; the only thing either one sees is the set of embeddings the current effective batch produced.

Why the cell site penalizes the sample-context bias. Proposition 1 is the second-moment part of the answer. With the head's batch normalization computed over each GPU's three samples, the twelve sample means of a batch span at most eight of the 64 head directions, so at the statistic's target at least 56 of them carry only cell-to-cell variation. Beyond second moments, samples that are shifted copies of one cloud make the batch a mixture of shifted clouds, which a test against a single Gaussian rejects. The pooled test is a statement about the union of the batch, so nothing in the objective aligns one sample to another or receives the sample a cell came from.

Table 4 measures how the statistic responds to the sample offsets on SIGCell's trained head. Batches of twelve internal samples with $2{,}048$ tokens each are passed through the head's linear map; each sample's mean offset from the batch mean is multiplied by κ; the result is standardized within groups of k samples, as batch normalization does; and the training statistic is evaluated. The learned scale and shift of the batch normalization stay within $.01$ of 1 and 0, so this standardized view is the tested one. With every sample standardized on its own ($k = 1$) the statistic is the same at every κ: the test cannot see the offsets. With groups of two or more it is nearly flat below the trained offsets, since removing them altogether lowers it by 7% in the training layout, and it rises steeply above them, 12.5-fold at $\kappa = 5$ for groups of three and more for larger groups. At convergence the term therefore acts as a barrier against larger offsets rather than a pull toward none, which is consistent with SIGCell's between-sample share settling just above its raw-marker level (Table 6). Batches of 48 samples with 512 tokens each respond less, because a mixture of many clouds looks more Gaussian.

C COHORTS

Mass cytometry. All three cohorts are measured by mass cytometry (cytometry by time of flight, CyTOF). Antibodies against cellular proteins are each conjugated to a distinct heavy-metal isotope; the stained cells pass one at a time through an inductively coupled plasma, which atomizes and ionizes each cell, and a time-of-flight mass spectrometer counts the ions of every isotope. Because the isotopes are separated by mass rather than by emission spectrum, several dozen markers can be read per cell with little overlap between channels. A measured sample is therefore a matrix with one row per cell and one column per marker, up to hundreds of thousands of cells for a tube of whole blood, with no order among the rows: the unordered set of cells that every model in this

Table 4: **The cell statistic under scaled sample offsets.** SIGCell's trained head; each sample's mean offset in a batch is multiplied by κ ($\kappa = 1$ is the trained model) before standardization within groups of k samples; mean training statistic over 40 batches of twelve samples (10 batches of 48). $k = 3$ is the training layout. The last column is the between-sample share of the standardized cloud at $\kappa = 1$, the quantity bounded by Proposition 1.

k	$\kappa = 0$	$\kappa = 1$	$\kappa = 2$	$\kappa = 3$	$\kappa = 5$	rise $1 \to 5$	between-sample share
twelve samples, 2,048 *tokens each*							
1	.0041	.0041	.0041	.0041	.0041	1.0×	0
2	.0042	.0044	.0070	.0137	.0380	8.6×	.022
3	.0042	.0045	.0083	.0185	.0563	12.5×	.030
12	.0042	.0047	.0106	.0279	.0933	19.9×	.042
48 *samples,* 512 *tokens each*							
1	.0019	.0019	.0019	.0019	.0019	1.0×	0
3	.0019	.0019	.0027	.0053	.0151	8.1×	.031
12	.0019	.0019	.0033	.0077	.0248	13.2×	.043
48	.0019	.0019	.0034	.0084	.0277	14.6×	.046

paper encodes. Marker intensities are ion counts and are conventionally arcsinh-transformed before analysis, as they are in all three cohorts.

Table 5: The three cohorts at a glance. The internal corpus is seen without labels by every pretrained model; its labeled subset, less the 126 repeated acquisitions of technical-control donors, is the thirteen-diagnosis task (1,532 samples). Neither external cohort contributes a cell to any model's pretraining. Donor counts are the grouping units of every cross-validation split in this paper.

cohort	institution	samples	donors	what it is used for
Internal	A	1,872 / 1,658 labeled	841	pretraining; 13-class probe; transfer source
Geanon	B	92	58	severity probe; transfer target; cell-type read-out
Penttilä	C	124	48	transfer target; cell-type read-out

Class distribution. Figure 4 shows every class of every cohort at a glance, and Table 5 their sizes. The pretraining corpus is 1,872 whole-blood samples from one institution; 1,658 of them carry one of thirteen diagnoses: Healthy 340, multiple sclerosis 230, lymphoma 219, premature infant 152, ANCA vasculitis 143, long COVID 127, melanoma 121, acute COVID 108, recovered COVID 103, sepsis 46, hidradenitis suppurativa 39, multiple myeloma 20, term infant 10. Of the Healthy samples, 126 are repeated acquisitions of technical-control donors; the thirteen-diagnosis probe is fitted on the 1,532 samples that remain once they are set aside, so its Healthy class holds 214. They come from 841 donors, and 71% of the samples belong to a donor with more than one draw, which is why every split in this paper is donor-grouped. The Geanon cohort is 92 samples from 58 donors: Healthy 15, moderate COVID 28, severe COVID 14, severe COVID with end-organ damage 35. The Penttilä cohort is 124 samples from 48 donors: Healthy 8, ward COVID 7, ICU admission 20, ICU midpoint 52, ICU discharge 37. The source-fitted transfer task uses the 214 Healthy samples that are not repeated acquisitions of a technical-control donor, together with the 108 acute-COVID samples, from 151 donors, against 15 versus 77 and 8 versus 116 externally; COVID prevalence therefore rises from 34% at the source to 84% and 94% at the targets by study design.

Demographics. The deposited metadata are uneven, and we report only what they contain. Internally, sex is recorded for 1,801 of 1,872 samples (963 male, 838 female) and age for 1,809. For Geanon neither sex nor age is deposited. For Penttilä sex is recorded for all eight healthy donors (3 male, 5 female) and for 108 of the 116 COVID samples (86 male, 22 female), age is not deposited, and treatment is recorded for 108 samples (methylprednisolone 53, supportive care only 30, methylprednisolone with anti-IL-6 16, with anti-IL-1 9). No demographic variable is used by any model or probe.

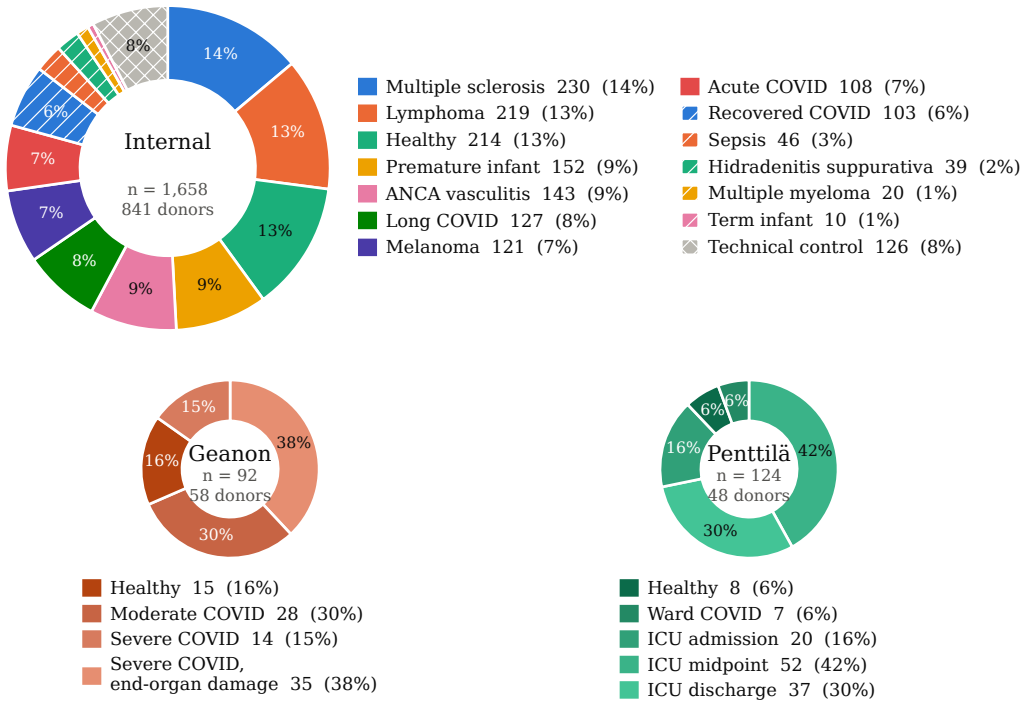

Figure 4: **Diagnosis composition of the three cohorts.** One pie per cohort, one wedge per deposited class label, ordered by size clockwise from the top, with every label, count and share given in the legend beside or below its pie. The internal cohort's thirteen diagnoses have no order, so they take a categorical palette; eight hues is as far as color alone separates, and the five smallest reuse the first five hues with a hatch, so the legend names each wedge. The two external cohorts' classes are ordered, by severity and by position in the intensive-care course, so they take one hue each, shaded dark to light in that clinical order from Healthy, with the legend in the same order. The three cohorts are labeled on different scales, which is why no single task spans them. The internal pie covers all 1,658 labeled samples, and its gray cross-hatched wedge, *technical control*, is the 126 repeated acquisitions of the technical-control donors: they carry the Healthy label and are set aside from the thirteen-diagnosis probe of Table 1, which leaves 1,532. The transfer rule of Section 5.2 is fitted on the internal Healthy and acute-COVID wedges alone, where COVID prevalence is 34%, and is then applied unchanged where it is 84% and 94%.

How far apart the three institutions are. All measurements below compare each external cohort against the internal cohort, with two donor-disjoint halves of the internal cohort as the floor. Per-marker Wasserstein distance in internal standard-deviation units, median over the thirty markers: floor .015, Geanon .405, Penttilä .294. Jensen–Shannon divergence between gated cell-type compositions: floor .0003, .050, .042; the neutrophil share rises from about 60% internally to 75–78%. A classifier of cohort identity built from marker intensities and composition reaches balanced accuracy 1.000 on both external cohorts against a floor of .504. Yet the individual cells are not new: nearest-neighbor manifold coverage of the internal cells is 82.8% and 89.1% against an internal held-out floor of 84.9%, while the density of the shared space is tilted (two-dimensional histogram intersection .291 and .627 against .853). The shift reweights shared cell states: the samples are perfectly distinguishable, while their individual cells are covered by the internal cohort, so a cell-level representation has something to transfer.

Those numbers are over every sample, and the external cohorts are 84% and 94% COVID-19, so by themselves they leave open the objection that the case mix, not the institution, is what a classifier sees. Figure 5 removes that objection by repeating the measurement on the healthy donors alone. Restricted that way the panel still moves: the median marker sits .55 and .38 internal standard deviations from the internal cohort against a .06 floor between donor-disjoint halves of it, every marker

clears that band for Geanon and 27 of 30 do for Penttilä, and the shift survives inside each gated cell type, from .35 to .97 cell-type standard deviations against floors of at most .19. The healthy-only medians are larger than the all-sample ones above because they are measured on different subsets: the all-sample figure pools 1,658 against 92 and 124 samples, the healthy-only figure 214 against 15 and 8.

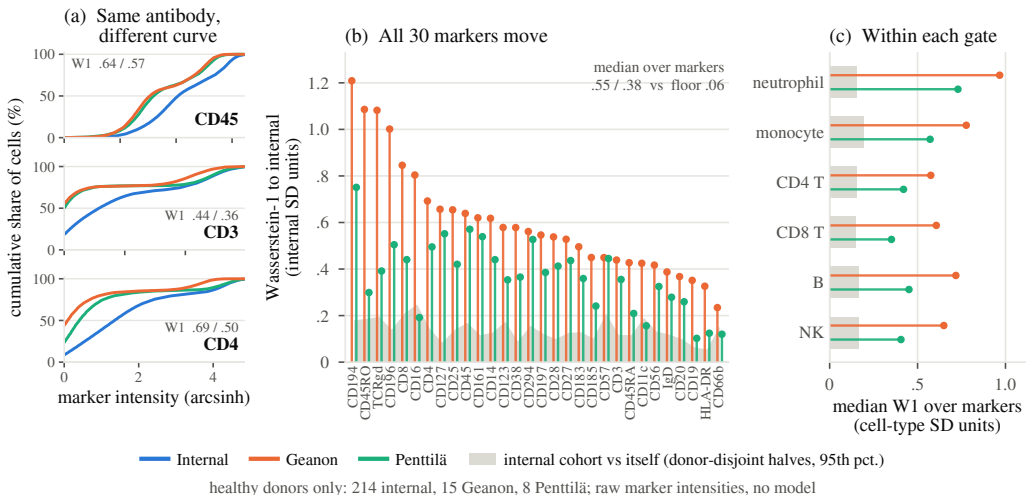

Figure 5: **Healthy donors differ across the three institutions on the same commercial panel.** Healthy donors only, 214 internal against 15 from Geanon and 8 from Penttilä, with 2,000 cells drawn per sample, so that the external cohorts' COVID-19 case mix cannot pose as an institution difference. Raw marker intensities and the deposited gate labels; no model is involved anywhere in this figure. (a) Cumulative distribution of three markers chosen in advance, the pan-leukocyte CD45 and the lineage markers CD3 and CD4, one curve per institution; the number is the 1-Wasserstein distance of each external curve from the internal one in internal standard deviations. Half of the external cells record a CD3 value of exactly zero, against a fifth internally. (b) The same distance for every marker on the panel, against the band spanned by donor-disjoint halves of the internal cohort (95th percentile over 20 splits). The median marker moves .55 and .38 internal standard deviations, ten and seven times the .06 the internal cohort moves against itself; all 30 markers clear the band for Geanon and 27 of 30 for Penttilä. (c) The same statistic restricted to the cells of one gated cell type, in that cell type's own standard deviations, median over markers. The shift survives inside every gate, from .35 to .97 against floors of at most .19, so it is not a difference in which cells were sampled but in how the same cells were stained and acquired.

All three cohorts are acquired with the same commercial 30-marker panel, the Maxpar Direct Immune Profiling Assay (Bagwell et al., 2020); where a cohort added markers, they are dropped, so every model reads the same 30 columns aligned by marker name. Each file also carries cell-type labels from manual gating, the standard expert procedure that assigns every cell to a population through a sequence of thresholds on its marker intensities. The fourteen gated cell types that all three cohorts share are the labels of the cell-type read-out and the composition that the manual-gating baseline of Table 1 reads.

D Effect of cell-SIGReg on the cell embedding

Figures 6–8 show the sample-context bias directly, on the internal cohort and on the two external ones. In Figure 6 every panel is the same 11,165 cells: thirty healthy internal donors, one whole-blood sample each, up to twenty-five cells per donor and deposited gate, so no panel can differ because it saw different cells. The five representations are the raw markers and the cell tokens of four training conditions. The layout is a UMAP and therefore says nothing quantitative; the number printed in each panel is computed in the native token space and is what the ranking rests on.

Under reconstruction alone, under self-distillation and under sample-SIGReg only, every donor is its own island and *all fifteen* of a cell's nearest neighbors come from that donor, against a chance level of .03. Those conditions keep the cell-type structure but repeat it inside each island, which is why the same-type agreement stays near .67 while the read-out of Section 5.4 fails to transfer: the coordinates are sample-specific. Adding cell-SIGReg drops same-donor agreement to .64, merges the islands and raises same-type agreement to .70. Raw markers sit at .27 and .95, so the donor islands come from the encoder.

Self-distillation is the published set encoder, not one of our conditions, and it carries the sample bias as completely as reconstruction alone does, so the sample bias is a property of masked set encoders and not an artifact of this ablation.

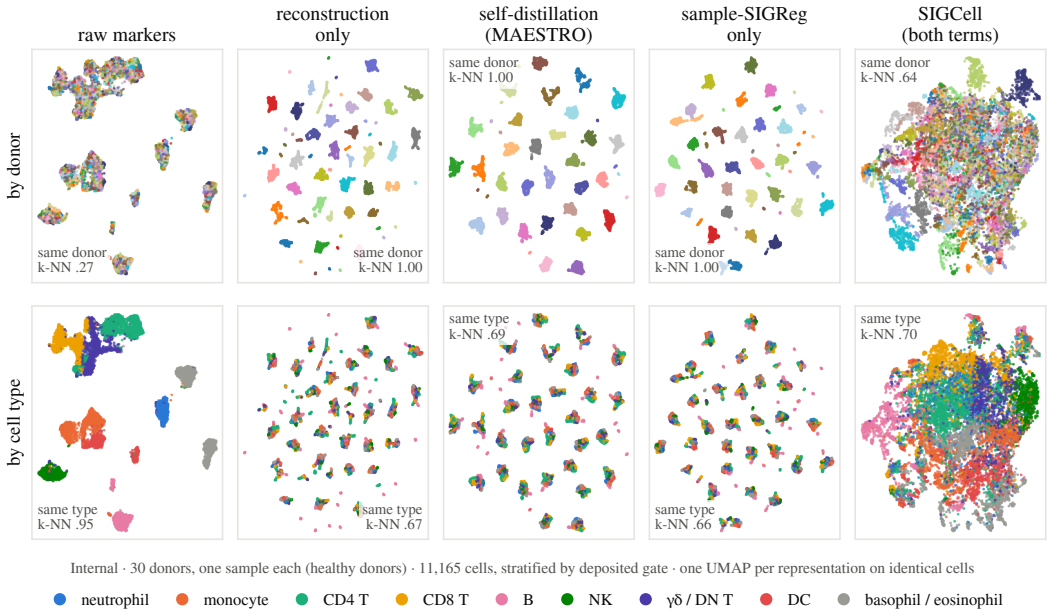

Figure 6: **What cell-SIGReg does to the cell embedding.** The same 11,165 cells from thirty healthy internal donors under five representations: the thirty raw markers, and the cell tokens of encoders trained with reconstruction only, with self-distillation (the retrained MAESTRO of Table 3), with sample-SIGReg only, and with both terms. One UMAP per representation; the top row colors a cell by its donor, with one color per donor and no legend, the bottom row by cell type. The number in each panel is the share of a cell's fifteen nearest neighbors *in the native space*, not in the layout, that come from the same donor (chance .03) or carry the same cell-type label (chance .12). Without the cell term every neighbor is a same-donor neighbor in all three conditions; with it, .64.

The quantitative counterpart is Table 6: donor-half separation falls from .89 to .40 internally against a raw-marker floor of .20, while the cell-type read-out fitted across samples rises. The same measure computed between healthy internal and healthy Geanon cells, a cohort separation that the table does not show, falls from 1.00 to .82 against a floor of .79.

GPCT's cells also attend to one another, yet its tokens carry only a weak sample bias: .22–.32 of their variance lies between samples and a context swap moves them by 0.8 of their spread, close to SIGCell and far from the masked set encoders trained without the cell term, and a read-out fitted across samples stays at the raw-marker ceiling. Its donor-half separation, .66, lies between SIGCell's and theirs. Encoding cells jointly therefore does not by itself produce a sample bias of the size seen in masked set encoders.

The sample bias before training. On a fixed set of cells, 10,000 from each of 48 internal, 24 Geanon and 24 Penttilä samples, a randomly initialized encoder places .949, .971 and .956 of its token variance between samples, where the raw markers of the same cells give .150, .125 and .108. Without the cell term the share rises to about .99 within ten epochs and ends at .96, .93 and .93

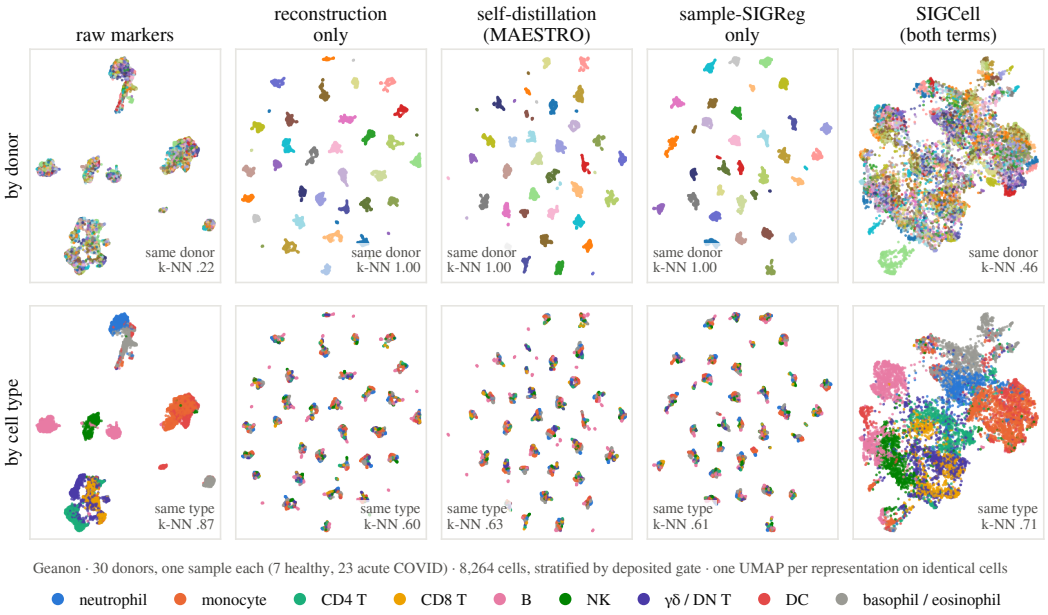

Figure 7: **The cell embedding on Geanon.** As Figure 6, on 8,264 cells from thirty Geanon donors, seven healthy and twenty-three with acute COVID-19, none of whose cells any encoder saw in training. Without the cell term every nearest neighbor is a same-donor neighbor in all three conditions, including the published self-distilled encoder; with it, .46.

Table 6: **The sample-context bias and its removal.** Between-sample share of token variance; token displacement when the same cells are re-encoded inside another sample's cells; donor-half separation among a cell's fifteen nearest neighbors of the same cell type, with the 45 healthy internal donors (one sample each) split in two, 0 being chance and 1 no mixing; and the cell-type read-out on the internal cohort, fitted across samples or inside one sample, with the task, cells and classifier of Table 7 (mean over its ten split seeds). The geometry columns are single fits.

	between-sample share			context	donor-half	cell-type read-out	
latent-space control	internal	Geanon	Penttilä	swap	separation	across	inside one
reconstruction only	.96	.93	.93	6.4	.89	.23	.80
self-distillation (MAESTRO)	.98	.97	.97	7.5	.89	.54	.82
sample-SIGReg only	.98	.96	.96	8.9	.89	.18	.80
both (SIGCell)	.24	.17	.12	1.0	.40	.77	.77
raw markers (floor)	.19	.11	.10	0	.20	.97	.83

under reconstruction alone and at .97, .96 and .95 under sample-SIGReg alone; with it, the share falls to .34–.45 within ten epochs and ends at .22, .17 and .13, while the participation ratio of the within-sample token covariance grows from 8 to 64. The sample bias is therefore present before any learning (Figure 9): the objectives without the cell term keep it, and the cell term removes it.

How the token-geometry columns of Table 6 are computed. The three geometry columns are label-free: none fits a classifier, each is one computation on stored tokens, and the first two are the quantities that the token-geometry rows of Tables 8 and 9 repeat for the sweep points.

Between-sample share. Every sample of the cell-type read-out protocol is encoded as one 10,000-cell set, which gives $n_s = 10{,}000$ tokens $h_{s,i} \in \mathbb{R}^{384}$ ($\mathbb{R}^{128}$ for GPCT) for each of S samples: 334 internal samples, 91 Geanon and 124 Penttilä. Geanon's 92nd sample holds 7,459 cells, fewer than one 10,000-cell set, so it is left out of every measurement that encodes a sample as such a set,

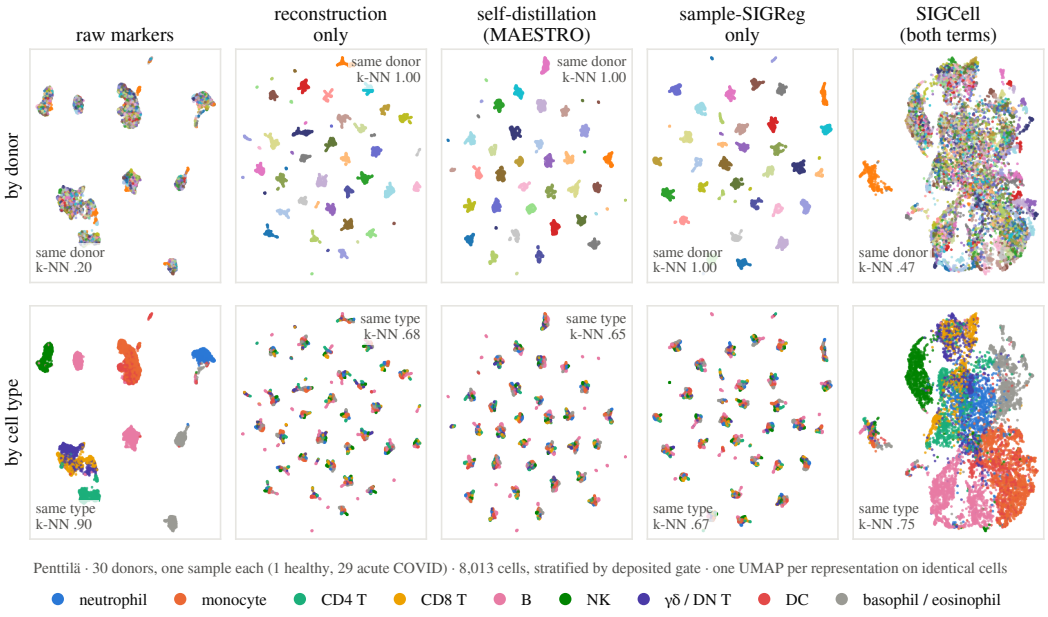

Figure 8: **The cell embedding on Penttilä.** As Figure 6, on 8,013 cells from thirty Penttilä donors, one healthy and twenty-nine with acute COVID-19. The pattern is the same: 1.00 same-donor agreement in every condition without the cell term, .47 with it, and the highest same-type agreement of the three cohorts, .75.

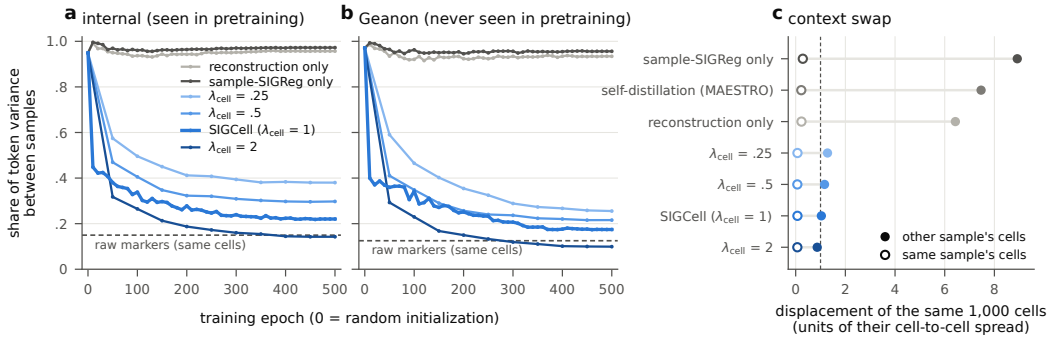

Figure 9: **The sample-context bias is present at initialization. (a, b)** Between-sample share of token variance (Equation 8) at random initialization (epoch 0) and every ten epochs of pretraining, on a fixed set of 10,000 cells from each of 48 internal and 24 Geanon samples; the dashed line is the raw markers of the same cells. Table 6 uses every sample, so its values differ in the second digit (internally, SIGCell .24 against .22 here and the raw markers .19 against .15). Epoch 0 is one random draw of the architecture, shared by all lines. **(c)** Context-swap displacement (Equation 9) with internal partner samples: filled, the same cells re-encoded among another sample's cells, the values of Tables 6 and 8; open, among other cells of their own sample. The dashed vertical line marks one cell-to-cell spread. One pretraining run per model.

here and in the cell-type read-outs. With $\bar{h}_s$ the mean token of sample s and $\bar{h}$ the grand mean, the column is the between-group term of a one-way decomposition of the total token variance, traced over the token dimensions,

$$\text{share} \;=\; \frac{\sum_s n_s \,\|\bar{h}_s - \bar{h}\|^2 \;-\; \sum_s \hat{\sigma}_s^2}{\sum_{s,i} \|h_{s,i} - \bar{h}\|^2}, \qquad \hat{\sigma}_s^2 = \frac{1}{n_s - 1} \sum_i \|h_{s,i} - \bar{h}_s\|^2, \tag{8}$$

where the second term of the numerator removes the part of the variance of the sample means that is due to averaging finitely many cells, and the ratio is clipped at 0. It is 0 when every sample draws its tokens from one common cloud and 1 when the sample alone fixes a cell's token. The raw-marker row applies the same formula to the thirty arcsinh-transformed marker values of the same cells; it is not 0, because samples differ in composition and staining, and it is the share the data carry before any encoder.

Context-swap displacement. Twenty-four internal samples are drawn at random. From each, 1,000 probe cells P and two disjoint sets C_1 and C_2 of 4,500 further cells are drawn. The encoder is run on the set $P \cup C_1$, which gives reference tokens $h_i^{(0)}$ for the probe cells and their cell-to-cell spread $\rho = \big(\frac{1}{1000}\sum_{i\in P}\|h_i^{(0)} - \bar{h}^{(0)}\|^2\big)^{1/2}$. The same probe cells are then re-encoded in the set $P \cup B$, where B holds 4,500 cells of one other sample, which gives $h_i^{(B)}$, and

$$\text{displacement} \;=\; \frac{1}{\rho}\Big(\frac{1}{1000}\sum_{i\in P}\big\|h_i^{(B)} - h_i^{(0)}\big\|^2\Big)^{1/2}, \tag{9}$$

averaged over the probe samples and two partner samples B each. Only the accompanying cells change; the probe cells' marker values do not, so a representation computed cell by cell, the raw markers included, has a displacement of 0 by construction. The table reports partners drawn from the internal cohort; partners drawn from Geanon and Penttilä give 6.4 and 6.5 under reconstruction only, 9.1 and 7.4 under self-distillation, 8.9 and 9.0 under sample-SIGReg only, 1.1 and 1.0 for SIGCell, and 0.8 and 0.7 for GPCT. Two controls fix the reading. Re-encoding the probe cells with C_2, other cells of the *same* sample, moves them by 0.23, 0.23, 0.29, 0.07 and 0.05 spreads in the same order, so the displacement comes from the change of sample and not from the change of cells. And the cosine between the mean displacement $\frac{1}{1000}\sum_{i\in P}(h_i^{(B)} - h_i^{(0)})$ and the shift of the context cells' own mean token, $\bar{h}_B - \bar{h}_{C_1}$, is .98–.99 without the cell term, .73 with it and .59 for GPCT: the probe cells are carried to wherever the new context's cells land, which is what "toward the token cloud of the new context" means in Section 5.4.

Donor-half separation. This column is computed on the healthy-donor token caches: 6,000 selected cells of 45 healthy internal donors, one sample each, encoded in 10,000-cell contexts and split into two donor-disjoint halves of 3,048 and 2,952 cells. Within each of five shared cell types (B, CD4 T, CD8 T, monocyte, neutrophil), an equal number of cells is drawn from each half, and for every cell the fraction of its fifteen nearest neighbors in the native token space, among cells of the same cell type, that come from its own half is recorded as its purity. With two balanced groups chance is .5, and separation is $(\text{purity} - \text{chance})/(1 - \text{chance})$, averaged over cell types with weights proportional to their cell counts: 0 when the cells of a cell type from the two donor groups are fully mixed, 1 when they form two islands. Both halves are healthy and the neighborhood is restricted to one cell type, so separation above the raw-marker floor of .20 is structure the encoder adds that sorts cells by the donors they came from. The Geanon values quoted above come from the same computation with the two cohorts in place of the two halves, the same 6,000 internal cells against 6,000 cells of the 15 healthy Geanon donors, so they measure cohort separation, not a split inside Geanon.

The same construction on the two external cohorts is Figures 7 and 8, and it is the part that matters for transfer, because no encoder saw a cell of either cohort. The draw differs in one respect: those cohorts have too few healthy donors to sample thirty, so their thirty donors are drawn from all their samples and differ in disease state as well as in identity. Even so, the three conditions without the cell term still place *every* one of a cell's fifteen nearest neighbors in its own donor, 1.00 on both cohorts as on the internal one, so the tokens are sorting by donor and not by disease. With the cell term the figure falls to .46 and .47, lower than the .64 it reaches internally, while same-type agreement rises to .71 and .75 against .60–.68 without it. Raw markers sit at .22 and .20. On cohorts no encoder saw in training, the sample bias is thus present in the published encoder too, and only the cell term removes it. A pointwise encoder has no sample bias, so it transfers cell-type read-outs trivially, but it also lacks the set-level information that makes set encoders strong in distribution.

The cell-level effect across the design space. Every condition without the cell term falls far below every condition with it on the source-fitted cell-type read-out: .13–.19 against .64–.69 on Geanon and .19–.57 against .76–.79 on Penttilä, where chance is .071.

Centering recovers the transfer. What fails without the cell term is the transfer of a read-out across samples, because each sample has its own coordinate frame: subtracting a sample's own token mean, a label-free diagnostic, recovers a large part of the loss. With the cell term the source-fitted and the centered read-outs coincide, on unseen cohorts as well.

Three read-outs on the same cells. Table 7 repeats the cell-type read-out of Table 8 and scores the same cells two more ways: after every sample's own mean token has been subtracted (*centered*), and with a classifier fitted on the other cells of the same sample (*inside one sample*). First, the cell term does not add cell-type information: inside one sample the read-out is .80 without it and .76–.77 with it internally, .62 against .59 on Geanon and .59 against .56–.57 on Penttilä. Second, what it raises is the read-out across samples, monotonically in λ_{cell} on all three cohorts: from .18 at $\lambda_{\text{cell}} = 0$ to .62, .73, .77 and .83 internally, from .13 to .72 on Geanon and from .19 to .80 on Penttilä. Third, without the cell term subtracting each sample's mean recovers .26 to .36 of read-out yet stays below what every cell-term weight reaches, whereas with the cell term the centered and across-sample read-outs differ by at most .08; from $\lambda_{\text{cell}} = 1$ on, the classifier fitted on other samples matches or beats one fitted on the sample's own cells (.77 against .77 internally, .64 against .59 on Geanon, .76 against .56 on Penttilä), because in a shared frame it can pool the rare cell types of many samples. The cell term thus puts the tokens of different samples into one coordinate frame; a sample's mean offset is a large part, though not all, of what separated the frames without it.

Table 7: **Cell-type read-out at three levels.** The cells, source split, classifier and metric of Table 8: balanced accuracy over the fourteen gated types on the selected cells, a mean ± SD over ten split seeds, each of which re-draws the cell selection and the grouped source half. *Across samples*: the classifier fitted on the internal source half and applied unchanged; these rows are the cell-type read-out of Table 8. *Centered*: a classifier of the same form fitted and applied after every sample's own mean token is subtracted, a label-free diagnostic rather than a proposed method. *Inside one sample*: a classifier fitted on the same sample's other cells, never on the scored ones. Because the scored cells are drawn per type across a cohort, rare types weigh as much as common ones while one sample holds few of them, so the inside read-out is limited by a single sample's data and is not an upper bound.

	reconstruction	λ_{cell}, at $\lambda_{\text{sample}} = .05$				
	only	0	.25	.5	1 (SIGCell)	2
internal						
across samples	.229±.016	.176±.015	.618±.009	.726±.012	.770±.014	.826±.009
centered	.578±.011	.536±.008	.697±.010	.767±.011	.785±.010	.812±.009
inside one sample	.803±.001	.802±.001	.758±.002	.768±.002	.769±.002	.771±.002
Geanon						
across samples	.164±.008	.133±.011	.510±.026	.601±.019	.635±.016	.721±.009
centered	.426±.020	.397±.016	.567±.010	.640±.014	.684±.012	.707±.010
inside one sample	.618±.001	.618±.001	.588±.001	.589±.001	.587±.001	.592±.001
Penttilä						
across samples	.260±.018	.191±.014	.643±.022	.732±.008	.757±.006	.801±.010
centered	.589±.010	.550±.011	.677±.008	.757±.008	.765±.005	.790±.009
inside one sample	.588±.001	.586±.001	.557±.001	.563±.001	.563±.001	.568±.001

E THE WEIGHT SWEEPS IN NUMBERS

Figure 10 draws the two sweeps; this section gives the numbers behind them, in Table 8 for the cell weight and Table 9 for the sample weight, together with the paired changes that Figure 11 shows. The two tables also report what each weight does to the tokens, which the figure does not show. Figure 11 gives what each term is worth *given* the other, seed by seed. With the sample term already on, turning the cell term on is worth $+.105$ balanced accuracy on severity grading and $+.129$ and $+.091$ on the two transfer cohorts; with the sample term off it is worth $-.037$, $+.020$ and $+.020$, so most of its value appears only in the other term's presence. The sample term on its own costs balanced accuracy on all three external read-outs, $-.062$, $-.049$ and $-.036$, while gaining $+.015$ in

distribution. The interaction is positive on all three external columns, $+.141$, $+.109$ and $+.071$, and slightly negative internally, so the pair's effect is not the sum of the two terms' effects.

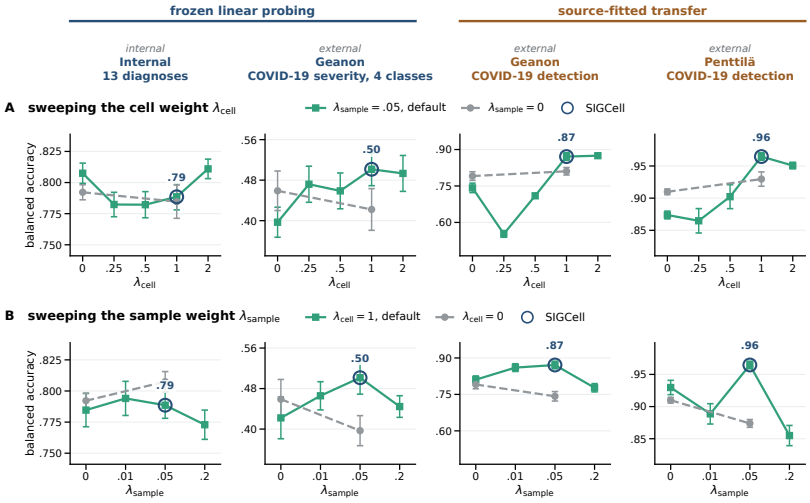

Figure 10: **What each weight does to the sample-level read-outs.** Balanced accuracy against the swept weight, mean $\pm$ SD over ten split seeds. In **A** the cell weight is swept with the sample weight held at its default (green) or at zero (gray); **B** exchanges the roles, so the gap between the lines is the effect of the *held* term and each line's shape the effect of the swept one; the ring marks SIGCell, fixed in advance. Y-ranges differ by column: read the shape and the gap, not the height. Appendix E gives the numbers and the paired changes.

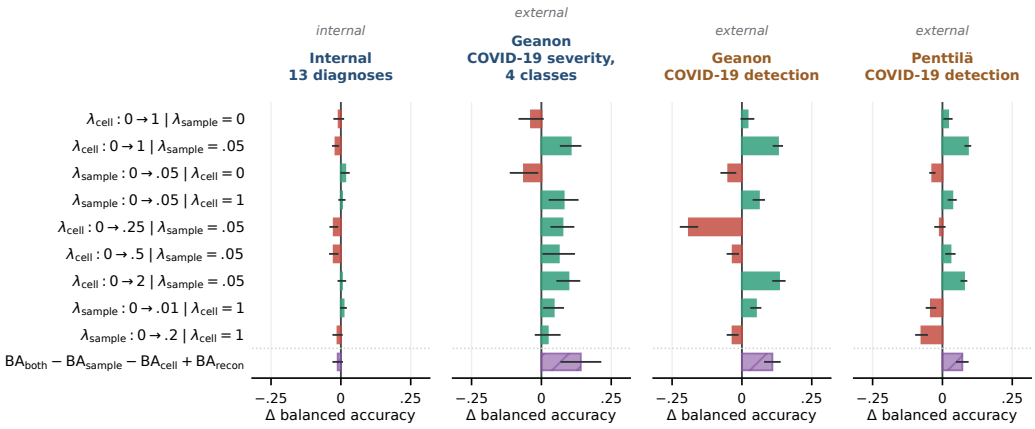

Figure 11: **Paired conditional changes.** Each bar turns one weight on while the other is held where the label says, and is paired seed by seed, so the bar is the mean of ten differences and the whisker their SD. Green is a gain, red a loss. The last row is the interaction between the two terms at the default setting, $\mathrm{BA}_{\mathrm{both}} - \mathrm{BA}_{\mathrm{sample}} - \mathrm{BA}_{\mathrm{cell}} + \mathrm{BA}_{\mathrm{recon}}$, the balanced accuracy with both terms, less that with the sample term alone and with the cell term alone, plus that with reconstruction alone: it is positive on every external column, which is the statement that the two terms are not additive there. One shared Δ scale across the columns.

The default is the best of the four on the severity probe and on both source-fitted columns, and a larger weight harms every read-out, including the cell-type read-out, which falls by 12 to 16 points on the three cohorts. Turning the term off entirely is the best setting for the internal cell-type read-out. Unlike the cell weight, the sample weight leaves the token geometry where it is, consistent

Table 8: **Cell-weight sweep**, at the default $\lambda_{\text{sample}} = .05$. Every read-out is reported as balanced accuracy so that the blocks are comparable, each a mean ± SD over ten split seeds, with the best value in a row in **bold**; the last block is the token geometry of Table 6, defined in Appendix D, which is not an accuracy and where lower is better down to the raw-marker level of Table 6; the between-sample share, which falls below that level at $\lambda_{\text{cell}} = 2$, carries no bold. The flagship value was fixed before any external cohort was scored, so this is a sensitivity analysis, not a selection. The two intermediate sweep points are read from a second extraction of the same checkpoints, which agrees with the first to one sample where both exist. λ_{sample} is swept in Table 9, and Figure 10 draws both sweeps.

λ_{cell}	0	.25	.5	1 (SIGCell)	2
frozen linear probing (balanced accuracy)					
internal, 13 diagnoses	.807±.008	.782±.010	.782±.011	.789±.011	**.811**±.008
Geanon, COVID-19 severity	.397±.030	.472±.036	.459±.035	**.501**±.032	.493±.035
source-fitted transfer (balanced accuracy)					
Geanon	.742±.020	.553±.014	.710±.010	.871±.016	**.875**±.008
Penttilä	.874±.006	.865±.019	.902±.019	**.965**±.006	.951±.005
cell-type read-out (balanced accuracy)					
internal	.176±.015	.618±.009	.726±.012	.770±.014	**.826**±.009
Geanon	.133±.011	.510±.026	.601±.019	.635±.016	**.721**±.009
Penttilä	.191±.014	.643±.022	.732±.008	.757±.006	**.801**±.010
token geometry (not an accuracy; lower is better down to the raw markers)					
between-sample share	.98	.40	.31	.24	.15
context-swap displacement	8.9	1.3	1.2	1.0	**0.9**

Table 9: **Sample-weight sweep**, at $\lambda_{\text{cell}} = 1$. Laid out like Table 8: every read-out is balanced accuracy so that the blocks are comparable, each a mean ± SD over ten split seeds, with the best value in a row in **bold**; the last block is the token geometry of Table 6, defined in Appendix D, which is not an accuracy and where lower is better. The default .05 is the value we kept, and it is the best of the four on three of the seven accuracy rows, including both transfer cohorts. The geometry block carries no bold because the sample weight barely moves it: the between-sample share stays inside .21–.25 and the displacement inside 1.0–1.2 across the whole sweep, against .15–.98 and 0.9–8.9 for the cell weight in Table 8. Figure 10 draws both sweeps.

λ_{sample}	0	.01	.05 (SIGCell)	.2
frozen linear probing (balanced accuracy)				
internal, 13 diagnoses	.785±.013	**.794**±.014	.789±.011	.773±.012
Geanon, COVID-19 severity	.422±.041	.466±.028	**.501**±.032	.445±.021
source-fitted transfer (balanced accuracy)				
Geanon	.811±.015	.860±.016	**.871**±.016	.777±.017
Penttilä	.930±.011	.889±.016	**.965**±.006	.855±.016
cell-type read-out (balanced accuracy)				
internal	**.801**±.011	.794±.011	.770±.014	.609±.011
Geanon	**.691**±.010	.685±.015	.635±.016	.519±.015
Penttilä	.787±.007	**.803**±.008	.757±.006	.610±.013
token geometry (not an accuracy; lower is better)				
between-sample share	.22	.25	.24	.21
context-swap displacement	1.0	1.0	1.0	1.2

with the division of labor Section 6 describes. The term is therefore kept at its published default rather than tuned.

How Figure 10 is drawn. Each column is headed by its cohort and its task, and by whether that cohort is the internal one or an external one: the first two columns are the frozen linear probing of

Section 5.1 and the last two the source-fitted transfer of Section 5.2, so the second and third columns are the *same* external cohort under two different tasks, not two cohorts. Penttilä's deposited classes get no frozen linear probing column, for the reason given in Section 4. Tick positions are equally spaced, the lines guide the eye, and only observed combinations are drawn.

Frozen linear probing under the cell weight. The severity probe is flat or noisy between .25 and .5, peaks at the default and falls slightly at 2, while the internal probe is highest at the two ends.

The sample term's external cost, by metric. Sample-SIGReg alone is harmful on every external read-out, $-.062$ severity balanced accuracy and $-.008$ and $-.029$ source-fitted AUROC, while being the best condition in distribution.

Proposition 1 across the cell weight. In the training layout, with batch normalization over each GPU's three samples, the tested cloud keeps .077, .047, .030 and .018 of its variance between samples as λ_{cell} goes $.25 \to 2$, below the bound of $8/64 = .125$; standardizing the twelve samples of a batch together instead gives .107, .064, .042 and .024, below $11/64$. The measurement is described in Appendix B (Table 4).

F SAMPLE EMBEDDING GEOMETRY

Figure 12 lays out the sample embeddings of the six models of Table 1, the ones whose sample representation is fixed by their own design, one column per model and one row per cohort, colored by that cohort's own classes. The last column is manual gating, where the sample is nothing but the frequencies of its fourteen gated cell types. A UMAP is a neighbor-preserving picture and not a metric: cluster sizes, gaps and the distances between clusters are artifacts of the layout. The number under each panel is therefore computed in the full z-scored latent and not on the coordinates. It is the 10-nearest-neighbor label purity, the mean fraction of a sample's ten nearest neighbors carrying its own class, with every other sample of the same donor blocked from being a neighbor and the chance level recomputed under that blocking.

The internal column orders the models the way Table 1 does, from .73 down to .34 against a chance level of .10, with one exception: on this statistic SIGCell is above the self-distilled encoder (.73 against .70) while under frozen linear probing it is slightly below (.789 against .791 balanced accuracy). The two statistics measure different things, and they disagree within the variation of either. Manual gating is last at .34, and its panel is the one to look at before reading anything into a layout: thirteen diagnoses are drawn on top of each other because fourteen frequencies cannot hold them apart, which is the same fact its row of Table 1 reports as .488 balanced accuracy. On both external cohorts every model sits between .39 and .48 against chance levels of .28 and .29, a band narrow enough that the layout should not be read as a ranking; the severity column of Table 1 is where that comparison is made.

G THE MODELS COMPARED

Every pretrained model shares one convention (Section 4, Table 12): one corpus of 1,872 unlabeled samples on one 30-marker panel, 10,000 cells drawn per sample per step, 500 epochs, effective batch 12, bf16. This section gives what that convention was applied to. Appendix I states how each implementation differs from its paper. Table 10 places the pretrained encoders of Section 2 on its two design axes and states which of them are compared here.

The compared models. Tables 11–13 give sizes and training settings; here each model is described as its paper defines it, and Appendix H states how it is adapted to the transfer task.

Pretrained encoders. MAESTRO (Lee et al., 2025) encodes a sample's cells with induced set-attention blocks and pools them into one sample embedding. It is pretrained by masking cells in similarity-ordered blocks and reconstructing the whole set from the pooled embedding under an optimal-transport loss, together with self-distillation between a student that sees a subset of the cells and an EMA teacher that encodes the full set. cyMAE (Kim et al., 2024) is a masked autoencoder over the markers of one cell: each marker is a token, and pretraining reconstructs masked marker

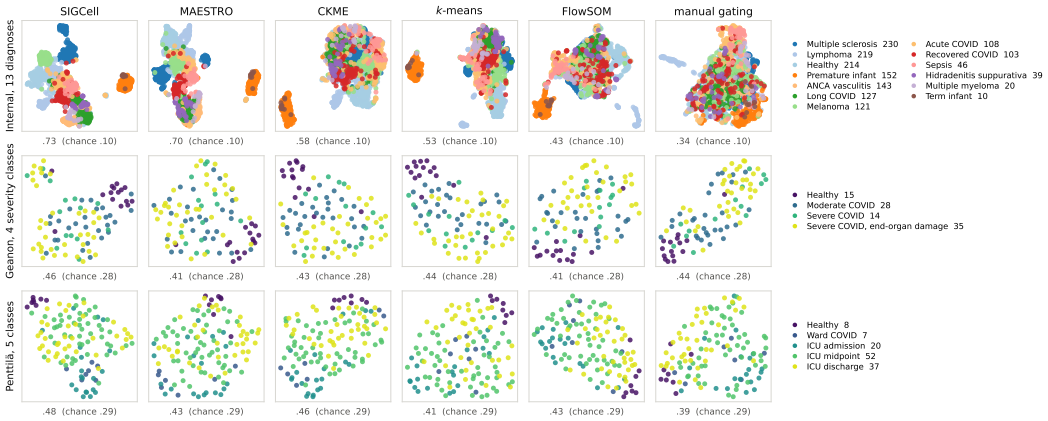

Figure 12: **Sample embeddings, per cohort and per model.** Rows are cohorts, columns are the six models of Table 1 that have a sample embedding of their own, manual gating's being the frequencies of its fourteen gated cell types; the other pretrained models form theirs with each task's labels, so they are absent here, as they are from Table 1. Points are samples, colored by the cohort's own classes, and the two external cohorts' ordered classes use one ramp in clinical order. The internal row is the 1,532 labeled samples that the thirteen-diagnosis probe uses, with the technical controls set aside. Under each panel: 10-nearest-neighbor label purity computed in the full z-scored latent, not on the layout, with same-donor neighbors blocked, and the chance level under that blocking in brackets.

Table 10: **Pretrained cytometry encoders on the two design axes of Section 2.** *Cells*: whether a cell is encoded in the context of the other cells of its sample; EventHorizon first encodes each cell from its own markers and then lets the cells of a sample attend to one another. *Sample representation*: fixed by pretraining, or formed with the labels of each task, and how. Only a representation fixed by pretraining enters frozen linear probing (Table 1); every compared model enters source-fitted transfer in its own adaptation regime (Table 2, Appendix H). DinoFlow (O'Fallon et al., 2026), which also pretrains a tube-level embedding and has released no code, is omitted because we could not establish how its transformer combines cells.

model	cells	sample representation	in this paper
SIGCell	in sample context	fixed by pretraining	Tables 1 and 2
MAESTRO	in sample context	fixed by pretraining	Tables 1 and 2
EventHorizon	in sample context	fixed by pretraining	not compared: no released code
GPCT	in sample context	decoder trained per task	Table 2
cyMAE	alone	pooling rule selected per task	Table 2
CytoBERT	alone	encoder fine-tuned per task	Table 2
ImmuneFM	alone	encoder fine-tuned per task	not compared: see Section 2

values from the others; a sample is summarized by pooling its cell embeddings. GPCT (Zhuang et al., 2026) embeds each cell as a token that also records which markers were measured, lets the cells of a sample attend to one another, and pretrains by imputing masked marker values together with their within-sample percentile ranks and log densities; a sample-level prediction comes from a decoder in which a class token cross-attends to the cell tokens, trained for each task. CytoBERT (Qadri et al., 2026) treats each cell independently as a sequence of marker–expression tokens with discretized expression and pretrains by predicting masked expression bins; a class token gives a cell-level representation, and the encoder is fine-tuned for each task.

Training-free baselines. CKME (Shan et al., 2022) represents a sample by the kernel mean embedding of its cell distribution, approximated with random Fourier features of a radial basis function kernel, so the sample vector is the average of fixed random features over its cells; a linear classifier on it is then the average of per-cell scores, which is what makes the model transparent. k-means and

Table 11: **What each pretrained model is.** Parameter counts were obtained by instantiating each model from its own recorded hyperparameters and summing its parameters, and cross-checked against its checkpoint. The two columns separate the training-time network from the part kept after pretraining: SIGCell's reconstruction decoder and cell-SIGReg projection head, MAESTRO's decoder, projection head and its entire exponential-moving-average (EMA) teacher, and the masked-prediction heads of cyMAE, GPCT and CytoBERT all exist only during pretraining. The kept part stays frozen in every model except CytoBERT, which fine-tunes it end to end (Table 14). CytoBERT's masked-prediction head holds 1,290 of its 1,196,298 parameters, so both counts round to 1.20M. *Read-out* is the width of the vector read by the part fitted for each task: the sample embedding for SIGCell and MAESTRO, the flattened marker tokens max-pooled over cells for cyMAE, each cell token for GPCT and the ℓ_2-normalized mean CLS state for CytoBERT. SIGCell's kept part ends in an affine layer after pooling (65,792 parameters), which gives the embedding the unconstrained location and scale that SIGReg acts on.

		parameters		read-out
model	backbone	training	kept	dim
SIGCell	5 ISABs, width 384, 16 inducing points	16.2M	9.46M	256
MAESTRO	5 ISABs, width 384, 16 inducing points; EMA teacher tower	32.5M	9.39M	256
cyMAE	6 layers, width 30, 6 heads; marker tokens	74k	68k	900
GPCT	4 layers, width 128, 4 heads; cell tokens	612k	538k	128
CytoBERT	6 layers, width 128, 8 heads; binned markers	1.20M	1.20M	128

Table 12: **How each pretrained model was trained.** All five ran 500 epochs over the same 1,872 samples at effective batch 12, which is 78,000 optimizer steps, in bf16 mixed precision at seed 206.

model	optimizer	learning rate	schedule	decay
SIGCell	AdamW	1×10^{-4}	cosine, no warmup	10^{-3}
MAESTRO	AdamW	1×10^{-4}	cosine, no warmup	10^{-3}
cyMAE	AdamW	1.5×10^{-4}	40 warmup epochs, cosine	.05
GPCT	AdamW	5×10^{-4}	5 warmup epochs, linear	5×10^{-3}
CytoBERT	AdamW	1×10^{-4}	none	.01

FlowSOM (Van Gassen et al., 2015) cluster the cells of the corpus without labels, FlowSOM with a self-organizing map followed by consensus metaclustering, and represent a sample by the proportion of its cells in each cluster. Manual gating is the standard analysis in immunology: an expert assigns cells to populations through a sequence of thresholds on marker intensities, and a sample is the vector of its population frequencies.

Models trained from scratch. CellCnn (Arvaniti & Claassen, 2017) applies a bank of learned filters to every cell and pools the responses over a random subset of cells, so that a small responsive population can drive a sample-level prediction; it was designed to detect rare disease-associated subsets. CytoSet (Yi & Stanley, 2021) is a permutation-invariant set network that predicts clinical outcomes from random subsets of a sample's cells. CytoDx (Hu et al., 2019) is a gating-free two-stage linear model: a cell-level model in which every cell inherits its sample's label, then a sample-level model on the average cell prediction. Deep Sets (Zaheer et al., 2017) maps each element with one network, averages, and maps the average with a second network; Set Transformer (Lee et al., 2019) replaces the average with attention through induced set-attention blocks and pooling by multihead attention, the architecture MAESTRO and SIGCell build on. MAESTRO used both as baselines for cytometry.

Feature widths of the training-free baselines. CKME represents a sample by the empirical mean over its cells of $\cos$ and $\sin$ of 1,000 frozen random projections, a 2,000-dimensional vector; k-means by the proportions of its cells in 100 clusters; FlowSOM by the proportions in 40 metaclusters; manual gating by the proportions of 14 gates. The clusterings are fitted once on the corpus without any label, and the probe is the only thing refitted inside a fold.

Table 13: **The baselines.** The first block is training-free: nothing is learned by gradient descent, and the only quantities carried are a frozen random draw or a clustering fitted once on the corpus without labels, so the number that matters is the width of the composition vector a probe reads. Manual gating stores nothing at all: the sample is the frequency vector of the gates the files already carry. The second block is trained from scratch on each task's own labels, so its parameter counts depend on the number of classes; the counts given are for the two-class transfer task of Table 2, and its cost is the mean fit over the ten split seeds. Set Transformer is MAESTRO's encoder at random initialization, which is why it has about twenty times the parameters and fitting time of Deep Sets, and more than a thousand times the parameters and a hundred times the fitting time of the other three.

model	what is fitted	parameters	cost
CKME	nothing; 1,000 frozen random features + a bandwidth	0	one CPU pass
k-means	100 centroids, once, without labels	0	CPU
FlowSOM	a 10×10 map, then 40 metaclusters	0	CPU
manual gating	nothing; the gates the files already carry	0	none
CytoDx	two stacked linear models, per task	68	0.2 min
CellCnn	an ensemble of one-layer filter networks, per task	435–732	0.4 min
CytoSet	one permutation-equivariant block, per task	8.2k	0.3 min
Deep Sets	two MLPs around a mean, widths $384/256$, per task	506k	2.2 min
Set Transformer	MAESTRO's encoder from random init, per task	9.39M	49.8 min

Hardware. All pretraining ran on one node with four NVIDIA A100-SXM4-80GB GPUs, three samples per GPU (the SIGCell run of record kept no log; its device is inferred from sibling runs on the same node). The supervised baselines of Table 13 were fitted on the same node, and the training-free baselines run on CPU.

Comparison models. We do not use released pretrained checkpoints. Instead, we retrain each comparison model from the authors' released implementation on the same corpus and marker panel. All pretrained models use 500 epochs and an effective batch size of 12 samples. Where paper and released code differ, we follow the paper, except for MAESTRO's backbone (Appendix I), and otherwise keep the released defaults. This controls the pretraining data, marker panel, and training schedule across models. Appendix I lists all remaining deviations from each model's original setup.

Splits and probes. For every within-dataset evaluation, we use five-fold cross-validation grouped by donor, so that all samples from the same donor remain in the same fold. We repeat the split with ten random seeds and use identical folds for every model. The probe consists of a standard scaler followed by ℓ_2-regularized multinomial logistic regression with balanced class weights and $C = 1$; both scaler and classifier are fit using training folds only. Metrics are computed from the pooled out-of-fold predictions for each seed.

Source-fitted transfer. For the frozen encoders, each split seed divides the 322 internal samples of the COVID-19 detection task (Appendix C) into five donor-grouped folds. The scaler and classifier are fitted on each training part, about four fifths of the samples; each of the five rules is applied to the whole of both external datasets, and the seed's value is the mean of the five rules' metrics. The models that train a network select their number of epochs on such folds and are then refit on all 322 samples (Appendix H). The decision threshold is fixed in advance at probability .5, and all fitted quantities are frozen before either external dataset is scored; no external label is used for fitting or recalibration. Fitting the frozen rows once on all 322 samples instead, as the rows that train a network are fitted, gives a deterministic rule with no variation over seeds and leaves SIGCell first in every column of Table 2: .958 AUROC and .888 balanced accuracy on Geanon, .998 and .987 on Penttilä.

Cell budgets. Frozen representations are extracted from up to 100,000 cells per sample, averaged over five draws at a fixed seed. The tuned models are trained on a shared 20,000-cell pool per sample

and, like the frozen representations, scored on up to $100{,}000$ cells of each test sample. The cell-type read-out runs at $10{,}000$-cell contexts.

Reporting. Unless stated otherwise, predictive results are reported as the mean and standard deviation over ten seeds. The source of variation differs by evaluation: seeds define dataset folds for frozen linear probing, source folds for source-fitted transfer, independent downstream training runs for methods that retrain a network, and cell subsampling together with the grouped source split for cell-type read-outs. Contrasts are computed as seed-wise paired differences and reported with their own standard deviation.

What a seed re-partitions, in full. Frozen linear probing re-partitions the evaluated cohort into five donor-grouped folds. The source-fitted transfer re-partitions the *source* into five donor-grouped folds, fits its rule on each training part, applies it to the whole external cohort with the threshold untouched and averages the five rules' metrics, so its SD says how much the rule depends on which source samples were available; the rows that train a network repeat the training itself at each seed. The cell-type read-out re-draws both the cell selection and the grouped source half.

H What is frozen and what is fitted with labels in Table 2

Table 14 states, for every row of Table 2, which part of the model is carried over from pretraining untouched and which part is fitted on the 322 labeled source samples; the paragraphs that follow give each recipe. Two things hold for every row and are stated once. The fitted quantity is a COVID-19 detection rule on the source, applied to each external cohort as it is: the decision threshold is a probability of $.5$, nothing on the target side is standardized, recalibrated or selected on target labels, and the external files are opened only after the fitted model has been saved. Where a row trains a network, except CellCnn, CytoSet and CytoDx, which keep their reference model selection (below), the number of training epochs is the only quantity selected: it is the epoch that minimizes the mean validation balanced log-loss over donor-grouped five-fold cross-validation on the source, after which the network is refit on all source samples for that many epochs; every other hyperparameter is fixed in advance by the model's own recipe. The loss of these networks is class-balanced cross-entropy; CellCnn and CytoSet balance the classes by drawing the same number of training subsets per class, and CytoDx is unweighted. Every model is scored on up to $100{,}000$ test cells per sample, drawn in its own context size, with the predicted probabilities averaged over the draws (Appendix G).

SIGCell and MAESTRO. Nothing pretrained is touched. Each sample is encoded by the frozen network into its 256-d embedding, from the cell budget of Appendix G, and the only fitted objects are a standard scaler and an ℓ_2-regularized logistic regression with balanced class weights and $C = 1$, the probe of Section 4: 257 parameters for the two-class task. For SIGCell the embedding is the output of the affine layer after pooling; for MAESTRO it is the pooled output of the student encoder, the quantity its paper reads with a linear probe. A split seed re-partitions the source into five donor-grouped folds and refits the probe on each training part, which is the only source of the row's SD.

cyMAE. The pretrained cell encoder is frozen. Following its paper's subject representation, each cell's thirty marker tokens of width 30 are flattened into one 900-d vector, and the sample is the element-wise maximum of that vector over its cells; the same scaler and logistic regression are fitted on it. Max was chosen over mean and min pooling by donor-grouped cross-validation of the source task alone, before any external file was read.

GPCT. The pretrained encoder stays frozen, as its paper recommends when labeled data are limited: the input projection, the mask embedding and the four encoder layers are loaded, their parameters excluded from the optimizer, and the encoder's dropout switched off so that it is deterministic. What is trained on the source labels is GPCT's own classification module on top of the frozen cell tokens of a $10{,}000$-cell context: a learned class token, a three-layer cross-attention decoder with eight heads and dropout $.2$ in which the class token attends to the cell tokens, and a three-layer MLP on the class token. During training a random 15% of markers are dropped per set, the paper's marker drop-out. AdamW with learning rate 2×10^{-5}, weight decay 10^{-3} and eight samples per step; at most 50 epochs, the count selected as above and re-selected at every split seed, after which the module is refit on all source samples.

Table 14: **Frozen versus fitted, row by row.** What each model of Table 2 keeps from pretraining and what it fits on the labeled source samples. The first block adapts a pretrained model; the second has no pretraining and fits everything.

model	frozen (from pretraining)	fitted on the source labels	training
SIGCell	the whole network: encoder blocks, pooling and the affine layer, so the 256-d sample embedding	a standard scaler and a class-balanced ℓ_2 logistic regression on the 256-d embedding	convex fit; no epoch selection
MAESTRO	the whole student encoder and its pooling, so the 256-d sample embedding	the same scaler and logistic regression	convex fit; no epoch selection
cyMAE	the cell encoder; each cell's thirty 30-d marker tokens flattened to 900-d and max-pooled over the sample's cells	the same scaler and logistic regression on the 900-d vector	convex fit; the pool (max) chosen by source cross-validation
GPCT	input projection, mask embedding and the four-layer encoder, dropout off	GPCT's classification module: a class token, a three-layer cross-attention decoder over the frozen cell tokens, and an MLP	AdamW, 10,000-cell contexts, epochs selected
CytoBERT	nothing	the whole encoder, fine-tuned end to end, plus a new head on the ℓ_2-normalized mean CLS state of 512 cells	AdamW, epochs selected
CellCnn	no pretraining	an ensemble of one-layer filter networks on 200-cell subsets	its own search and early stopping
CytoSet	no pretraining	one permutation-equivariant block and a two-layer decoder on 200-cell subsets	Adam, its own early stopping
CytoDx	no pretraining	a cell-level logistic regression, then a sample-level one on the mean cell probability	unregularized; nothing selected
Deep Sets	no pretraining	a per-cell MLP, mean over 10,000 cells, an MLP and a linear head, end to end	AdamW, epochs selected
Set Transformer	no pretraining	MAESTRO's encoder and pooling from random initialization, plus one linear layer, end to end	AdamW, epochs selected

CytoBERT. Nothing is frozen. Following its released fine-tuning script, the pretrained encoder is trained end to end together with a new head: for each of 512 cells drawn from the sample the CLS state of the final layer is taken, the 512 states are averaged, the mean is ℓ_2-normalized, and a Linear(128)–ReLU–Linear(2) head gives the logits. AdamW with learning rate 10^{-4} and weight decay .01, one sample per step, at most 50 epochs with the count selected as above; at test time the 100,000 cells of a sample are read in 196 draws of 512.

The models without pretraining. CellCnn, CytoSet and CytoDx run as ports of their reference implementations and keep their own model selection, which uses only the training samples. CellCnn fits an ensemble of one-layer filter networks (a per-cell linear filter bank, ReLU, top-k pooling over the cells of a 200-cell subset, softmax) on 1,000 subsets per class, trying 15 random configurations of filter count, pooling percentage and learning rate, each for at most 20 epochs with early stopping on a cell-level validation split of the training samples, and predicts with the three best by validation accuracy. CytoSet fits one permutation-equivariant block of width 64 with max pooling and a two-layer decoder on 200-cell subsets, Adam at 10^{-4}, at most 200 epochs with the same kind of early stopping. CytoDx fits two unregularized logistic regressions: a cell-level one in which every cell carries its sample's label (2,000 cells per sample), then a sample-level one on the mean cell

probability; nothing is selected. Deep Sets and Set Transformer follow MAESTRO's recipe for its from-scratch baselines: a mean-pooling network of two MLPs of widths $384/256$, or MAESTRO's own encoder and pooling at random initialization with one linear layer on the 256-d output, trained end to end on $10{,}000$-cell draws with AdamW at 10^{-4}, weight decay $.01$, four samples per step and at most 300 epochs, the count selected as above with patience 20 and refit on all source samples. Because these rows train their networks anew at every split seed, their SD in Table 2 includes the training itself, which the frozen rows' does not.

I IMPLEMENTATION DEVIATIONS

Four kinds of difference separate our comparison models from their papers, and one separates SIGCell's objective from SIGReg's; all five are deliberate or unavoidable.

Shared by construction. Every pretrained model is trained from scratch on this corpus and this panel for 500 epochs at effective batch 12, which is $78{,}000$ optimizer steps, instead of on its own corpus at its own budget.

Released code versus paper. Where a released implementation's defaults differ from its paper's text, we follow the paper: cyMAE's masking ratio and drop path and GPCT's depth, width and learning rate are the published values rather than the released code's. Those are the checkpoints of record, they are the rows of Table 2, and each ran its full 500 epochs at the shared budget.

MAESTRO's backbone. MAESTRO is the one exception to that rule. Its paper describes three ISABs in the encoder, three SABs in the decoder and a $1{,}024$-dimensional sample embedding (Lee et al., 2025); its released code has five encoder blocks, two decoder SABs, width 384 and a 256-dimensional embedding. We retrain MAESTRO in its released configuration.

The shared adaptation harness. Section 5.2 holds fixed everything outside each model's adaptation regime, from the source roster, donor groups and cell pools to the test cell budget. Where this departs from the papers, it departs in the same way for GPCT, CytoBERT, Deep Sets and Set Transformer: the number of epochs is chosen by donor-grouped cross-validation on the source with a shared criterion, the objective is class-balanced, predictions are averaged over up to $100{,}000$ test cells, and no learning-rate schedule is used. The individual papers use early stopping with their own patience, oversampling or plain cross-entropy, a single draw at test time, and their own schedules. CellCnn, CytoSet and CytoDx keep their reference model selection and objectives (Appendix H); for them the harness changes only the source roster, the cell pools and the averaging of predictions over test draws. The regimes themselves run from a linear layer on a frozen encoder through a trained cross-attention decoder to end-to-end fine-tuning. We therefore describe these rows as each model in its own adaptation regime inside a shared harness, and not as each model in its published protocol.

SIGReg's calibration at the cell site. SIGCell's own objective departs from SIGReg's at one point: the scale factor of the cell statistic is pinned to the effective batch, $n = 12$, instead of the number of points it tests, for the reason given in Appendix B.