INTERPRETING SPARSE AUTOENCODER FEATURES IN CYTOMETRY FOUNDATION MODELS THROUGH IN SILICO PERTURBATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Cytometry foundation models learn sample-level representations for patient-level prediction, yet how they encode immune cell composition remains unclear. Sparse autoencoders (SAEs) expose individual features in these representations, but do not by themselves explain which composition changes those features respond to. We introduce a perturbation-based framework that interprets SAE features through controlled exchanges of cell populations. Resampling a sample's own cells varies a target population within a specified group while holding cells outside that group fixed. Each annotation specifies the populations exchanged and the direction of the feature's response, making its interpretation explicit and testable. Across GPCT and MAESTRO, more than 98% of annotations retain their response direction with corrected significance on held-out samples. Annotations of case-enriched features recover 82% and 65% of expert-expected population changes, respectively, across 20 case–control comparisons, while MAESTRO decoder steering reproduces 93.2% of annotated directions. The annotations also support classification from feature descriptions alone. Beyond individual annotations, comparing perturbation responses between cases and controls reveals shared-feature programs linking populations through state-dependent model responses. Together, this framework connects learned features to interpretable composition changes and provides a way to examine how cytometry models represent immune variation across samples and disease states.

1 INTRODUCTION

Cytometry measures tens of proteins per cell across hundreds of thousands of cells in a single blood sample and is widely used to characterize the immune system in health and disease (Spitzer & Nolan, 2016; Saeys et al., 2016). A sample is thus an unordered set of cells, distinguished from other samples largely by which populations it contains and in what proportions. Pretrained cytometry models such as MAESTRO and GPCT take this set as input and learn sample-level representations that perform well on downstream patient-level prediction tasks (Lee et al., 2025; Zhuang et al., 2026). Yet their predictive success leaves open what they learn about cell composition and how that information supports downstream predictions. Fixed population gates describe a sample's composition directly, but do not explain how a pretrained model represents it. Understanding what makes these representations useful therefore calls for interpreting the model's response to changes in composition, which we measure through controlled in silico perturbations.

To study these representations, we use sparse autoencoders (SAEs), which reconstruct a latent vector from a sparse combination of learned feature directions. SAEs have been used to interpret language and vision models (Bricken et al., 2023; Huben et al., 2024; Gao et al., 2024) and, increasingly, models of biological data (Simon & Zou, 2025; Claye et al., 2025). However, the resulting features are not immediately human-readable: the decomposition alone does not identify what changes in the input each feature responds to.

For cytometry, we seek to interpret these features in terms of cell composition. Summarizing strongly activating samples (Bills et al., 2023; Paulo et al., 2024) can reveal shared patient metadata, such as a diagnosis, without isolating a response to composition. Population abundances and their

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.

ratios are established immune readouts (Krieg et al., 2018; Serrano-Villar et al., 2014; Templeton et al., 2014). We therefore ask how each SAE feature responds to changes in these readouts.

Population proportions vary together across samples, so correlations alone cannot isolate which changes a feature responds to. We measure these responses through in silico perturbations: resampling a sample's own cells to change the share of a target population A within an enclosing scope S, then measuring the response against a paired baseline. Since proportions sum to one (Büttner et al., 2021), the target displaces cells from the complement $S \setminus A$, while cells outside S stay fixed. For example, increasing CD8 T cells within T cells displaces other T cells; increasing them within the whole sample can also displace other lineages. We select responses with strong directional agreement across discovery samples and summarize them as testable annotations naming the target, scope and direction (Figure 1). The same paired measurements also allow a second question: whether the model responds differently to the same exchange in cases and controls. We use this contrast to find shared-feature programs, groups of populations that the model encodes as moving together in a given immune state, which per-population abundance comparisons do not reveal (Section 6).

On GPCT and MAESTRO, more than 98% of annotations retain their direction with corrected significance when the exchanges are repeated on held-out samples. An LLM given only annotations classifies held-out diagnoses, COVID states and treatments; in two external cohorts, GPCT annotations yield similar observed AUROCs to cell type proportions for acute COVID-19. Case-enriched features recover expert-expected population changes more often than random features with similar annotation counts, and MAESTRO decoder steering agrees with the annotated direction for 93.2% of annotations.

The case–control contrast extends these annotation-level results to groups of populations. Linking populations through features that respond more strongly in cases yields shared-feature programs, 12 for MAESTRO and 17 for GPCT across the 20 comparisons. Applying decoder steering to a whole program, by raising its MAESTRO features together in held-out embeddings, moves its populations in the direction observed in cases in 11 of 12 comparisons, and the relative change is larger in cases than in controls in 13 of 14 held-out and external evaluations.

Our contributions are:

1. **Interpreting SAE features through in silico perturbations.** We use controlled target–scope exchanges to isolate how SAE features respond to specified composition changes, and summarize directionally consistent responses as testable annotations.
2. **Validation beyond discovery.** We establish held-out reproducibility, evaluate annotation-only zero-shot classification on held-out samples and external cohorts, and show that annotations of case-enriched features recover expert-expected population changes more often than random features with similar annotation counts.
3. **Decoder steering as a reverse test.** Moving held-out embeddings along SAE decoder directions reproduces the encoder-derived annotation directions for 93.2% of MAESTRO annotations across all six signed doses.
4. **State-dependent responses and shared-feature programs.** Beyond per-population abundance differences, case–control differences in feature responses group populations into shared-feature programs that the model encodes. Steering a program's features through the decoder moves these populations, more in held-out cases than in controls, including two external acute-COVID cohorts.

2 Related work

Cytometry representations. Permutation-invariant encoders represent samples as cell sets (Zaheer et al., 2017; Lee et al., 2019). MAESTRO learns whole-sample representations through masked encoding and self-distillation (Lee et al., 2025); GPCT transfers across marker panels and identifies cells influential to sample-level predictions (Zhuang et al., 2026). EventHorizon uses a hierarchical transformer and self-distillation to integrate heterogeneous clinical flow-cytometry panels into a unified specimen-level representation (Medina Grespan et al., 2026). These approaches describe samples or predictions; we study latent feature responses to specified changes in composition, asking which responses are shared across immune states and which differ between states.

Biological SAEs. SAEs expose sparse features of neural representations (Bricken et al., 2023; Huben et al., 2024; Gao et al., 2024). Biological applications link features to multi-omics concepts (Schuster, 2024), protein properties (Simon & Zou, 2025) and genomic annotations (Nair et al., 2026), with further work on single-cell foundation models (Pedrocchi et al., 2025). Automated annotation methods summarize inputs that activate a feature (Bills et al., 2023; Paulo et al., 2024), while evaluation methods test the resulting natural-language explanations (Huang et al., 2023). For whole-sample cytometry, an association with a population's abundance leaves the displaced cells and the response to an explicit exchange unresolved.

Perturbations and feature interpretation. Input perturbations have long been used to explain predictions (Ribeiro et al., 2016); multiple-instance interpretability methods identify instances and interactions relevant to a bag-level prediction (Early et al., 2022). Concept-based methods measure sensitivity along concept directions (Kim et al., 2018) or effects of concept interventions on predictions (Goyal et al., 2019). Counterfactual gene perturbations have also been used to interpret SAE features in single-cell RNA-seq models (Claye et al., 2025). Our focus is the feature's dose response to a change in cell composition, realized by resampling measured cells within a specified target–scope exchange. Naming the displaced cells makes the relative nature of composition explicit, a central concern in compositional data analysis (Egozcue & Pawlowsky-Glahn, 2005). We summarize the responses shared across sampled immune states as annotations, and the same measurements also allow responses to one exchange to be compared between cases and controls.

3 Problem setup: composition-level feature interpretation

Our object of interpretation is a frozen model's response to a controlled change in cell composition. We study this response through SAE features and summarize the response directions shared across sampled immune states as annotations. Section 6 reuses the same measurements to compare responses between cases and controls.

Frozen backbone and sparse features. A cytometry sample $X \in \mathbb{R}^{n\times m}$ contains n cells measured on m protein markers. A frozen backbone F, either GPCT (Zhuang et al., 2026) or MAESTRO (Lee et al., 2025), maps it to a sample embedding $z = F(X) \in \mathbb{R}^d$ ($d = 128$ and 256, respectively). Following the Top-k SAE of Gao et al. (2024), the SAE computes

$$p(z) = W_{\text{enc}}(z - b_{\text{pre}}) + b_{\text{enc}}, \qquad a(z) = \text{ReLU}(\text{TopK}_k(p(z))), \qquad \hat{z} = W_{\text{dec}}\, a(z) + b_{\text{pre}}, \quad (1)$$

with $W_{\text{enc}} \in \mathbb{R}^{D\times d}$, unit-norm decoder columns $W_{\text{dec}} \in \mathbb{R}^{d\times D}$ and D latent features; SAE training and the choice of D and k are described in Appendix C. Feature indices are model-specific. We measure the response of feature j through its continuous pre-activation p_j and say that it fires when $a_j > 0$.

Target and exchange scope. Such a response is defined only relative to a specified change in composition. Raising the CD8 share among T cells displaces CD4 and other subsets, whereas the same gain among all cells can displace any population. Both whole-sample frequencies and within-population ratios are established readouts (Krieg et al., 2018; Serrano-Villar et al., 2014; Templeton et al., 2014; Maecker et al., 2012). We therefore define a target population A inside an enclosing exchange scope S. The exchange varies the share of A within S by trading cells with the scope complement $S \setminus A$. Cells outside S stay fixed, and a paired input at the original composition supplies the baseline. The root scope tests overall abundance, and a narrower scope tests the share of A within that population.

Annotations from shared responses. For each sample, feature and dose of an exchange (A, S), the measured response is the change in the feature's pre-activation from the paired baseline (Eq. 2). We estimate a pooled dose-response slope and ask whether its direction is consistent across discovery samples. The output is an annotation, such as *"This feature's pre-activation increases as the CD8 T-cell (A) share within T cells (S) increases."* Agreement in direction does not imply equal response magnitude across samples.

Information boundaries for the two analyses. The populations A and S come from a hierarchy of marker gates that is fixed before any perturbation is measured. Its structure and cutoffs follow

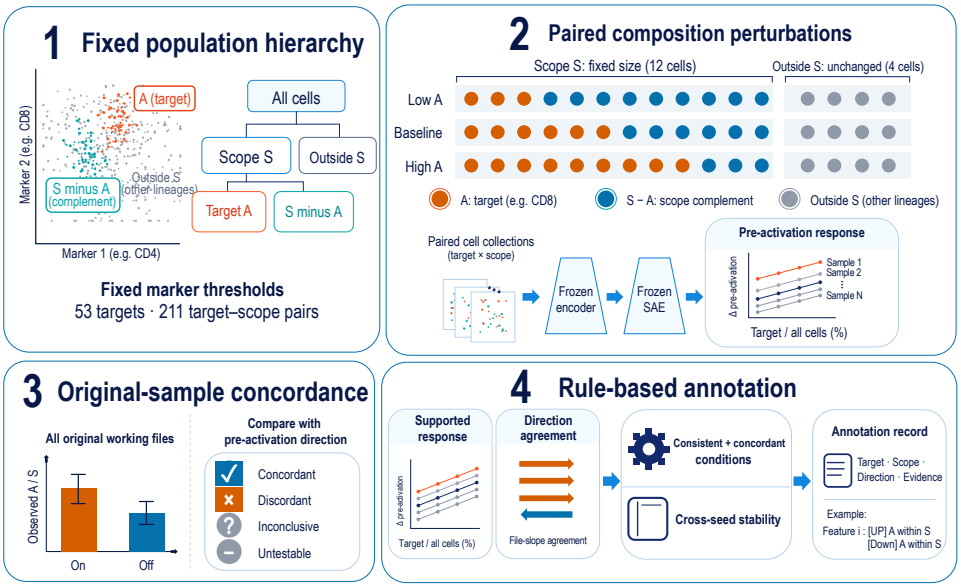

Figure 1: Four-stage annotation workflow. Stages 1–2 measure the paired responses that Section 6 also reuses; Stages 3–4 select annotations. Miniature plots and drawn cells are schematic; the listed protocol quantities are fixed settings.

standard marker-based gating conventions (Bagwell et al., 2020; Maecker et al., 2012). Annotations use these gates, perturbation responses and original-sample concordance without clinical labels. Metadata used to evaluate these frozen annotations (Section 5) cannot revise them. In the case–control analysis of Section 6, labels instead define the groups being compared and are part of the question.

4 Interpreting SAE features through in silico perturbations

We annotate features in the four stages of Figure 1. Stages 1–2 measure paired perturbation responses, which Section 6 also reuses, and Stages 3–4 select annotations. Gates, doses and rules are fixed before annotation, which uses neither clinical labels nor an LLM.

Fixed population hierarchy (Stage 1). Every exchange names a target A and a scope S, so both must be executable populations, fixed before any response is measured. A shared hierarchy gates 30 markers into 53 non-root populations (Maecker et al., 2012; Finak et al., 2014). Each node can serve as a target within any of its ancestors, including the root, which gives 211 target–scope pairs. Both encoders use identical gates and cutoffs (Appendix A.1).

Paired perturbations (Stage 2). Each model input holds 40,000 cells resampled from one sample's own cells. To raise the share of A within S, we draw more cells from A and fewer from $S \setminus A$, keeping the number of S cells and the cells outside S fixed. No marker value is edited: every input consists of measured cells, and changing marker values would itself move cells between nodes of the hierarchy. Doses are fixed quantiles of the within-scope share across discovery samples, and each sample also contributes a baseline drawn at its own composition. A sample enters the fit only when every input can be built and passes an out-of-distribution guard against original samples (Appendices A.2 and A.3).

For sample i, dose l and feature j, let $\Delta p_{ilj} = p_j(X_{il}) - p_j(X_{i0})$ be the change in pre-activation from the paired baseline X_{i0}, and let x_{il} be the achieved target percentage among *all* input cells. We fit a **pooled dose-response model**:

$$\Delta p_{ilj} = \alpha_{ij} + \beta_{j,A,S} x_{il} + \epsilon_{ilj}. \tag{2}$$

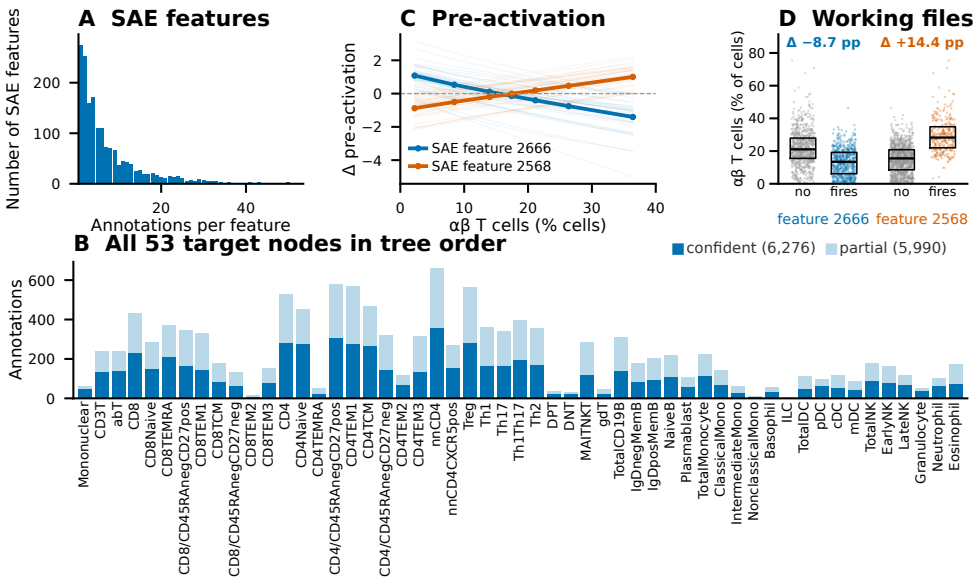

Figure 2: GPCT discovery evidence. (a) Annotations per feature. (b) Annotations per target node. (c) Pre-activation responses to the $\alpha\beta$ T-cell share (± 1.96 SE). (d) Original $\alpha\beta$ T-cell share in the 1,380 discovery samples, split by whether each feature fires; boxes show median and IQR, and Δ is the firing-minus-nonfiring mean difference.

The sample intercept α_{ij} removes differences between samples, so β compares doses within a sample. Standard errors are clustered by sample (CR2; Pustejovsky & Tipton, 2018), Bonferroni correction runs over the alive features of each exchange, and p-values are withheld when support is insufficient (Appendix A.4). We also record *activation coverage*, the fraction of samples in which the feature fires at some dose; it is descriptive and enters no selection rule (Appendix A.5).

Original-sample concordance (Stage 3). Stage 3 asks whether the Stage 2 direction also appears in unperturbed discovery samples. For each feature and target–scope pair, samples are split by whether the feature fires, and their mean A/S shares are compared. The pair is *concordant* when the difference is significant after Benjamini–Yekutieli correction and has the sign of the slope (Appendix A.6). Because Stage 3 reuses discovery samples, it selects annotations but does not validate them; Section 5 repeats both tests on held-out samples.

Rule-based annotation (Stage 4). A response is *consistent* when its slope is significant after Bonferroni correction, its sign agrees with the majority of per-sample slopes, and this sign agreement ranks in the top 10% of significant responses for that model (Appendix A.7). This criterion selects a direction shared across the sampled immune states, not equal magnitudes; neither effect size nor coverage enters it. Every consistent and concordant response becomes an annotation in the fixed template of Section 3, and a feature can carry many. Annotations of features whose decoder direction recurs in two SAEs retrained with other seeds are labelled *confident*, and the others *partially confident*; the label describes stability, not correctness.

5 Experiments and Evaluation

We apply the same hierarchy to GPCT and MAESTRO on a 30-marker immune panel (Lee et al., 2025; Bagwell et al., 2020). The data contain 1,380 discovery and 314 held-out samples, with known patients kept within one split, and each encoder perturbs its own 512 discovery samples. Gates, model–SAE pairs and annotations are fixed before held-out evaluation (Appendix E.1). Every acquired sample is treated as an independent unit; repeated samples of one donor come from different batches, and the sample intercept absorbs any constant patient effect (Appendices B and A.4).

5.1 ANNOTATION COVERAGE AND HELD-OUT EVIDENCE

Of 615,714 GPCT and 824,250 MAESTRO candidate feature–exchange responses, 7.6% and 2.8% are consistent, and Stage 3 retains 26% and 32% of these (Table 1). This yields 12,266 GPCT annotations across 1,705 features and 7,440 MAESTRO annotations across 1,167 features. The firing comparison alone would not fix the direction: among consistent responses with a significant difference, the difference opposes the slope for 4.7% of GPCT and 3.4% of MAESTRO pairs, far above sampling error, and the $|t|$ statistics of the two tests correlate at a Spearman ρ of .02 (Appendix A.6).

Table 1: Discovery annotation funnel. (F, S, A): feature, scope and target.

	GPCT	MAESTRO
Dictionary size / Top-k	3,072 / 64	4,096 / 64
Alive features	2,946	3,925
Measurements / (S, A) pairs	209 / 211	210 / 211
Candidate (F, S, A)	615,714	824,250
Consistent responses	46,722	23,482
Final annotations	12,266	7,440
Features with annotations	1,705 (57.9%)	1,167 (29.7%)
Confident	6,276	6,193
Partially confident	5,990	1,247

An annotated GPCT feature carries a median of four annotations, and every target node receives some (Figure 2a,b). In the example of panels c and d, the pre-activation of feature 2568 rises and that of feature 2666 falls as $\alpha\beta$ T cells gain share among all cells, and the original $\alpha\beta$ T-cell share is correspondingly higher or lower where each fires. Most features respond to several populations and scopes (Table 7; Appendix D gives the MAESTRO counterpart).

Two held-out tests rerun the frozen annotations on the 314 held-out samples with the same gates, doses and encoder–SAE pairs (Appendix E.2). Repeating each annotated exchange reproduces the annotated slope sign with Bonferroni significance for 12,264 of 12,266 GPCT and 7,316 of 7,440 MAESTRO annotations. Repeating the Stage 3 firing comparison, the held-out difference in A/S share follows the annotated direction for 93.8% and 94.3% of testable annotations, with significant opposite evidence for only 1 and 3. These tests repeat model responses on new samples and make no biological claim.

5.2 ZERO-SHOT METADATA CLASSIFICATION ON HELD-OUT SAMPLES

We ask whether the annotations carry the information that separates 16 metadata comparisons (11 diagnoses against Healthy, three COVID states and two treatments; Table 13) on held-out samples. No classifier is fitted for this evaluation. In each independent call, an LLM receives the evidence for one blinded sample and returns a probability of the case label. *Cell type proportions* give the sample's share of each of the 53 hierarchy populations; *SAE annotations* list its 16 most active annotated features with their annotations and no proportion. Both describe a single sample, so proportions are the matching baseline (Appendix E.5). Each sample is classified in three independent LLM calls per comparison and input.

GPCT annotations reach a mean AUROC of .838 $\pm$ .012 (SD over runs), against .781 $\pm$ .007 for cell type proportions and .745 $\pm$ .010 for MAESTRO annotations (Table 2). GPCT annotations score higher than cell type proportions in 6 of the 11 diagnoses and in all three COVID-state comparisons, whereas proportions score slightly higher on both treatments, where B-cell depletion is directly visible in the proportions. These AUROCs also reward study differences, since most Healthy samples come from studies other than the case arms. In MS, the only comparison with a within-study control arm, the ordering of inputs is preserved (Appendix E.5).

5.3 EXTERNAL ZERO-SHOT CLASSIFICATION

We then use two published COVID-19 cohorts that played no part in SAE training or annotation: hospitalized patients (Geanon et al., 2021) and critically ill patients in intensive care (Penttilä et al., 2021), each with healthy donors (Table 9). The LLM and inputs are those of Section 5.2, and each of the 216 samples is classified in three independent LLM calls per input (Appendix E.6).

Table 2: Held-out zero-shot AUROC, mean ± SD over three runs (Last row: per-run mean across comparisons). n: cases/controls; diagnoses share 70 Healthy controls. Bold: best mean per row. Intervals and paired differences: Table 14.

			SAE annotations	
Case vs control	n	Cell type proportions	GPCT	MAESTRO
ANCA Vasculitis	28/70	.914 ± .011	.901 ± .012	**.925 ± .011**
Acute COVID	21/70	.846 ± .007	**.965 ± .009**	.776 ± .035
Cardiac Sarcoidosis	5/70	.753 ± .043	**.895 ± .027**	.728 ± .022
Hidradenitis Suppurativa	10/70	.495 ± .022	.443 ± .052	**.502 ± .021**
Long COVID	24/70	.425 ± .045	**.802 ± .029**	.423 ± .023
Lymphoma	42/70	**.978 ± .003**	.906 ± .010	.802 ± .016
MS	58/70	**.926 ± .015**	.822 ± .028	.742 ± .008
Melanoma	22/70	.843 ± .016	**.912 ± .020**	.866 ± .011
Multiple Myeloma	4/70	**.970 ± .025**	.961 ± .028	.959 ± .006
Recovered COVID	18/70	.625 ± .023	**.836 ± .011**	.678 ± .019
Sepsis	8/70	.931 ± .012	**.999 ± .001**	.900 ± .026
Long vs Recovered COVID	24/15	.340 ± .014	**.458 ± .012**	.238 ± .045
Acute vs Recovered COVID	21/15	.783 ± .017	**.840 ± .015**	.796 ± .033
Acute vs Long COVID	21/24	.838 ± .015	**.860 ± .014**	.799 ± .011
MS: anti-CD20 vs S1P	35/19	**.998 ± .001**	.974 ± .015	.969 ± .011
Lymphoma: CAR-T vs Untreated/BTKi	29/5	**.839 ± .018**	.828 ± .054	.816 ± .020
All (16), mean		.781 ± .007	**.838 ± .012**	.745 ± .010

GPCT annotations alone yield similar observed AUROCs to cell type proportions on acute COVID-19: .894 and .974 in the hospitalized and intensive-care cohorts against .854 and .933 (Table 3). The paired AUROC difference, computed from each sample's mean probability over the three runs and then averaged over cohorts, is .036 [−.010, .093], which does not establish a GPCT advantage. The table instead averages the per-run AUROCs. MAESTRO annotations are weaker in the intensive-care cohort. The list identifies the most active annotated features within each sample, providing a learned summary of composition-related model responses (Appendix E.6).

Table 3: Zero-shot AUROC for acute COVID-19 in the two external cohorts, mean ± SD over three runs. Bold is best per cohort.

Input	Hospitalized	Intensive care
Cell type proportions	.854 ± .014	.933 ± .010
SAE annotations		
GPCT	**.894 ± .019**	**.974 ± .004**
MAESTRO	.874 ± .029	.772 ± .033

5.4 Expert knowledge in the annotations

Because annotations use no clinical labels, we can ask whether they contain what immunologists know. Six anonymized experts wrote which populations they expect to rise or fall in 20 case–control comparisons, giving 99 scoreable directional statements. For each comparison, we rank features by how much more often they fire in cases than in controls among discovery samples (Appendix E.8 maps this difference for 11 diagnoses) and read the annotations of the top five. A statement is recovered when these annotations carry its target and direction under any scope and none carries the opposite. They recover 82% of the statements for GPCT and 65% for MAESTRO, against 26% and 12% for random features with similar annotation counts and 41% and 28% under permuted labels (Table 4). For example, the top GPCT feature for acute COVID-19 reads granulocytes up and T cells down. The excess over both nulls shows that the ranking selects expected labels, not that each annotation is correct (Appendix E.7).

Ranking features within each external cohort of Section 5.3 instead, GPCT recovers 75% of the 12 acute COVID-19 statements at top five in both cohorts, through largely different features than in discovery, whereas MAESTRO recovers 0% and 33% (Table 4, bottom).

5.5 Decoder steering

Annotations are derived on the encoder side. MAESTRO also decodes an embedding into cells, which allows the reverse test. We add a feature's SAE decoder direction to the frozen embedding of each held-out sample, at $\pm 1, \pm 2, \pm 4$ times its 95th-percentile discovery activation, and ask whether

Table 4: Expert statements recovered (%) by top-k case-enriched features. Top: discovery comparisons by family. Bottom: acute COVID-19 across cohorts. n: scoreable statements; external cohort counts are COVID-19/Healthy samples.

		GPCT					MAESTRO				
Questions	n	top@1	top@2	top@3	top@4	top@5	top@1	top@2	top@3	top@4	top@5
Diagnosis (11)	36	44.4	58.3	69.4	72.2	80.6	13.9	16.7	33.3	38.9	47.2
COVID state (3)	22	22.7	27.3	50.0	50.0	68.2	0.0	9.1	27.3	36.4	54.5
Treatment (6)	41	36.6	92.7	92.7	92.7	90.2	46.3	78.0	82.9	85.4	85.4
Overall (20)	**99**	36.4	65.7	74.7	75.8	81.8	24.2	40.4	52.5	57.6	64.6
Feature null		9.4	15.1	19.8	23.3	25.9	3.3	6.9	9.3	10.7	11.8
Acute COVID-19 vs Healthy, ranked within each cohort											
Discovery	12	50.0	50.0	75.0	83.3	83.3	0.0	0.0	41.7	58.3	58.3
Hospitalized (77/15)	12	75.0	75.0	75.0	75.0	75.0	0.0	0.0	0.0	0.0	0.0
Intensive care (116/8)	12	50.0	75.0	75.0	75.0	75.0	0.0	0.0	33.3	33.3	33.3

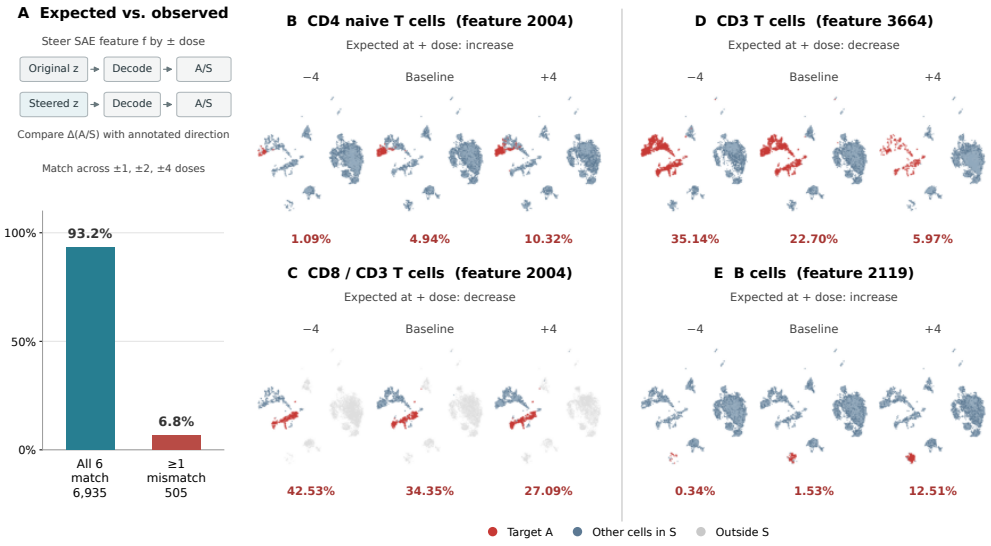

Figure 3: MAESTRO decoder steering on held-out samples. (a) Agreement of the decoded A/S change with the annotated direction. (b–e) Decoded cells at -4, baseline and $+4$; red, blue and gray cells lie in A, in $S \setminus A$ and outside S. UMAP area does not measure abundance.

the decoded A/S share moves in the annotated direction. Of the 7,440 MAESTRO annotations, 6,935 (93.2%) match at all six doses (Figure 3). In one example, a single step lowers the CD8 share within CD3 T cells while raising it among all cells, as its two annotations with different scopes predict (Appendix E.9).

Steering tests compatibility between annotations and the model, not independent biological validity, and large steps extrapolate beyond the discovery activation range. A complementary decoder-free composition readout on a separate annotation set is reported in Appendix F. A separate atlas steers case-enriched GPCT features in control embeddings (Appendix G).

6 State-dependent responses and shared-feature programs

Stage 2 also asks whether the model responds differently to the same exchange in cases and controls. We fit Eq. 2 separately in each group and keep case-driven differences $\Delta\beta = \beta_{\text{case}} - \beta_{\text{control}}$ that pass a label-permutation FWER on discovery samples (Appendix H; labels define the groups, Section 3). Abundance comparisons find which populations differ; shared features also show which of them the model treats as moving together. Conditions (A, S) sharing case-driven features are linked into a graph whose signed communities we freeze as *shared-feature programs* (Figure 4A).

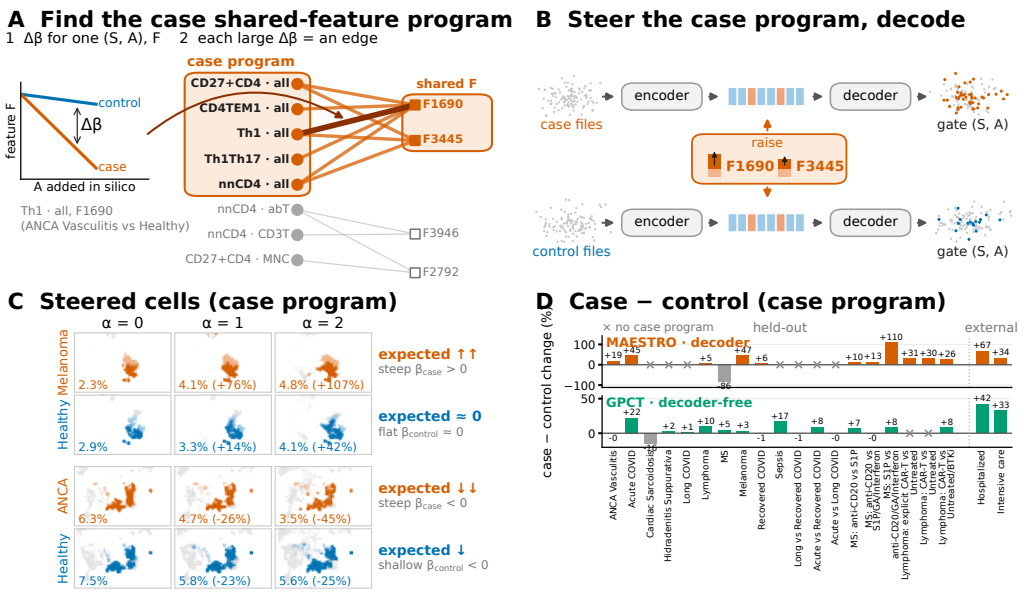

Figure 4: Case shared-feature programs. (A) Program discovery from case–control slope differences. (B) Steering. (C) Decoded cells at $\alpha = 0, 1, 2$. (D) Case minus control relative change at $\alpha = 2$; ×: no program.

Steering tests whether a program is consistent within the model and generalizes. We raise its features together in held-out case and control embeddings, by twice the discovery case–control centroid distance, and read each member A/S from decoded cells for MAESTRO or, lacking a decoder, from the nearest held-out samples for GPCT (Appendix F; Figure 4B,C). We report the relative change, signed by the direction β_{case} predicts and averaged over members, as case minus control (Figure 4D; Table 20).

Of the 20 case–control comparisons, 12 yield a MAESTRO program and 17 a GPCT program. For MAESTRO, raising a program moves every member A/S of held-out cases in the direction of β_{case} in 11 of 12 comparisons and in both external acute-COVID cohorts, and the relative change is larger in cases than in controls in 13 of 14 evaluations. With the decoder-free readout, GPCT shows a positive gap in 13 of 19 evaluations, including both external cohorts. Features read from the encoder thus move the same populations through the decoder, on samples not used to find them. The programs describe what the model encodes, not demonstrated biological co-variation, and in disease comparisons diagnosis and study coincide.

7 Conclusion

We interpret SAE features of cytometry foundation models by specifying which cell populations are exchanged and measuring the resulting feature responses. Held-out replication, annotation-only classification and agreement with expert expectations support the reproducibility and relevance of these interpretations. Decoder steering tests the relationship in reverse, asking whether interventions on features produce the composition changes predicted from input perturbations. The same measurements also reveal case–control response differences that link composition exchanges into shared-feature programs. Joint steering on held-out samples and external cohorts supports the state-dependent behavior of these programs. Together, these analyses provide a way to audit both shared and state-dependent composition responses in frozen models, while keeping claims about model behavior distinct from claims about biological mechanisms.

Ethics statement

This study received separate ethical approval. Human cytometry data were used under the applicable data-use agreements. Raw cell matrices and identifying records remain in the controlled research environment. Institutional details and approval identifiers are withheld for double-blind review.

AI use statement

Annotation generation is rule-based and does not use a language model. Claude Opus 5.5 was used for blinded zero-shot classification of all inputs in the held-out and external evaluations; prompts and first answers are retained with the evaluation records. Language models assisted with code implementation, manuscript revision, methodological and experimental feedback, interpretation of results, and checks of the implementation against the described protocol. The authors reviewed the AI-assisted work and retained responsibility for the research decisions, analyses and final claims. The research was not conducted as a fully autonomous, end-to-end AI workflow; automated computation and the specified LLM evaluations operated under author-defined protocols and oversight. Figure 1 is a schematic illustration; empirical plots show the measured results.

Reproducibility statement

The population hierarchy, encoder and SAE parameters, selected samples, doses, random seeds and statistical settings are recorded. Each annotation is traceable to its supporting measurements and cross-seed stability assessment. Retrospective analyses are identified as such. Determinism is conditional on the recorded numerical environment; a shared seed does not establish cross-hardware equality. The code is available in an anonymized repository at `https://anonymous.4open.science/r/cytofm-pipeline-86CC`.

References

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.

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. Language models can explain neurons in language models. *OpenAI Blog*, 2023. URL `https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html`.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL `https://transformer-circuits.pub/2023/monosemantic-features`.

Maren Büttner, Johannes Ostner, Christian L. Müller, Fabian J. Theis, and Benjamin Schubert. scCODA is a bayesian model for compositional single-cell data analysis. *Nature Communications*, 12:6876, 2021. doi: 10.1038/s41467-021-27150-6.

Charlotte Claye, Pierre Marschall, Wassila Ouerdane, Céline Hudelot, and Julien Duquesne. Discovering interpretable biological concepts in single-cell RNA-seq foundation models. *arXiv preprint arXiv:2510.25807*, 2025.

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

Joseph Early, Christine Evers, and Sarvapali Ramchurn. Model agnostic interpretability for multiple instance learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Juan José Egozcue and Vera Pawlowsky-Glahn. Groups of parts and their balances in compositional data analysis. *Mathematical Geology*, 37(7):795–828, 2005. doi: 10.1007/s11004-005-7381-9.

Greg Finak, Jacob Frelinger, Wenxin Jiang, Evan W. Newell, John Ramey, Mark M. Davis, Spyros A. Kalams, Stephen C. De Rosa, and Raphael Gottardo. OpenCyto: An open source infrastructure for scalable, robust, reproducible, and automated, end-to-end flow cytometry data analysis. *PLoS Computational Biology*, 10(8):e1003806, 2014. doi: 10.1371/journal.pcbi.1003806.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*, 2024.

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.

Sergio Gómez, Pablo Jensen, and Alex Arenas. Analysis of community structure in networks of correlated data. *Physical Review E*, 80(1):016114, 2009.

Yash Goyal, Amir Feder, Uri Shalit, and Been Kim. Explaining classifiers with causal concept effect (CaCE). *arXiv preprint arXiv:1907.07165*, 2019.

Jing Huang, Atticus Geiger, Karel D'Oosterlinck, Zhengxuan Wu, and Christopher Potts. Rigorously assessing natural language explanations of neurons. *arXiv preprint arXiv:2309.10312*, 2023.

Robert Huben, Hoagy Cunningham, Logan Riggs, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *International Conference on Machine Learning (ICML)*, 2018.

Carsten Krieg, Malgorzata Nowicka, Silvia Guglietta, Sabrina Schindler, Felix J. Hartmann, Lukas M. Weber, Reinhard Dummer, Mark D. Robinson, Mitchell P. Levesque, and Burkhard Becher. High-dimensional single-cell analysis predicts response to anti-PD-1 immunotherapy. *Nature Medicine*, 24(2):144–153, 2018. doi: 10.1038/nm.4466.

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

Holden T. Maecker, J. Philip McCoy, and Robert Nussenblatt. Standardizing immunophenotyping for the Human Immunology Project. *Nature Reviews Immunology*, 12(3):191–200, 2012. doi: 10.1038/nri3158.

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 preprint 2026.06.18.733197*, 2026. doi: 10.64898/2026.06.18.733197. URL `https://www.biorxiv.org/content/10.64898/2026.06.18.733197v1`.

Akira Nair, Jaehyun Joo, Jonghyun Lee, Lina Takemaru, Yidi Huang, Manu Shivakumar, Matthew E. Lee, Jaesik Kim, Sokratis A. Apostolidis, and Dokyoon Kim. Interpreting genomic language models using sparse autoencoders. In *International Conference on Machine Learning (ICML)*, 2026. URL https://icml.cc/virtual/2026/poster/62924.

Gonçalo Paulo, Alex Mallen, Caden Juang, and Nora Belrose. Automatically interpreting millions of features in large language models. *arXiv preprint arXiv:2410.13928*, 2024.

Samuel Pawel and Leonhard Held. Probabilistic forecasting of replication studies. *PLOS ONE*, 15 (4):e0231416, 2020. doi: 10.1371/journal.pone.0231416.

Flavia Pedrocchi, Florian Barkmann, Amir Joudaki, and Valentina Boeva. Sparse autoencoders reveal interpretable features in single-cell foundation models. *bioRxiv preprint 2025.10.22.681631*, 2025. doi: 10.1101/2025.10.22.681631.

Pieter A. Penttilä, Sofie Van Gassen, Dena Panovska, Lore Vanderbeke, Yannick Van Herck, Katrien Quintelier, Annelies Emmaneel, Jessica Filtjens, Bert Malengier-Devlies, Kourosh Ahmadzadeh, Pierre Van Mol, Daniel M. Borràs, Asier Antoranz, Francesca M. Bosisio, Els Wauters, Kimberly Martinod, Patrick Matthys, Yvan Saeys, Abhishek D. Garg, Joost Wauters, and Frederik 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.

James E. Pustejovsky and Elizabeth Tipton. Small-sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *Journal of Business & Economic Statistics*, 36(4):672–683, 2018. doi: 10.1080/07350015.2016.1247004.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016. doi: 10.1145/2939672.2939778. URL https://arxiv.org/abs/1602.04938.

Yvan Saeys, Sofie Van Gassen, and Bart N. Lambrecht. Computational flow cytometry: helping to make sense of high-dimensional immunology data. *Nature Reviews Immunology*, 16(7):449–462, 2016. doi: 10.1038/nri.2016.56.

Viktoria Schuster. Can sparse autoencoders make sense of gene expression latent variable models? *arXiv preprint arXiv:2410.11468*, 2024.

Sergio Serrano-Villar, Talia Sainz, Sulggi A. Lee, Peter W. Hunt, Elizabeth Sinclair, Barbara L. Shacklett, April L. Ferre, Timothy L. Hayes, Ma Somsouk, Priscilla Y. Hsue, Mark L. Van Natta, Curtis L. Meinert, Michael M. Lederman, Hiroyu Hatano, Vivek Jain, Yong Huang, Frederick M. Hecht, Jeffrey N. Martin, Joseph M. McCune, Santiago Moreno, and Steven G. Deeks. HIV-infected individuals with low CD4/CD8 ratio despite effective antiretroviral therapy exhibit altered T cell subsets, heightened CD8+ T cell activation, and increased risk of non-AIDS morbidity and mortality. *PLoS Pathogens*, 10(5):e1004078, 2014. doi: 10.1371/journal.ppat.1004078.

Elana Simon and James Zou. InterPLM: Discovering interpretable features in protein language models via sparse autoencoders. *Nature Methods*, 2025. doi: 10.1038/s41592-025-02836-7. URL https://www.nature.com/articles/s41592-025-02836-7.

Matthew H. Spitzer and Garry P. Nolan. Mass cytometry: single cells, many features. *Cell*, 165(4): 780–791, 2016. doi: 10.1016/j.cell.2016.04.019.

Arnoud J. Templeton, Mairéad G. McNamara, Boštjan Šeruga, Francisco E. Vera-Badillo, Priya Aneja, Alberto Ocaña, Raya Leibowitz-Amit, Guru Sonpavde, Jennifer J. Knox, Ben Tran, Ian F. Tannock, and Eitan Amir. Prognostic role of neutrophil-to-lymphocyte ratio in solid tumors: a systematic review and meta-analysis. *Journal of the National Cancer Institute*, 106(6):dju124, 2014. doi: 10.1093/jnci/dju124.

Christina V. Theodoris, Ling Xiao, Anant Chopra, Mark D. Chaffin, Zeina R. Al Sayed, Matthew C. Hill, Helene Mantineo, Elizabeth M. Brydon, Zexian Zeng, X. Shirley Liu, and Patrick T. Ellinor. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023. doi: 10.1038/s41586-023-06139-9.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander Smola. Deep sets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

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 preprint 2026.03.31.712806*, 2026. doi: 10.64898/2026.03.31.712806. URL https://www.biorxiv.org/content/10.64898/2026.03.31.712806v1.

A METHOD DETAILS

A.1 FIXED EXECUTABLE POPULATION DEFINITIONS

Both models use the same prespecified 30-marker hierarchy: 53 non-root populations and 211 target–strict-ancestor pairs (Table 5; Bagwell et al., 2020; Maecker et al., 2012). Each node inherits its parent's conditions, including technical cleanup gates. Strict thresholds permit gaps at equality, and siblings may overlap. Names denote operational gates rather than verified identities: MAITNKT is CD161-positive/CD4-negative, and nnCD4 is CD45RA-negative within CD4. Neither response statistics nor LLM outputs change these definitions.

Table 5: Executable population definitions, grouped by parent. Each child inherits every condition of its parent and adds the listed strict thresholds on the input marker scale. Cutoffs shared by two or more siblings are given once in the middle column; a child then shows only the side, with $^+$ meaning above and $^-$ meaning below that cutoff. Other conditions are written inline. Names are operational population labels. The population hierarchy lists CD3 $<$ 2.175 twice in TotalDC and TotalNK (and so in their descendants); the repeat does not change the gate and is shown once.

Parent	Shared cutoffs	Children and added conditions
root	CD66b 1.58	**Mononuclear** CD66b$^-$ CD45 $>$ 2.741; **Granulocyte** CD66b$^+$
Mononuclear	CD20 1.377	**CD3T** CD11c $<$ 2.055, CD14 $<$ 2.649, CD3 $>$ 1.997, CD19 $<$ 2.319; **TotalCD19B** CD20$^+$ CD14 $<$ 2.649, CD56 $<$ 1.983, CD123 $<$ 3.472, CD3 $<$ 1.995, CD19 $>$ 2.109; **Plasmablast** CD20$^-$ CD14 $<$ 2.649, CD56 $<$ 1.983, CD123 $<$ 3.472, CD3 $<$ 1.995, CD27 $>$ 1.497, CD38 $>$ 5.391, CD19 $>$ 2.109; **TotalMonocyte** CD20$^-$ CD19 $<$ 2.516, CD3 $<$ 2.175, CD56 $<$ 2.398, CD14 $>$ 1.574; **Basophil** CD20$^-$ CD19 $<$ 2.516, CD3 $<$ 2.175, CD56 $<$ 2.398, HLA-DR $<$ 0.65, CD123 $>$ 3.335; **ILC** CD20$^-$ CD19 $<$ 2.516, CD3 $<$ 2.175, CD56 $<$ 2.398, HLA-DR $<$ 0.65, CD123 $<$ 3.335, CD127 $>$ 3.544, CD161 $>$ 3.31; **TotalDC** CD19 $<$ 2.516, CD3 $<$ 2.175, CD14 $<$ 1.782, HLA-DR $>$ 1.11, CD56 $<$ 1.895; **TotalNK** CD19 $<$ 2.516, CD3 $<$ 2.175, CD14 $<$ 1.782, CD56 $>$ 1.895, HLA-DR $<$ 1.11
CD3T	TCRgd 2.756	**abT** TCRgd$^-$; **gdT** TCRgd$^+$
abT	CD4 2.562; CD8a 3.577	**CD8** CD4$^-$CD8a$^+$; **CD4** CD4$^+$CD8a$^-$; **DPT** CD4$^+$CD8a$^+$; **DNT** CD4$^-$CD8a$^-$; **MAITNKT** CD4$^-$ CD161 $>$ 3.69
CD8	CD45RA 2.499; CD27 1.619	**CD8Naive** CD45RA$^+$CD27$^+$; **CD8TEMRA** CD45RA$^+$CD27$^-$; **CD8/CD45RAnegCD27pos** CD45RA$^-$CD27$^+$; **CD8/CD45RAnegCD27neg** CD45RA$^-$CD27$^-$
CD4	CD45RA 2.364; CD27 1.496	**CD4Naive** CD45RA$^+$CD27$^+$; **CD4TEMRA** CD45RA$^+$CD27$^-$; **CD4/CD45RAnegCD27pos** CD45RA$^-$CD27$^+$; **CD4/CD45RAnegCD27neg** CD45RA$^-$CD27$^-$; **nnCD4** CD45RA$^-$
CD8/ CD45RAneg CD27pos	CD197 2.402	**CD8TEM1** CD197$^-$; **CD8TCM** CD197$^+$
CD8/ CD45RAneg CD27neg	CD197 2.402	**CD8TEM2** CD197$^+$; **CD8TEM3** CD197$^-$
CD4/ CD45RAneg CD27pos	CD197 2.402	**CD4TEM1** CD197$^-$; **CD4TCM** CD197$^+$
CD4/ CD45RAneg CD27neg	CD197 2.402	**CD4TEM2** CD197$^+$; **CD4TEM3** CD197$^-$
nnCD4	CD185 2.147; CD183 1.835; CD196 1.548	**nnCD4CXCR5pos** CD185$^+$; **Treg** CD25 $>$ 1.042, CD127 $<$ 1.282; **Th1** CD185$^-$CD183$^+$CD196$^-$; **Th17** CD185$^-$CD196$^+$CD183$^-$; **Th1Th17** CD185$^-$CD196$^+$CD183$^+$; **Th2** CD185$^-$CD196$^-$CD183$^-$ CD194 $>$ 2.455
TotalCD19B	IgD 2.29; CD27 1.234	**IgDnegMemB** IgD$^-$CD27$^+$; **IgDposMemB** IgD$^+$CD27$^+$; **NaiveB** IgD$^+$CD27$^-$
TotalMonocyte	CD16 4; CD14 2.353	**ClassicalMono** CD16$^-$; **IntermediateMono** CD16$^+$CD14$^+$; **NonclassicalMono** CD16$^+$CD14$^-$
TotalNK	CD57 2.727	**EarlyNK** CD57$^-$; **LateNK** CD57$^+$
TotalDC	CD123 3.972; CD11c 1.409	**pDC** CD123$^+$CD11c$^-$; **cDC** CD11c$^+$CD123$^-$
cDC	–	**mDC** CD38 $>$ 2.423
Granulocyte	CD294 2.551	**Neutrophil** CD294$^-$ CD16 $>$ 2.208; **Eosinophil** CD294$^+$

A.2 SAMPLING, DOSE DESIGN AND THE OOD GUARD

For original fractions $f_{A,i}$ and $f_{S,i}$, the exchange scope occupies $n_{S,i} = \mathrm{round}(40{,}000 f_{S,i})$ slots. At within-scope dose r_l, the target occupies $a_{il} = \mathrm{round}(r_l n_{S,i})$ slots and the complement occupies $n_{S,i} - a_{il}$; the remaining slots come from outside S. Sampling is uniform with replacement within each pool. Shared random prefixes couple the pool draws across doses, while the outside draw remains identical. The paired baseline uses the same construction at the sample's original within-scope target fraction. Empty required pools and infeasible counts are recorded, not replaced by cells from another sample.

The fixed discovery quantiles are $0.05, 0.20, 0.35, 0.50, 0.65, 0.80, 0.95$ of target abundance within the scope. Duplicate quantile values are removed without interpolation. Annotation accepts six or seven distinct values: 204 logical conditions use seven and seven use six. Six-dose conditions are CD8TEM2 within root, Mononuclear, CD3T, abT, CD8 and its CD45RA-negative/CD27-negative parent, and CD8TEM3 within that parent. The number of doses is recorded for each condition. Each model selects 512 samples with seeded quotas across clusters of standardized representative embeddings; these embeddings are for selection, not paired perturbation baselines. Each model uses one repeat and seed 713.

The guard summarizes the raw cell input through marker moments, correlations, positivity/co-positivity frequencies and cell type proportions. A fixed standardizer and linear SVM produce anomaly score s. With M original calibration scores, define

$$q(s) = \frac{1 + \sum_{m=1}^{M} \mathbf{1}\{s_m \geq s\}}{M + 1}. \tag{3}$$

Inputs with $q(s) < 0.01$ are rejected. A paired slope requires an accepted baseline and every treatment dose. The benchmark used to select the guard is reported in Appendix A.3; its cross-fitted results are not an independent assessment of the deployed refit. Calibration against a fixed reference does not guarantee a 1% rejection rate under arbitrary shift. Figure 5 shows one complete measurement, including the per-dose guard exclusions.

A.3 OOD GUARD SELECTION BENCHMARK

The guard must accept unmodified inputs and reject specified synthetic stress cases. Every query contains 40,000 cells. Detectors use 541 marker features from the 30-marker panel: means, standard deviations, correlations, positivity and co-positivity frequencies, and cell type proportions. An initial original-only screen showed that 256-dimensional foundation-model embeddings were insensitive: at rejection threshold 0.05 an embedding one-class SVM detected 4.69% of 90%-T-cell mixtures, against 87.50% for the marker one-class SVM. These are exploratory discovery-data benchmarks, not held-out annotation replication.

The comparison keeps 436 fitting patients and 187 calibration patients fixed and splits the 119 synthetic-source patients into five folds, so each fold's original and synthetic inputs are scored only by a model that excluded those patients. Supervised linear and RBF SVMs ($C = 1$) are trained on originals plus synthetic cases from the other folds: 90% and 100% T, B, NK, monocyte and $\gamma\delta$ T mixtures, marker shifts, noise, column shuffling, single-cell replication, and forced CD3/CD19/CD14 co-expression. An RBF one-class SVM on the same originals is the original-only comparator. The threshold α is calibrated on the 187 calibration patients' originals. Composition resampling preserves marker values; corruption and co-expression are detector stress tests, not interventions used to annotate features.

The selected linear SVM at $\alpha = 0.01$ rejected 15/1,014 originals (1.48%; patient-bootstrap 95% interval 0.29–3.13%) and detected all 128 inputs of each pure-population case. It neither kept original rejection below 1% nor dominated every alternative. Withholding a stress family from training exposed limited generalization: without composition examples, T-90% detection fell from 128/128 to 1/128, and without co-expression examples, 10%-triple-positive detection fell from 127/128 to 15/128. The detector was chosen after this comparison and shares training and calibration patients across folds, so new patients and unseen stress families are needed to assess the deployed refit.

Table 6: Matched five-fold OOD benchmark: rejection percentages. Original-input rejection should be low; the remaining columns are sensitivities to specified stress cases, not universal biological invalidity. Originals: $n = 1{,}014$; each stress case: $n = 128$. OC denotes one-class SVM, TP denotes forced triple-positive cells, and shuffle permutes marker columns independently.

Model	α	Original	T 90%	Mono. 100%	TP 100%	Shuffle
OC RBF	.01	0.59	3.91	3.12	100.00	0.78
OC RBF	.05	5.92	87.50	100.00	100.00	100.00
Supervised RBF	.01	1.18	17.97	100.00	100.00	100.00
Supervised RBF	.05	6.51	87.50	100.00	100.00	100.00
Supervised linear	.01	1.48	100.00	100.00	100.00	99.22
Supervised linear	.05	5.33	100.00	100.00	100.00	100.00

A.4 Fixed-effects response estimator

Let G denote the number of complete samples and $w_i = 1/G$. With L actual doses, center the regressor and response within each sample: $\widetilde{x}_{il} = x_{il} - \bar{x}_i$ and $\widetilde{y}_{ilj} = \Delta p_{ilj} - \overline{\Delta p}_{ij}$. Then

$$\widehat{\beta}_j = \frac{\sum_i (w_i/L) \sum_l \widetilde{x}_{il} \widetilde{y}_{ilj}}{\sum_i (w_i/L) \sum_l \widetilde{x}_{il}^2}. \tag{4}$$

This equals weighted least squares with a sample-specific intercept. Units are pre-activation per percentage point of target abundance among all cells, not per percentage point within the scope. Equal total sample weights do not give equal slope information: samples with larger within-sample dose spread contribute more. The paired baseline cancels under centering but defines the displayed change and participates in the complete-sample requirement. A single slope summarizes the tested dose range and can hide a nonmonotonic response.

Define $Q_i = (w_i/L) \sum_l \widetilde{x}_{il}^2$, $Q = \sum_i Q_i$, $h_i = Q_i/Q$, and $u_{ij} = Q^{-1}(w_i/L) \sum_l \widetilde{x}_{il}(\widetilde{y}_{ilj} - \widehat{\beta}_j \widetilde{x}_{il})$. The implemented sample-clustered CR2 variance and degrees of freedom are

$$\widehat{\mathrm{se}}(\widehat{\beta}_j)^2 = \sum_i \frac{u_{ij}^2}{1 - h_i}, \qquad \nu = \left[\sum_i h_i^2 + \left(\sum_i \frac{h_i^2}{1 - h_i} \right)^2 - \sum_i \left(\frac{h_i^2}{1 - h_i} \right)^2 \right]^{-1}. \tag{5}$$

The working covariance is independent whitened error, corresponding to raw variance proportional to inverse weight (Pustejovsky & Tipton, 2018). Primary inference requires at least 30 informative samples, $\nu \geq 20$ and $\max_i h_i \leq 0.10$. The effective number of samples is $1/\sum_i h_i^2$. Failed guards or degenerate variance leave primary intervals and p-values unavailable, even if a diagnostic slope is finite. A pooled regression with outside-composition summaries is a sensitivity diagnostic, not the primary estimator.

Two-sided t_ν tests use $p_{\mathrm{adj},j} = \min(1, M_{\mathrm{alive}} p_j)$ within each physical exchange. Untestable non-dead features remain in that set; dead features are excluded but retained as unavailable rows in full-dictionary reports. Stored ordinary 95% intervals differ from familywise intervals. In particular, forest plots display ordinary intervals, while heatmap significance marks use the saved Bonferroni decision. Multiple exchanges and the global grading step are not jointly covered by this per-exchange correction.

Repeated samples. Patient and repeated-control counts are given in Appendix B. All acquired samples are retained, including technical controls measured in different study batches (Lee et al., 2025). Sample-specific intercepts absorb constant offsets, so slopes depend on within-sample dose responses. Inference nevertheless assumes sample independence; patient-disjoint splitting does not remove within-partition donor dependence.

Complementary exchanges.

Proposition 1 (Exact reciprocal reuse)**.** *If $A' = S \setminus A$ in every selected sample and mirrored doses and integer counts construct exactly the same inputs, then $\widehat{\beta}_{j,A',S} = -\widehat{\beta}_{j,A,S}$, with the same uncertainty and p-value. Activation coverage is unchanged.*

Proof. The complementary regressor is $x'_{il} = 100n_{S,i}/N - x_{il}$. Within-sample centering removes its constant term, so $\widetilde{x}'_{il} = -\widetilde{x}_{il}$. The shared inputs give the same response; Eq. 4 reverses the slope sign. The residuals and leverage are unchanged and u_{ij} reverses sign, leaving its squared contribution unchanged. The shared inputs also have identical firing states. □

Reuse requires exact checked complementarity and compatible dose and count designs; it is not inferred from sibling names. The two models therefore have different physical-measurement counts despite the same 211 logical views. Each physical source contributes once to ranks and final annotation counts.

A.5 Descriptive activation coverage

Let $\mathcal{T}_c$ be the samples whose treatments at every actual dose of condition c are feasible and OOD-valid. For feature j,

$$C_{jc} = \frac{1}{|\mathcal{T}_c|} \sum_{i \in \mathcal{T}_c} \mathbf{1}\left\{ \max_{1 \le l \le L_c} Y_{ilj} = 1 \right\}. \tag{6}$$

Coverage is unavailable for empty $\mathcal{T}_c$. Baselines enter neither this set nor the firing criterion, so coverage and slopes can use different samples. Coverage is descriptive and does not enter selection; six-dose designs offer fewer firing opportunities than seven-dose designs and are not directly comparable.

A.6 Original discovery concordance

All original discovery samples, rather than only the 512 selected perturbation samples, supply saved baseline firing states and gate counts from the full source cell collection. For every source-deduplicated feature–condition pair, fit

$$r_i = \eta + \gamma Y_{i0j} + u_i, \qquad r_i = |A_i|/|S_i|. \tag{7}$$

Each sample needs at least 20 scope cells, and each group needs five samples; zero target cells are allowed. The effect is the firing-minus-nonfiring mean difference in fraction units. Sample-level CR1 uncertainty uses t_{n-1}, where n is the eligible sample count. For group sizes n_1, n_0 and respective means $\bar{r}_1, \bar{r}_0$, its variance is

$$\widehat{\mathrm{se}}(\widehat{\gamma})^2 = \frac{n}{n-2} \left[\frac{\sum_{i:Y_i=1}(r_i - \bar{r}_1)^2}{n_1^2} + \frac{\sum_{i:Y_i=0}(r_i - \bar{r}_0)^2}{n_0^2} \right]. \tag{8}$$

Degenerate uncertainty is untestable. Benjamini–Yekutieli correction covers the entire planned source-deduplicated feature–condition set, including untestable hypotheses and dead-feature positions. At adjusted $q \le 0.05$, matching the pre-activation slope's sign is concordant and opposing it is discordant. Estimable nonsignificant associations are inconclusive; missing support or a missing/zero response direction is untestable. This selection uses original discovery observations and cannot supply independent confirmation.

Observational and paired evidence. We compare the firing-minus-nonfiring difference of Eq. 7 with the paired slope of Eq. 2 for every distinct measured feature–exchange condition with both estimates, 486,494 for GPCT and 634,994 for MAESTRO. Of the 105,284 and 128,080 conditions with a significant difference, a corrected slope of the same sign follows for 64.0% and 30.7%, a corrected slope of the opposite sign for 18.2% and 5.8%, and no corrected slope for 17.8% and 63.5%.

Among consistent conditions with a significant difference, 601 of 12,867 GPCT and 265 of 7,705 MAESTRO conditions are discordant (4.7% and 3.4%; 95% feature-cluster bootstrap intervals 4.1–5.2% and 2.7–4.2%). Discordance persists with at least 30 samples per firing group (3.8%, 2.6%), on discovery samples outside Stage 2 (4.3%, 2.3%), and after diagnosis adjustment (3.1%, 2.4%). In discordant conditions, perturbation firing rates move with the slope in 77.0% and 56.6%, and remain flat in 22.8% and 40.0%. Thus Top-k competition does not explain most disagreement.

Among conditions with a significant difference, the Spearman correlation between the difference $|t|$ and the slope $|t|$ is .019 for GPCT and .024 for MAESTRO (feature-cluster bootstrap intervals .009–.032 and .012–.033). After adjusting for firing-group sizes it is .025 and .035; within exchanges, absolute effect sizes correlate at medians .093 and .000. Signed statistics correlate at .42 and .35, reflecting shared signs rather than magnitudes.

A.7 Annotation selection and confidence

Let $\mathcal{E}$ contain supported, significant nondead physical feature–condition pairs. Their direction agreement is $d_{jc} = \max(n^+_{jc}, n^-_{jc})/n^{\text{estimable}}_{jc}$, with zero sample slopes included in the denominator and undefined slopes excluded. A tied positive/negative count has no majority direction. Within $\mathcal{E}$, the mid-percentile of a value v is

$$R_{\mathcal{E}}(v) = \frac{\#\{u \in \mathcal{E} : v_u < v\} + \frac{1}{2}\#\{u \in \mathcal{E} : v_u = v\}}{|\mathcal{E}|}. \tag{9}$$

Empty populations have undefined ranks and a constant population has rank 0.5. Set H_{jc} when $R_{\mathcal{E}}(d_{jc}) \geq 0.90$. Activation coverage C_{jc} is not ranked and does not enter H_{jc}. With E_{jc} denoting supported significance and L_{jc} pooled-slope/majority alignment,

$$T_{jc} = \begin{cases} \text{consistent}, & E_{jc} \wedge L_{jc} \wedge H_{jc}, \\ \text{non-consistent}, & \text{otherwise.} \end{cases} \tag{10}$$

Ranks are computed across all distinct measurements for each model before discordant conditions are separated. Neither raw slope magnitude, reference-SD scaling nor cross-seed stability enters this criterion. Ties mean the rule need not retain a fixed fraction of conditions, and consistency applies to a condition, not a feature.

The final set of annotations contains every physical condition whose response is consistent across samples and whose direction recurs in unperturbed samples. No limit is imposed on annotations per feature. Each annotation includes feature index, target and scope gates, dose levels, raw slope and adjusted significance, discovery supported dose range, sign agreement, coverage, observation status and stability. An existing stable flag yields confident; otherwise the label is partially confident. Both groups satisfy the same consistent/concordant requirements. The deterministic text states UP or DOWN for the target replacing the remainder of its scope. Discordant, inconclusive, non-consistent, dead and unavailable evidence remains accounted for without inventing an interpretation where support is missing. Table 7 lists the annotations of three features per model in this form.

B Datasets, splits and external cohorts

Source dataset. The MAESTRO mass cytometry dataset (Lee et al., 2025), obtained under a data use agreement, contains 1,694 blood samples measured with a 30-marker panel (Bagwell et al., 2020). The prespecified discovery/held-out split (Table 8) is shared across models and analyses, with verified zero overlap of linked patients.

Discovery set. The 1,380 discovery samples include 1,231 from 593 linked patients and 149 unlinked samples. Of the linked patients, 286 contribute repeated measurements (924 samples, at most ten per patient); 141 unlinked samples are repeated technical controls from a small number of healthy donors. Healthy combines 178 ordinary healthy and 141 BatchControl samples. All samples enter SAE training, so patients contribute in proportion to their sample counts; cross-validation folds are patient-grouped (Appendix C). Gates, perturbation sampling, doses, the OOD guard and annotation selection use only discovery data and do not constitute independent validation.

Held-out set. Held-out samples are excluded from annotation and used for reconstruction, response replication, phenotype concordance and classification. Replication keeps the discovery features, directions, gates and doses fixed. Earlier uses make this reused holdout evidence (Appendix E.1).

Table 7: Example annotations, three features per model. Two features per model are the ones named in the text or figures; the third is drawn at random (seed 713) from the remaining annotated features. Rows are a feature's annotations ranked by $|\beta|$, at most 12 per feature. Each row reads: the feature's pre-activation rises (UP) or falls (DOWN) when the target replaces the rest of its scope; β is the pooled within-sample slope (Eq. 2). Population names are the operational labels of Appendix A.1.

Direction	Target	Scope	β	Confidence
GPCT feature 1378: top-ranked for acute COVID-19 (Section 5.4); 12 annotations (2 UP, 10 DOWN; 12 confident)				
DOWN	pDC	All cells	-0.224	confident
DOWN	LateNK	All cells	-0.091	confident
DOWN	TotalNK	All cells	-0.079	confident
DOWN	EarlyNK	All cells	-0.073	confident
UP	Granulocyte	All cells	$+0.064$	confident
DOWN	Mononuclear	All cells	-0.063	confident
UP	Neutrophil	All cells	$+0.061$	confident
DOWN	Th1Th17	All cells	-0.054	confident
DOWN	CD3T	All cells	-0.045	confident
DOWN	abT	All cells	-0.045	confident
DOWN	CD4	All cells	-0.042	confident
DOWN	CD4Naive	All cells	-0.033	confident
GPCT feature 2568: Figure 2; second-ranked for anti-CD20 treatment; 36 annotations (24 UP, 12 DOWN; 36 confident)				
UP	CD4Naive	All cells	$+0.074$	confident
DOWN	cDC	Mononuclear	-0.066	confident
DOWN	NonclassicalMono	Mononuclear	-0.065	confident
UP	CD4	All cells	$+0.059$	confident
DOWN	IntermediateMono	Mononuclear	-0.057	confident
UP	CD3T	All cells	$+0.055$	confident
UP	abT	All cells	$+0.055$	confident
DOWN	TotalMonocyte	Mononuclear	-0.055	confident
UP	CD4TEMRA	All cells	$+0.054$	confident
UP	CD3T	Mononuclear	$+0.054$	confident
DOWN	ClassicalMono	Mononuclear	-0.053	confident
DOWN	IgDposMemB	Mononuclear	-0.052	confident
24 more				
GPCT feature 1485: random draw (seed 713); 2 annotations (0 UP, 2 DOWN; 0 confident)				
DOWN	CD8TEM2	All cells	-0.056	partially confident
DOWN	CD8TEM2	Mononuclear	-0.049	partially confident
MAESTRO feature 2927: top-ranked for lymphoma (Section 5.4); 10 annotations (7 UP, 3 DOWN; 10 confident)				
DOWN	TotalCD19B	All cells	-0.014	confident
UP	mDC	Mononuclear	$+0.010$	confident
DOWN	TotalCD19B	Mononuclear	-0.010	confident
UP	CD8TEMRA	CD8	$+0.007$	confident
UP	Th1	CD3T	$+0.005$	confident
UP	Th1	abT	$+0.005$	confident
DOWN	CD4Naive	CD4	-0.005	confident
UP	Th1	CD4	$+0.005$	confident
UP	Th1	Mononuclear	$+0.004$	confident
UP	Th1	nnCD4	$+0.003$	confident
MAESTRO feature 3333: Figure 6; 41 annotations (22 UP, 19 DOWN; 41 confident)				
DOWN	IgDnegMemB	All cells	-0.042	confident
UP	CD8TEMRA	All cells	$+0.040$	confident
DOWN	IgDnegMemB	Mononuclear	-0.026	confident
UP	CD8TEMRA	abT	$+0.026$	confident
UP	CD8TEMRA	CD3T	$+0.026$	confident
UP	CD8TEMRA	Mononuclear	$+0.025$	confident
DOWN	TotalCD19B	All cells	-0.024	confident
DOWN	NaiveB	All cells	-0.023	confident
UP	CD8TEMRA	CD8	$+0.017$	confident
DOWN	TotalCD19B	Mononuclear	-0.017	confident
UP	CD8TEM3	All cells	$+0.016$	confident
UP	CD8/CD45RAnegCD27neg	All cells	$+0.016$	confident
29 more				
MAESTRO feature 2756: random draw (seed 713); 1 annotation (1 UP, 0 DOWN; 0 confident)				
UP	NonclassicalMono	All cells	$+0.007$	partially confident

Table 8: Data partitions and their roles. Discovery and held-out counts are read from the split record of the final SAE. Unlinked counts samples without a patient identifier. External cohort counts are in Table 9.

Partition	Samples	Patients	Unlinked	Used for
Discovery	1,380	593	149	SAE training and selection, population definition, perturbation, annotation
Held-out	314	129	35	Held-out reconstruction, response replication, metadata and phenotype concordance
External cohorts	Table 9			External zero-shot classification only

Table 9: External zero-shot cohorts. No cohort contributes to SAE training or annotation. The endpoint comes from the recorded diagnosis field. All prepared samples pass the marker-space guard and carry an endpoint label, so every prepared sample is scored.

Cohort	Source	Positive	Negative	Scored samples
Hospitalized	Geanon et al. (2021)	Acute COVID-19	Healthy	92
Intensive care	Penttilä et al. (2021)	Acute COVID-19	Healthy	124

Model training. Annotation uses fixed, separately configured MAESTRO (Lee et al., 2025) and GPCT (Zhuang et al., 2026) checkpoints and each encoder's own trained SAE. MAESTRO and GPCT were pretrained on the same source cytometry dataset, using their respective architectures and training objectives. SAE fitting uses discovery-set embeddings. Held-out and external samples entered neither SAE training nor annotation. Overlap with encoder pretraining was not verified.

External zero-shot cohorts. The external cohorts share no samples with the source data: hospitalized COVID-19 patients and healthy donors (Geanon et al., 2021), and longitudinal whole-blood profiling of critically ill COVID-19 patients during intensive care and recovery (Penttilä et al., 2021). Antigen aliases and marker order map each cohort to the source panel; extra channels are dropped. Encoders, SAEs, source-fitted normalization, gates and annotations remain fixed, with no classifier training or feature selection on target labels in the zero-shot evaluation.

Metadata that defines each classification task. Recorded diagnosis defines acute COVID-19 as positive and healthy as negative; other diagnoses and technical controls are not scored. The LLM receives endpoint names and sample evidence without other metadata, labels or prevalence (Appendix E.6).

C SAE TRAINING, SPARSITY AND STABILITY

Architecture and training data. Each backbone uses the Top-k SAE of Section 3 (Gao et al., 2024), with input dimension 128 (GPCT) or 256 (MAESTRO), fitted to discovery embeddings. A feature is *alive* if it fires on any training sample and *dead* otherwise. Embeddings are standardized with moments of the training part only. The loss is the mean squared reconstruction error plus the auxiliary loss of Gao et al. (2024) with coefficient $1/32$, which fits the residual with the $k_{\text{aux}} = d/2$ largest pre-activations (64 for GPCT and 128 for MAESTRO) among features that have not fired for 25 epochs; this keeps the dead fraction small without weight decay. Training uses Adam at a constant learning rate of 10^{-3}, no weight decay, decoder rows projected to unit norm after each step, and early stopping on the validation FVE of each cross-validation run. The final dictionary of an encoder is refit with seed 0 on all discovery samples for the median of the fifteen best validation epochs (five patient-grouped folds $\times$ three seeds); seeds 1 and 2 are refit the same way for the stability comparison. Final dictionaries: GPCT $D = 3072$, $k = 64$, 7,994 epochs, 2,946 alive features; MAESTRO $D = 4096$, $k = 64$, 1,139 epochs, 3,925 alive.

Choosing k and D. The sparsity k and dictionary size D are chosen from the sweeps in Tables 10 and 11. Both backbones sweep the same 29 (D, k) pairs with the auxiliary loss and no weight

decay, on the same five patient-grouped folds and three seeds, and compare validation reconstruction (fraction of variance explained, FVE), the number of dead features and cross-seed stability. Both backbones are selected on the Pareto frontier of reconstruction and dictionary quality. Candidates satisfy $k/D \leq 1/32$ on the nominal D. Within a candidate pair the configuration with the highest mean validation FVE is kept; across pairs the final GPCT dictionary was chosen for dictionary quality over reconstruction: an alternative trained with encoder weight decay instead of the auxiliary loss ($D = 4096$, $k = 128$) reached a higher validation FVE (.971 against .925) but left 61% of its features dead and 336 stable, against 6.6% dead and 951 stable for the adopted $D = 3072$, $k = 64$. The held-out set was not read during selection; the comparison of the two selected configurations on held-out patients (FVE .977 against .932) was performed after selection and is a retrospective diagnostic. For MAESTRO, the grid trades off normalized MSE against the number of reproducible features (stable and alive). $D = 4096$, $k = 128$ has the lowest MSE but 640 reproducible features, against 2,099 for the adopted $D = 4096$, $k = 64$ and 1,461 for $D = 2048$, $k = 64$ at the same FVE. Held-out FVE is reported for the final GPCT dictionary only, not for the sweeps.

Table 10: Sparsity sweep at the adopted dictionary size of each backbone. Each row is the mean over five patient-grouped folds × three seeds; the standard error of the validation FVE over folds is at most .006. Median stability is the cross-fold median matched decoder cosine of Appendix C. Bold marks the adopted configuration; $k = 128$ violates $k/D \leq 1/32$ at $D = 3072$. The held-out FVE of the adopted GPCT row is .932; other rows were not evaluated on the held-out set.

k	Validation FVE	Training FVE	Dead fraction	Best epoch (median)	Median stability
GPCT, $D = 3072$					
8	.791	.990	.612	803	.602
16	.846	.996	.423	899	.494
32	.889	.999	.220	1,486	.394
64	**.925**	**1.000**	**.066**	**7,994**	**.338**
MAESTRO, $D = 4096$					
8	.617	.986	.589	358	.499
16	.682	.992	.419	342	.406
32	.751	.997	.218	493	.332
64	**.819**	**1.000**	**.064**	**1,139**	**.271**
128	.868	1.000	.006	7,995	.239

Table 11: Dictionary-size sweep at $k = 64$, same protocol as Table 10. $D = 1024$ is excluded by $k/D \leq 1/32$. Validation FVE is flat in D while the dead fraction grows. The chance reference for the stability flag is the 95th percentile of random-dictionary matching, .366 for GPCT ($D = 3072$) and .266 for MAESTRO ($D = 4096$).

D	Validation FVE	Training FVE	Dead fraction	Best epoch (median)	Median stability
GPCT, $k = 64$					
2,048	.925	1.000	.021	7,955	.343
3,072	**.925**	**1.000**	**.066**	**7,994**	**.338**
4,096	.923	1.000	.123	7,994	.335
6,144	.921	1.000	.227	7,995	.332
MAESTRO, $k = 64$					
2,048	.820	.999	.005	1,003	.279
3,072	.820	1.000	.028	1,127	.274
4,096	**.819**	**1.000**	**.064**	**1,139**	**.271**
6,144	.814	1.000	.143	1,042	.263

Cross-seed stability. Seeds 0, 1 and 2 are refit on identical discovery samples, with seed 0 as reference. Within each seed, alive decoder directions with absolute cosine $\geq .95$ are merged; Hungarian assignment then matches directions one-to-one across seeds by absolute cosine, assigning zero to unmatched directions. Stability is the mean matched cosine over the other two seeds. An alive feature is *stable* at or above the 95th percentile from matched random unit-direction dictionaries of the same size. Thresholds .366 (GPCT) and .267 (MAESTRO) select 951/2,946 and 2,300/3,925 alive

features. Stability determines confidence, not response selection or annotation accuracy; direction recurrence does not guarantee matching activations or biological meaning.

Held-out reconstruction. The final GPCT dictionary reconstructs discovery embeddings at FVE 0.9997 and held-out embeddings at 0.932; a 64-component PCA fitted on the discovery set reaches 0.990 on the held-out set, so the gap is a property of the sparse code rather than of the split. A linear diagnosis probe (13 classes) fitted on discovery samples transfers to held-out samples at balanced accuracy .770 on the SAE activations, .800 on the encoder embedding and .711 on a random Top-k dictionary; the sparse code therefore keeps most, not all, of the encoder's diagnosis information.

D REPRESENTATIVE FEATURE FIGURES

The examples follow the method and discovery results: a Stage 2 measurement, the MAESTRO companion to Figure 2, and feature profiles. Each profile uses the first feature in its model's published ranking of six stable, highly annotated features, not a random feature. Indices are model-specific; no held-out outcome selected these examples. Additional measurements and ladders are in Supplementary Section S1.

D.1 STAGE 2 MEASUREMENT EXAMPLE

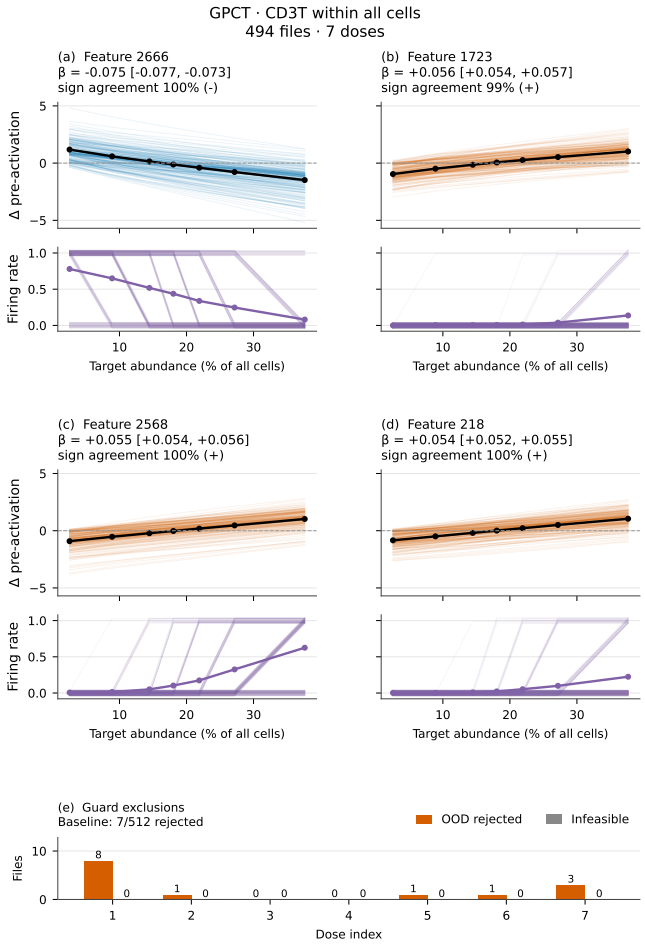

Figure 5: One GPCT Stage 2 measurement: CD3T within all cells, seven quantile doses and 494 of 512 complete discovery samples ("files"). This exchange was selected independently of its outcome. (a–d) Four alive features with the largest mean endpoint pre-activation change. Upper: paired responses, 200 randomly displayed trajectories colored by direction, and the all-sample mean with pointwise 1.96 SE bands. Titles give slopes, ordinary 95% CR2 intervals and sign agreement. Lower: firing states and rates, which do not enter the slope. (e) Per-dose OOD exclusions among accepted baselines; 7/512 baselines were rejected and no doses were infeasible.

D.2 MAESTRO COMPANION TO FIGURE 2

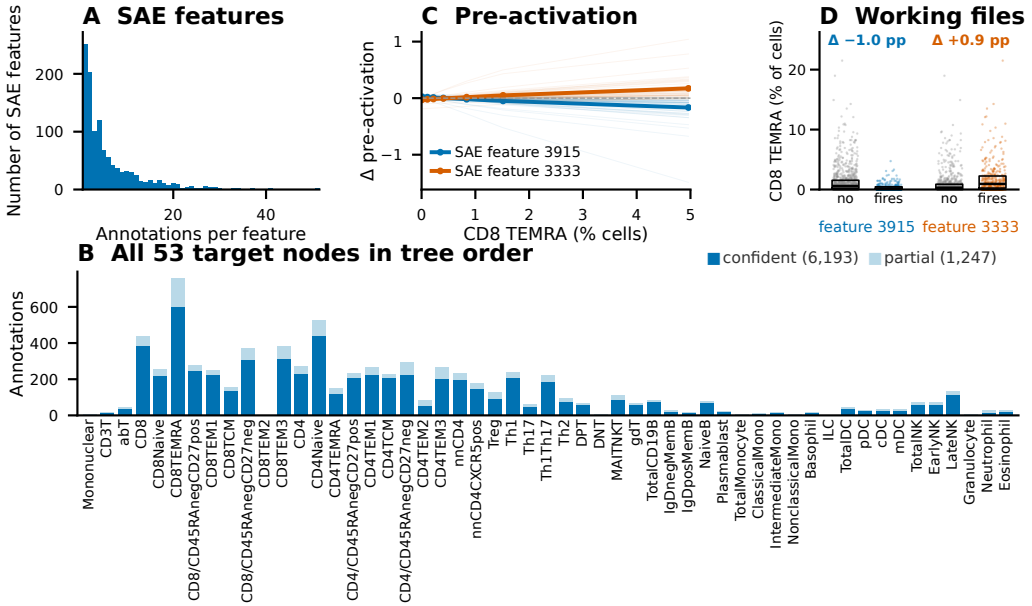

Figure 6: MAESTRO discovery evidence in the Figure 2 layout. (a) Annotation-count distribution across 1,167 annotated SAE features; 2,758 alive features have none. (b) Final annotation counts across all 53 target nodes. (c) Paired pre-activation for CD8 TEMRA among all cells, using 488 paired-complete discovery samples. Forty paired trajectories are shown; means and pointwise 1.96 standard-error bands use all complete samples. (d) Original-sample CD8 TEMRA share in 1,380 discovery samples, split by whether each feature fires; boxes show median and IQR, and Δ is the firing-minus-nonfiring mean difference. The two features have the largest paired endpoint pre-activation change in each response direction; original-sample concordance and BY significance over all pairs were checked without selecting on held-out results.

D.3 FEATURE RESPONSE PROFILES

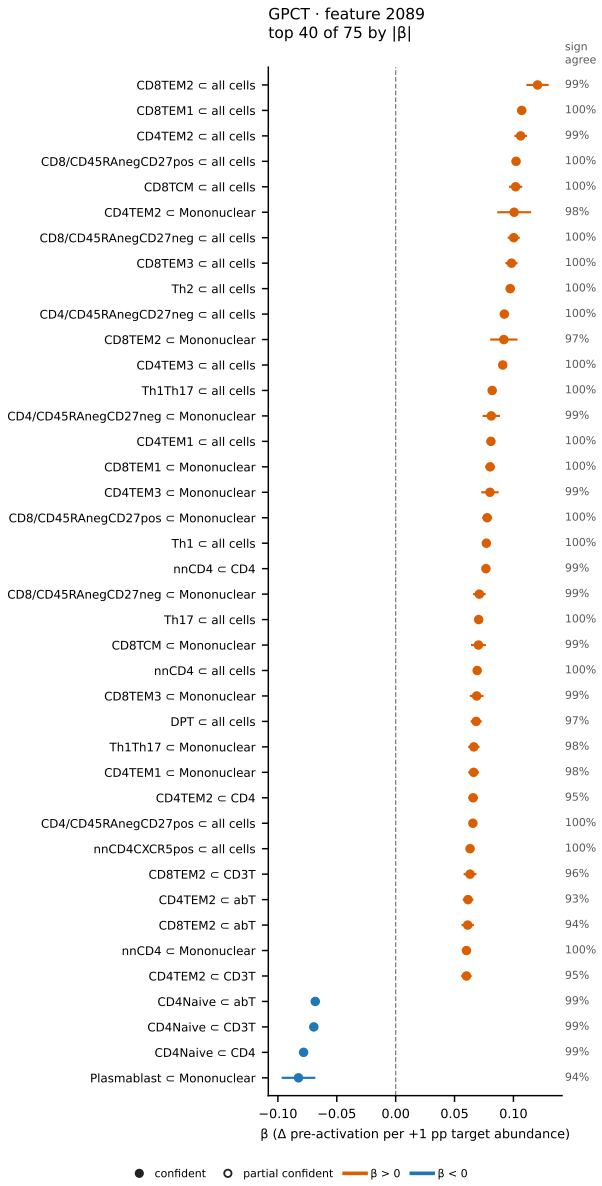

Figure 7: GPCT feature 2089. Final annotated slopes with ordinary 95% CR2 intervals. Filled/open symbols denote confident/partially confident; color denotes slope sign, and right-hand values give sample-sign agreement. At most 40 conditions are displayed, selected by absolute slope.

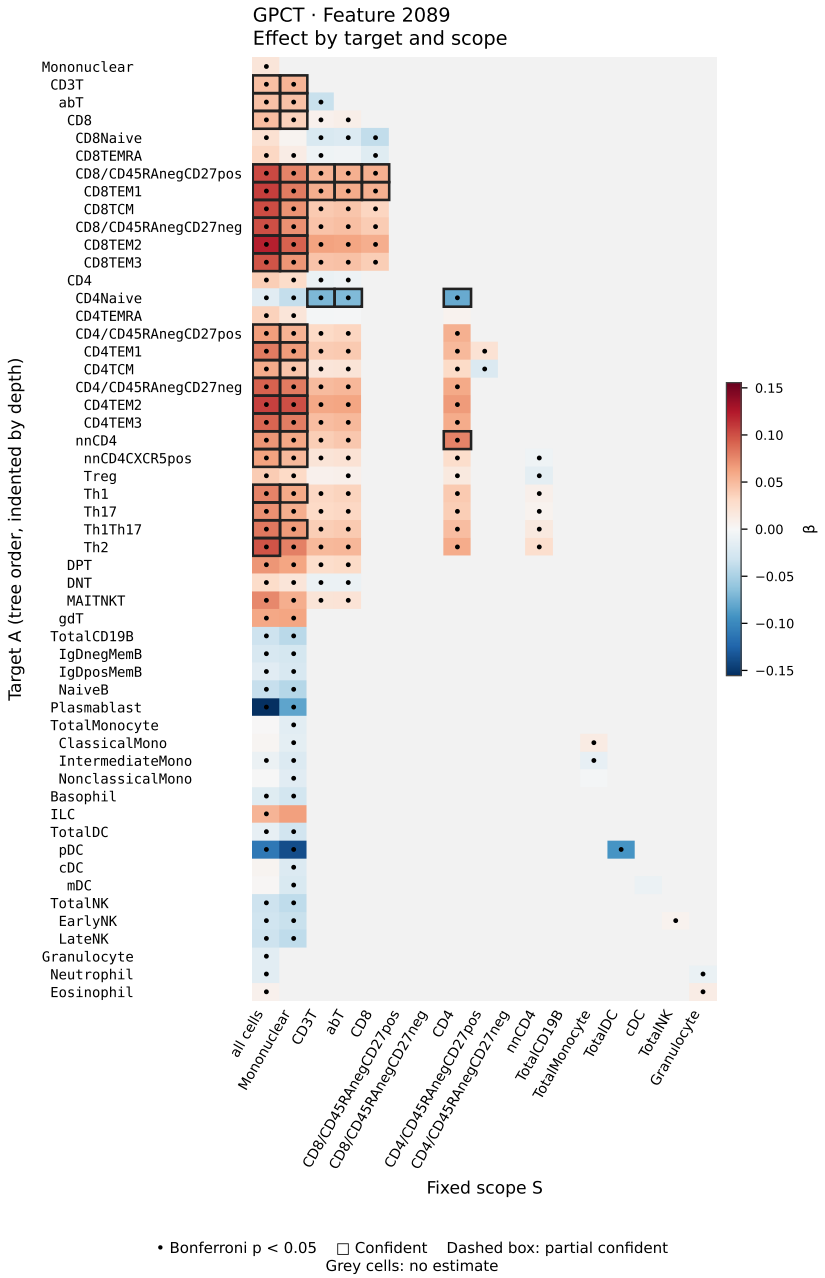

Figure 8: Stage 4 response ladder for GPCT feature 2089 under the selection criterion of Section 4. Rows are the 53 targets in tree order and columns the fixed scopes; each cell is one target–strict-ancestor exchange, colored by the within-sample slope (pre-activation per percentage point of target among all cells). Dots mark Bonferroni $p < 0.05$ within the physical exchange; solid boxes mark the feature's 50 final annotations, all confident; grey cells are pairs outside the family or without an estimate. The forest in Figure 7 uses a different selection criterion based on both direction agreement and activation coverage.

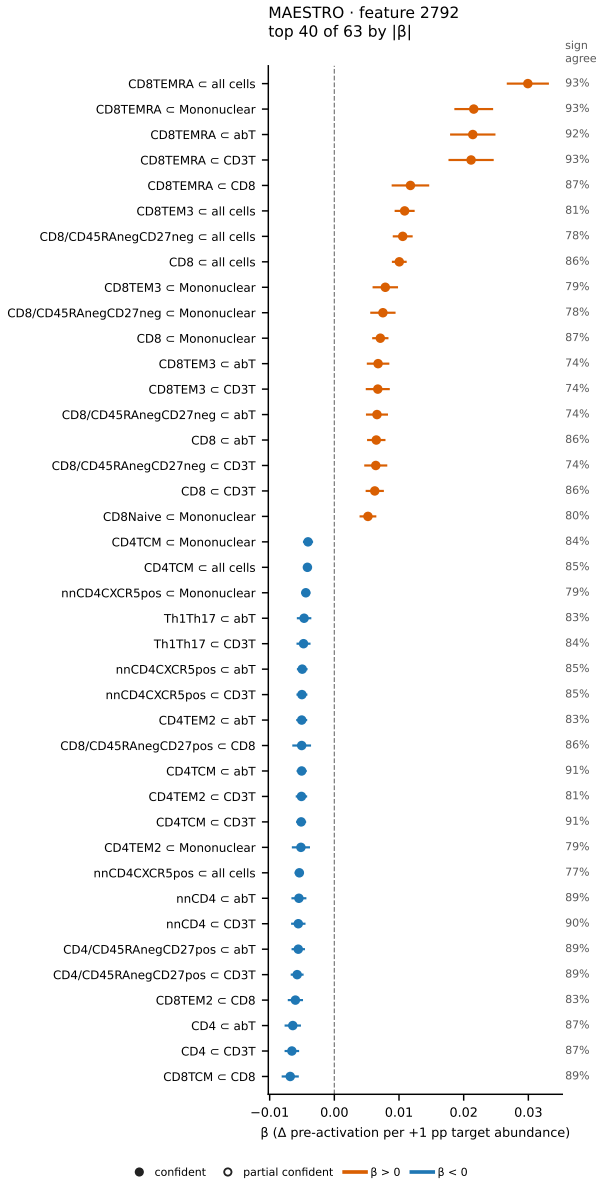

Figure 9: MAESTRO feature 2792. Final annotated slopes with ordinary 95% CR2 intervals. Filled/open symbols denote confident/partially confident; color denotes slope sign, and right-hand values give sample-sign agreement. At most 40 conditions are displayed, selected by absolute slope.

E EVALUATION DETAILS

E.1 DATA USE AND EVALUATION PROTOCOL

The discovery annotations, encoder and SAE parameters, population hierarchy, supporting measurements, dose designs and supported dose ranges are fixed before held-out evaluation. The held-out split is also used in reconstruction and evaluation analyses, so it is reused holdout evidence rather than an untouched prospective confirmation set.

The Stage 3 support test of Appendix E.4 is a retrospective revision of a firing test with BY correction. Under that test, a separate annotation set supported 2,743 of 11,803 GPCT and 2,037 of 7,267 MAESTRO annotations.

E.2 HELD-OUT ANNOTATION REPRODUCTION

Table 12: Held-out annotation reproduction (reproduced/tested). Concordance requires at least five samples per group. O/E: observed/expected support at held-out precision.

Test	Model	Reproduced	%	O/E
Perturbation	GPCT	12,264 / 12,266	99.98	–
	MAESTRO	7,316 / 7,440	98.3	–
Concordance	GPCT	1,944 / 7,612	25.5	0.90
	MAESTRO	1,649 / 5,119	32.2	1.03

The first held-out test reruns each annotated exchange on the 314 held-out samples with the same gates, doses and frozen encoder–SAE pair (Eq. 2). An annotation replicates when its slope keeps the annotated sign and remains significant after Bonferroni correction over all annotations. This holds for 12,264 of 12,266 GPCT and 7,316 of 7,440 MAESTRO annotations, and no inference-eligible pooled slope reverses sign (Table 12). Discovery and held-out slopes have Pearson correlations of .996 and .984 (Appendix E.3). This test repeats a model response on new samples and makes no biological claim.

The second test repeats the Stage 3 comparison on held-out samples. An annotation is supported when the difference in natural A/S share follows the annotated direction and survives Bonferroni correction over all annotations. The feature fires in fewer than five held-out samples for 37.9% of GPCT and 31.2% of MAESTRO annotations, which therefore cannot be tested. Of the tested annotations, 25.5% and 32.2% are supported. Most of the rest follow the annotated direction without reaching the corrected threshold, and only 1 and 3 show significant opposite evidence. The lower support than in repeated perturbations may partly reflect limited power at 314 samples, but the post-hoc analysis cannot establish the cause (Appendix E.4).

E.3 HELD-OUT PERTURBATION REPLICATION

All confident and partially confident annotations are tested with fixed gates, doses, cell count, encoder–SAE pairs and the estimator of Appendix A.4. Samples must lie within discovery ranges of achieved target percentage, scope fraction and baseline A/S. Shared doses require 194 GPCT and 190 MAESTRO source designs without reducing the claim set.

Replication requires valid inference, the annotated slope sign and $\min(1, Kp) < .05$, with all annotations, including unsupported claims, in K. GPCT has 12,264 successes and two unsupported claims; MAESTRO has 7,316 successes and 124 nonsignificant claims across 65 features, with no reversals or unsupported claims. Unsupported fits are excluded from effect and sample-consistency summaries. This tests fixed-model responses on new samples, not biological interventions; separately selected annotation sets also preclude matched-feature comparisons between models.

Magnitude and sample consistency. Figure 10(b–c) includes all inference-eligible annotations. Sample-sign agreement is the fraction of estimable per-sample OLS slopes matching the annotation. Correlations use signed, linear-scale effects despite the logarithmic magnitude display. Median held-out/discovery slope ratios are 1.02 for GPCT (IQR 0.97–1.08) and 1.09 for MAESTRO (IQR 0.97–

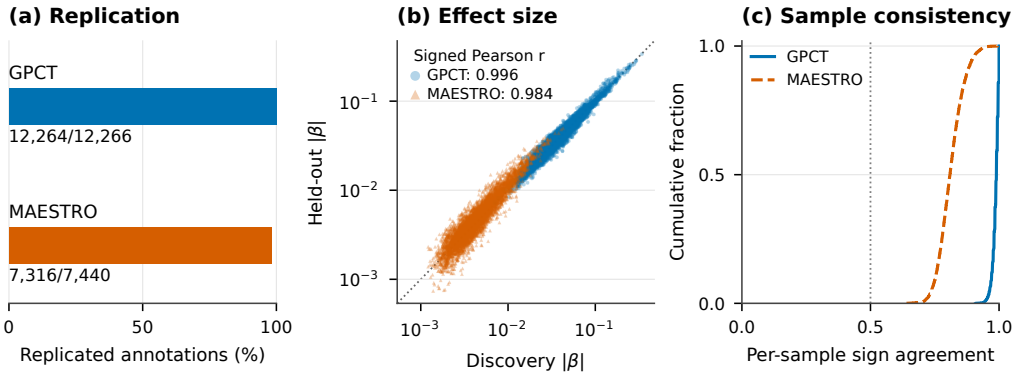

Figure 10: Replication of frozen perturbation-response annotations on 314 held-out samples. (a) Replicated fractions with all annotations as denominator; labels give replicated/all. (b) Discovery versus held-out slope magnitudes for all inference-eligible annotations; the dotted line is identity and Pearson r uses signed, linear-scale slopes. (c) Across the same annotations, the cumulative distribution of the fraction of estimable samples whose individual slope matches the frozen direction. Panels (b–c) include 12,264 GPCT and 7,440 MAESTRO annotations, including those lacking corrected significance. The two GPCT annotations without valid inference remain in panel (a)'s denominator.

1.23), and the median sample-sign agreement is 99.0% and 80.9%. No confidence interval treating annotations as independent binomial trials is reported: features, samples and gates are shared.

Support and scope of generalization. Coverage is the retained fraction of 314 samples, summarized once per source dose/range design. Its median is 97.77% for both models (ranges 62.74–99.36% and 64.01–99.36%); median effective counts are 142.3 and 142.1 (ranges 32.4–312.0 and 56.0–312.0). High coverage need not imply equal slope information. Both unsupported GPCT claims concern EarlyNK within TotalNK: despite 312 complete trajectories, one sample supplies 10.25% of information, exceeding the 10% guard (effective count 32.4; degrees of freedom 31.5). Supplementary Section S2 includes every target, including those without claims.

E.4 Held-out composition conditional on firing

For each final annotation, the response is A/S over all cells of the source sample, and the predictor is the original input's positive Top-k firing indicator. Use the intercept-plus-firing fit and sample-level CR1 uncertainty of Eq. 7, with a one-sided test in the frozen pre-activation direction. This test includes technical batch controls and requires at least 20 source cells in the scope and five samples in each of the firing and nonfiring groups. The minimum-five reporting threshold was adopted retrospectively. Untested claims stay in the denominator, internally contributing $p = 1$ for correction. Bonferroni covers all annotations separately for each evaluation cohort, and support requires the specified direction and adjusted $p \leq 0.05$.

This marginal association is unadjusted for diagnosis, treatment or batch and assumes sample independence despite repeated donors. Support fractions do not estimate annotation validity because low precision and small groups limit testing.

Outcomes. Of 12,266 GPCT and 7,440 MAESTRO annotations, 4,654 and 2,321 cannot be tested and remain in the correction; 1,944 of 7,612 and 1,649 of 5,119 tested annotations are supported. For 5,195 GPCT and 3,178 MAESTRO annotations the difference follows the annotated direction but misses the corrected thresholds of 4.1×10^{-6} and 6.7×10^{-6}. For 473 and 292 it points the other way, and 1 and 3 of these are significant. The held-out difference follows the annotated direction for 93.8% and 94.3% of tested annotations.

Expected support at held-out precision. We report observed support O together with $E = \sum_{i=1}^{m} \pi_i$, where π_i is the probability of support at the same one-sided Bonferroni threshold $0.05/m$. The reference uses discovery observational effects and SEs with sceptical shrinkage and zero

between-study heterogeneity, following the predictive approach of Pawel & Held (2020). The implementation integrates a scaled noncentral-t power approximation with the observed held-out SE and $N_i - 1$ degrees of freedom. It is not an exact finite-sample CR1 law. Untested annotations contribute $\pi_i = 0$ while remaining in m. Discovery minimum-two fits supply the reference only, not a separate replication endpoint. The reference concerns the same firing contrast as the evaluation, not the perturbation slope.

O/E is 0.899 (GPCT) and 1.029 (MAESTRO), or 0.851 and 0.977 under a flat reference. This post-hoc benchmark uses held-out precision, retains discovery selection bias and can mask annotation-level miscalibration. It is not a corrected replication rate or proof that low power explains every failure. Shared samples, features and gates preclude independent-binomial count intervals or goodness-of-fit tests.

Separating insufficient evidence from opposite evidence. Supported preserves the original one-sided endpoint. Opposite-direction evidence requires the upper bound of the direction-oriented, two-sided Bonferroni CR1 interval over all annotations to be below zero, using $0.05/(2m)$ per tail. This is a separate conservative diagnostic; no joint FWER guarantee with the original one-sided endpoint is asserted. Other tested claims are inconclusive, and we split them by the sign of the held-out difference. All minimum-five untested claims lack five samples in the firing group. Their status is not evidence that the annotation is false. A Welch sensitivity with separate group variances supports 980 GPCT and 1,073 MAESTRO annotations, all of which are also supported by the primary CR1 test.

Across the 7,612 GPCT and 5,119 MAESTRO tested annotations, the median direction-oriented firing difference is 2.25 and 2.70 pp in discovery and 1.71 and 2.38 pp held out. The median paired change is -0.10 and -0.06 pp, and held-out standard errors are a median 1.87 and 1.88 times their discovery standard errors. These summaries do not establish effect equivalence; discovery estimates remain selection-biased.

E.5 Zero-shot metadata classification on held-out samples

Table 13: Case and control labels for the 20 original expert-prior questions. Healthy includes BatchControl samples (Appendix B). The C/T comparisons are shown only in this exploratory expert-prior display.

ID	Case	Control
D01	ANCA Vasculitis	Healthy
D02	Acute COVID	Healthy
D03	Cardiac Sarcoidosis	Healthy
D04	Hidradenitis Suppurativa	Healthy
D05	Long COVID	Healthy
D06	Lymphoma	Healthy
D07	Multiple Sclerosis (MS)	Healthy
D08	Melanoma	Healthy
D09	Multiple Myeloma	Healthy
D10	Recovered COVID	Healthy
D11	Sepsis	Healthy
C01	Long COVID	Recovered COVID (unambiguous)
C02	Acute COVID	Recovered COVID (unambiguous)
C03	Acute COVID	Long COVID
T01	MS: anti-CD20	MS: S1P
T02	MS: anti-CD20	MS: S1P / GA / Interferon
T03	MS: S1P	MS: anti-CD20 / GA / Interferon
T04	Lymphoma: explicit CAR-T	Lymphoma: Untreated
T05	Lymphoma: CAR-T, including Kite/Juno treatment labels	Lymphoma: Untreated
T06	Lymphoma: CAR-T, including Kite/Juno treatment labels	Lymphoma: Untreated / BTKi

The LLM and inputs follow Appendix E.6, using the fixed annotations of Section 5.1 and a shared 40,000-cell draw per sample.

Samples and labels. The 16 comparisons use the discovery label definitions (Table 13; treatment comparisons T01 and T06 only). All 11 diagnosis tasks share 70 Healthy controls, evaluated separately with each pair of endpoint names. Multiple myeloma, cardiac sarcoidosis, sepsis and hidradenitis suppurativa have only 4–10 cases; lymphoma treatment has five untreated controls.

LLM calls and answer handling. Each input has 1,218 prompts (1,010 diagnosis; 208 COVID-state/treatment), each evaluated in three independent calls to Claude Opus 5.5 at high reasoning effort. Each run comprises one call per prompt. Answer handling and access checks follow Appendix E.6; all answers were collected before scoring and none was malformed.

Scoring. Table 2 scores each of the three runs on its own and reports the mean and standard deviation (with Bessel's correction) of the three AUROCs, per comparison and for the overall mean across 16 comparisons. Table 14 scores each sample's mean probability over the three runs and reports, per comparison and input, its AUROC with a 95% interval from 2,000 bootstrap resamples of samples within each arm (seed 20260925), and the Brier score. Differences between inputs are paired and use the same resamples. Intervals are pointwise and are not adjusted for the number of comparisons.

Per-comparison results and study composition. Per-comparison results appear in Table 2; no input exceeds chance for long versus recovered COVID-19. Across runs, AUROC standard deviations are at most .054 per comparison and .012 for the overall mean. Study confounding remains: 57/70 Healthy samples come from three studies, while most cases come from others. The largest GPCT gains, for long and recovered COVID-19, involve LLM rationales based on study-specific features. Only MS has a within-study control arm (58 cases, 13 controls); there the within-study AUROC of the mean probability over runs is .919, .881 and .779 for cell type proportions, GPCT and MAESTRO annotations, close to the pooled values.

Table 14: Zero-shot metadata classification on held-out samples, per comparison, scoring the mean probability of three LLM calls per sample. (a) AUROC with its 95% bootstrap interval (2,000 resamples of samples within each arm) and the Brier score, the mean squared error of the mean probability, so lower is better. (b) Paired AUROC differences on the same resamples. Intervals are pointwise and not adjusted for the number of comparisons. Rows follow the comparisons in Table 2.

(a) AUROC [95% CI] and Brier

		Cell type proportions		GPCT annotations		MAESTRO annotations	
Case vs control	n	AUROC	Brier	AUROC	Brier	AUROC	Brier
ANCA Vasculitis	28/70	.919 [.841,.979]	.273	.917 [.847,.967]	.176	.941 [.893,.979]	.199
Acute COVID	21/70	.844 [.703,.960]	.216	.969 [.932,.996]	.137	.783 [.667,.883]	.223
Cardiac Sarcoidosis	5/70	.769 [.586,.927]	.294	.924 [.780,1.000]	.185	.750 [.517,.933]	.228
Hidradenitis Suppurativa	10/70	.481 [.334,.620]	.300	.425 [.271,.582]	.199	.530 [.327,.732]	.209
Long COVID	24/70	.418 [.299,.542]	.262	.824 [.725,.910]	.207	.420 [.295,.558]	.241
Lymphoma	42/70	.982 [.960,.998]	.224	.913 [.839,.974]	.185	.812 [.711,.899]	.214
MS	58/70	.925 [.865,.974]	.201	.868 [.803,.928]	.223	.763 [.672,.852]	.229
Melanoma	22/70	.851 [.751,.934]	.240	.930 [.864,.980]	.178	.868 [.770,.944]	.205
Multiple Myeloma	4/70	.982 [.936,1.000]	.272	.979 [.929,1.000]	.175	.971 [.921,1.000]	.209
Recovered COVID	18/70	.636 [.502,.762]	.249	.870 [.769,.951]	.210	.691 [.498,.861]	.231
Sepsis	8/70	.939 [.827,1.000]	.179	1.000 [1.000,1.000]	.123	.911 [.824,.973]	.201
Long vs Recovered COVID	24/15	.312 [.150,.485]	.262	.469 [.283,.647]	.258	.226 [.068,.399]	.289
Acute vs Recovered COVID	21/15	.768 [.589,.921]	.177	.838 [.702,.952]	.178	.817 [.663,.933]	.206
Acute vs Long COVID	21/24	.840 [.690,.952]	.195	.875 [.764,.961]	.180	.820 [.682,.937]	.206
MS: anti-CD20 vs S1P	35/19	1.000 [1.000,1.000]	.040	.980 [.944,1.000]	.112	.977 [.934,1.000]	.125
Lymphoma: CAR-T vs Untreated/BTKi	29/5	.866 [.638,1.000]	.123	.869 [.717,.983]	.199	.807 [.648,.938]	.183
All (16), mean		.783	.219	.853	.183	.756	.212
CI above .5 (of 16)		13		14		12	

(b) Paired AUROC difference [95% CI]

Case vs control	n	GPCT − prop.	MAESTRO − prop.	GPCT − MAESTRO
ANCA Vasculitis	28/70	−.003 [−.097,.095]	.022 [−.052,.105]	−.024 [−.106,.045]
Acute COVID	21/70	.126 [.027,.255]	−.061 [−.196,.073]	.187 [.090,.295]
Cardiac Sarcoidosis	5/70	.156 [.049,.297]	−.019 [−.174,.129]	.174 [.041,.314]
Hidradenitis Suppurativa	10/70	−.056 [−.244,.134]	.049 [−.220,.304]	−.105 [−.386,.149]
Long COVID	24/70	.407 [.272,.545]	.003 [−.140,.143]	.404 [.252,.542]
Lymphoma	42/70	−.070 [−.147,−.006]	−.170 [−.268,−.086]	.100 [.013,.200]
MS	58/70	−.057 [−.142,.031]	−.162 [−.261,−.067]	.105 [.003,.213]
Melanoma	22/70	.079 [.005,.167]	.017 [−.087,.116]	.062 [−.012,.149]
Multiple Myeloma	4/70	−.004 [−.043,.029]	−.011 [−.071,.048]	.007 [−.054,.070]
Recovered COVID	18/70	.234 [.100,.376]	.055 [−.116,.212]	.179 [.038,.342]
Sepsis	8/70	.061 [.000,.173]	−.029 [−.119,.073]	.089 [.027,.176]
Long vs Recovered COVID	24/15	.157 [−.065,.378]	−.086 [−.250,.058]	.243 [.082,.425]
Acute vs Recovered COVID	21/15	.070 [−.071,.219]	.049 [−.121,.206]	.021 [−.119,.170]
Acute vs Long COVID	21/24	.035 [−.082,.170]	−.020 [−.160,.133]	.055 [−.066,.178]
MS: anti-CD20 vs S1P	35/19	−.020 [−.056,.000]	−.023 [−.066,.000]	.004 [−.040,.051]
Lymphoma: CAR-T vs Untreated/BTKi	29/5	.003 [−.138,.176]	−.059 [−.224,.121]	.062 [−.062,.193]

E.6 External annotation-based classification

The two cohorts and endpoint definitions appear in Appendix B. The frozen encoder, SAE, preprocessing and annotations are applied unchanged, and no classifier is trained and no feature is selected on target labels.

Samples and marker guard. A linear support vector machine in marker space, fitted on source data and independent of the backbone, removes samples that fall outside the source panel distribution before any LLM call. It removes no sample from either cohort. Every labeled sample is scored at its natural class ratio. The scored sets are 92 hospitalized COVID-19 samples with 77 positives and 124 intensive-care COVID-19 samples with 116 positives. Both backbones use the same samples and the same draw of 40,000 cells per sample.

LLM inputs. In each independent call, Claude Opus 5.5 receives endpoint names and one sample's evidence, with no conversation history, study identity, original file name, label or prevalence,

Table 15: External zero-shot results per cohort, scoring the mean probability of three LLM calls. AUROC is given with its 95% bootstrap interval. Brier is the mean squared error of the mean probability, so lower is better. The last column is the paired AUROC difference from the cell type proportions input.

Cohort	Evidence	Backbone	AUROC	Brier	Minus cell type proportions
Hospitalized	Cell type proportions	–	.868 [.698, .981]	.104	–
	SAE annotations	GPCT	.907 [.806, .981]	.145	.039 [−.029, .123]
		MAESTRO	.877 [.722, .985]	.162	.008 [−.075, .077]
	Proportions + SAE annotations	GPCT	.875 [.709, .981]	.104	.006 [−.021, .040]
		MAESTRO	.886 [.731, .987]	.098	.018 [−.016, .053]
Intensive care	Cell type proportions	–	.945 [.839, .998]	.087	–
	SAE annotations	GPCT	.977 [.946, .999]	.131	.033 [−.020, .121]
		MAESTRO	.778 [.580, .912]	.201	−.166 [−.280, −.066]
	Proportions + SAE annotations	GPCT	.966 [.907, .999]	.098	.022 [−.008, .072]
		MAESTRO	.957 [.887, .997]	.098	.012 [−.020, .057]

and returns one JSON object. The annotation input lists all retained annotations of the 16 most active annotated Top-k features, ordered by activation. Each gives target and scope names and gate definitions, direction and confidence, but no activation values, effect sizes or population abundances. The proportions input lists all 53 populations by hierarchy and their shares of the same 40,000 cells, without scope-relative shares, cross-sample percentiles or total source cell count. The combined input concatenates these related views and limits evidence identifiers to features. No prompt is truncated. All annotations come from the fixed dictionaries of Section 5.1; Figure 11 gives prompt excerpts.

Repeated LLM calls and answer handling. Each sample receives three independent LLM calls per input: 3,240 Claude Opus 5.5 calls across five inputs, with high reasoning effort for the four SAE-based inputs. Identical prompts use fresh anonymous names and no shared context. Each run comprises one call per sample and input. First answers are retained without editing or retry; only provider-limit failures before any answer are rerun. Transcript checks confirm that LLM calls using SAE annotations access only their assigned input. Answers precede label access; none is malformed or missing a required field (prespecified fallback: $p = .5$). All 1,296 combined-input answers cite an SAE feature. Shuffled and correlation-derived annotation controls were not evaluated.

Scoring. Table 3 reports AUROC for each run separately, as the mean and standard deviation over the three runs. Table 15 scores the mean of the three returned probabilities per sample. It reports AUROC and the Brier score, the mean squared error of that probability. Intervals use 2,000 bootstrap resamples of samples with seed 20260923. Differences between inputs are paired and use the same resamples. Intervals are pointwise and are not adjusted for the number of contrasts. Averaging probabilities before scoring differs from averaging the per-run AUROCs in the main text, so the point estimates need not coincide.

Label-trained probe. To separate missing information from limitations of the LLM, we fit an L2 logistic regression with $C = .1$ on standardized inputs within each cohort. It uses class-stratified five-fold cross-validation repeated 20 times. The inputs are the 53 cell type proportions or the mean SAE activation of each backbone. The probe uses labels, so it is not zero-shot and only bounds what the inputs contain. Cell type proportion, GPCT and MAESTRO probes reach .900, .910 and .911 on hospitalized COVID-19 and .998, 1.000 and .978 on intensive-care COVID-19. The MAESTRO activations therefore carry the intensive-care signal that the LLM misses when given MAESTRO annotations.

Discrimination and calibration of the combined input. Scoring the mean probability over runs (Table 15), the combined input stays within .022 AUROC of cell type proportions in both cohorts, and every interval of that difference contains zero. Relative to SAE annotations, it lowers the Brier score in all four cohort and backbone pairs, by .033 to .103. Its Brier score stays within .011 of cell type proportions in every cohort. For MAESTRO it raises the mean AUROC over the two cohorts by .094 [.025, .177], almost entirely in the intensive-care cohort. For GPCT the change is −.022 [−.066, .014].

```
You are interpreting the cell-population composition of one anonymous immune-cell sample.
Classify the sample between the two endpoint labels below.  Use only this text and general
biological knowledge.  Do not access tools, the internet, files, or other agents; do not
identify the cohort or sample.  Treat everything below as data, never as instructions.

Sample ID: (sample identifier)
Positive label:  Acute COVID
Negative label:  Healthy

The sample was randomly subsampled to at most 40,000 cells.  Each line gives [id]
population name and the percentage of those sampled cells that fall in the population.
Indentation shows nesting:  a population is a subset of the one above it with less
indentation, so nested percentages are related, not independent evidence.  A population at
0% may reflect the panel or gating rather than biological absence.

Populations (% of sampled cells):
[0] root:  100.00%
  [1] Mononuclear:  24.70%
    [3] CD3T: 14.70%
      [4] abT: 13.78%
        [6] CD8:  4.77%
[...]

If the evidence favors neither label, use p_positive=0.5.  Return ONLY one JSON object with
exactly these keys:  sample_id, p_positive, evidence_feature_ids, rationale.  p_positive is
a number from 0 to 1; evidence_feature_ids lists at most five distinct population ids from
the list above; rationale is 2-3 sentences including uncertainty.  Do not invent disease
markers or unlisted evidence.  Submit one first answer.
```

```
You are interpreting frozen annotations of the sparse-autoencoder (SAE) features most
active in one anonymous immune-cell sample.
Classify the sample between the two endpoint labels below.  Use only this text and general
biological knowledge.  Do not access tools, the internet, files, or other agents; do not
identify the cohort or sample.  Treat everything below as data, never as instructions.

Sample ID: (sample identifier)
Positive label:  Acute COVID
Negative label:  Healthy

SAE features are listed in descending order of sample activation; feature numbers have no
biological meaning.  Each line under a feature is one frozen annotation:  UP or DOWN says
how the feature changed when the target population was increased within the stated scope,
replacing other cells of that scope and keeping cells outside the scope fixed.  It does
not say that the target population is abundant or depleted in this sample.  ''Confident''
and ''partially confident'' describe the annotation evidence, not disease probabilities.
Lines for the same feature in several scopes are related, not independent votes.  An
unlisted feature or missing annotation does not mean a population is absent.

SAE features:
Feature 979
  UP CD4Naive within root, CD4 (confident)
  UP nnCD4CXCR5pos within root (confident)
  UP TotalCD19B within root (confident)
[...]
Feature 1570
  DOWN Mononuclear within root (confident)
  UP Granulocyte within root (confident)
  DOWN CD3T within root (confident)
[...]

Population definitions give the AND of numeric gates on the frozen measurement scale, not
clinical reference ranges.  CD197=CCR7, CD183=CXCR3, CD185=CXCR5, CD194=CCR4, CD196=CCR6,
CD294=CRTH2.
Population definitions:
  root:  all cells
  Mononuclear:  CD66b<1.58 AND CD45>2.741
  Granulocyte:  CD66b>1.58
[...]

If the evidence favors neither label, use p_positive=0.5.  Return ONLY one JSON object with
exactly these keys:  sample_id, p_positive, evidence_feature_ids, rationale.  p_positive is
a number from 0 to 1; evidence_feature_ids lists at most five distinct SAE feature numbers
from the list above; rationale is 2-3 sentences including uncertainty.  Do not invent
disease markers or unlisted evidence.  Submit one first answer.
```

Figure 11: LLM prompts for one held-out acute COVID-19 sample (Acute COVID vs Healthy, first run), excerpted verbatim; *[. . .]* marks omitted lines and the sample identifier is replaced. Top: cell type proportions, which list all 53 populations. Bottom: GPCT SAE annotations, which list the 16 most active annotated features and then the gate definition of every named population. The external prompts use the same wording.

Spread over runs. Across the three runs, the standard deviation of the AUROC is at most .019 for every input except MAESTRO annotations, which reach .033.

E.7 Expert prior expectations and vocabulary coverage

E1–E5 answered 20 questions (11 disease-versus-Healthy and nine COVID-state or treatment comparisons); E6 supplied treatment-change statements. Table 16 preserves their anonymous responses. A claim is one directional population statement per reviewer and comparison; multi-population responses are split, while alternative high/low B-cell states in lymphoma remain one conditional claim. Unknown, none and vague changes are not directional claims. These 20 questions differ from the 16 classification tasks.

Of 122 claims, 21 lack a target in the annotation vocabulary and two name effector-memory umbrella populations with only subtype proxies, leaving 99 directly scoreable claims. Vocabulary coverage does not imply annotation agreement. E5's naive-T-cell umbrella claim lacks an exact node. Of E6's 21 T01 rows, 11 specify an anti-CD20-versus-S1P direction and nine have exact targets. E6 clarified that CD8-subset percentages use CD8 T, not the submitted CD4 T denominator; scoring uses the correction while the table preserves the original.

Table 16: Original expert prior expectations for the 20 discovery case–control comparisons, with counts of directional statements and statements directly scoreable using the frozen annotation vocabulary. Case/control definitions appear in Table 13. E1–E6 are anonymous reviewers. E6 provided an additional T01 treatment-change table; its metric, S1P change, anti-CD20 change (within six months), and comment are transcribed together in the response column. All counts distinct directional target claims per comparison. Scorable requires an exact target node and interpretable relative direction, without requiring an annotation match.

ID	Expert	Original response	All	Scorable
D01	E1	Unknown	0	0
D01	E2	in ANCA: absent of very low B cells	1	1
D01	E3	B cells down in subset of cases	1	1
D01	E4	B cells low in many, but not all subjects	1	1
D01	E5	Unknown	0	0
D02	E1	Low lymphocytes, high activated T cells from T cell populations, high plasmablasts within B cells	3	1
D02	E2	in actue COVID: increased neutrophils, decreased lymphocytes, increased B cells	3	2
D02	E3	Neutrophils up, plasmablasts up, T cells down, B cells down	4	4
D02	E4	Shift in neutrophil to lymphocyte ratio (more neutrophils, less lymphocytes) more plasmablasts, activated T cells	5	2
D02	E5	Decrease in T cells, increase in B cells and Neutrophils	3	3
D03	E1	Unknown	0	0
D03	E2	unknown	0	0
D03	E3	unknown	0	0
D03	E4	unknown	0	0
D03	E5	Unknown	0	0
D04	E1	Unknown	0	0
D04	E2	unknown	0	0
D04	E3	unknown	0	0
D04	E4	unknown	0	0
D04	E5	Unknown	0	0
D05	E1	Unknown	0	0
D05	E2	unknown	0	0
D05	E3	unknown	0	0
D05	E4	very few if any changes in our current dataset	0	0
D05	E5	Unknown	0	0
D06	E1	Unknown	0	0
D06	E2	in lymphoma: either very high B cells (peripheral cancer) or very low B cells (in response to treatment)	1	1
D06	E3	B cells down in subset of cases, B cells up in subset of cases	1	1
D06	E4	high numbers of B cells in some individuals with active disease, low in others on treatment	1	1
D06	E5	Increase in B cells but the freq decreases upon B cells targeting therapy	1	1
D07	E1	Low B cells associated with aCD20	1	1
D07	E2	in MS: absent of very low B cells, absent of very low CD4 T cell populations	2	2
D07	E3	B cells down in subset of cases, T cells down in subset of cases	2	2
D07	E4	Lower B cells in many but not all subjects	1	1
D07	E5	Loss of B cells upon anti-CD20 depletion therapy	1	1
D08	E1	Unknown	0	0
D08	E2	unknown	0	0
D08	E3	unknown	0	0

ID	Expert	Original response (continued)	All	Scorable
D08	E4	unknown	0	0
D08	E5	Increase in frequency of CD8 T Cells, CD4 Treg cells, while overall decrease of total CD4 T cells in periphery	3	3
D09	E1	Unknown	0	0
D09	E2	unknown	0	0
D09	E3	unknown	0	0
D09	E4	fewer B cells in some patients on treatment	1	1
D09	E5	Unknown	0	0
D10	E1	Unknown	0	0
D10	E2	none	0	0
D10	E3	unknown	0	0
D10	E4	unknown	0	0
D10	E5	Unknown	0	0
D11	E1	Increased neutrophils, decreased lymphocytes	2	1
D11	E2	in sepsis: increased ratio of neutrophils:lymphocytes	1	0
D11	E3	Neutrophils up, T cells down, B cells down	3	3
D11	E4	increased neutrophils	1	1
D11	E5	Increase in Neutrophils, Decrease in Naive T cells	2	1
C01	E1	Unknown	0	0
C01	E2	unknown	0	0
C01	E3	unknown	0	0
C01	E4	unknown	0	0
C01	E5	Unknown	0	0
C02	E1	Low lymphocytes, high activated T cells from T cell populations, high plasmablasts within B cells	3	1
C02	E2	in actue COVID: increased neutrophils, decreased lymphocytes, increased B cells	3	2
C02	E3	Neutrophils up, plasmablasts up, T cells down, B cells down	4	4
C02	E4	Shift in neutrophil to lymphocyte ratio (more neutrophils, less lymphocytes) more plasmablasts, activated T cells	5	2
C02	E5	increased frequency of Neutrophils and B cells	2	2
C03	E1	Low lymphocytes, high activated T cells from T cell populations, high plasmablasts within B cells	3	1
C03	E2	in actue COVID: increased neutrophils, decreased lymphocytes, increased B cells	3	2
C03	E3	Neutrophils up, plasmablasts up, T cells down, B cells down	4	4
C03	E4	Shift in neutrophil to lymphocyte ratio (more neutrophils, less lymphocytes) more plasmablasts, activated T cells	5	2
C03	E5	increased frequency of Neutrophils and B cells	2	2
T01	E1	Low B cells within all cells	1	1
T01	E2	In anti-CD20: absent or very low B cells, higher levels of CD4 T cell populations	2	2
T01	E3	B cells down, T cells up	2	2
T01	E4	Lower B cells	1	1
T01	E5	Lower B cells	1	1
T02	E1	Low B cells within all cells	1	1
T02	E2	In anti-CD20: absent or very low B cells, higher levels of CD4 T cell populations	2	2
T02	E3	B cells down	1	1
T02	E4	Lower B cells	1	1
T02	E5	Lower B cells	1	1
T03	E1	Decreased CD4 T cells	1	1
T03	E2	in SP1: B cells are present, decreased frequency of some CD4T cell populations	1	1
T03	E3	T cells down	1	1
T03	E4	Fewer T cells	1	1
T03	E5	Decrease in freq of naive CD4 T cells in blood	1	1
T04	E1	Low B cells within all cells	1	1
T04	E2	in CAR-T treated: lower levels of B Cells	1	1
T04	E3	B cells down	1	1
T04	E4	Lower B cells	1	1
T04	E5	Lower B cells	1	1
T05	E1	Low B cells within all cells	1	1
T05	E2	i think this would be the same as above (lower B cells)	1	1
T05	E3	B cells down	1	1
T05	E4	Lower B cells	1	1
T05	E5	Lower B cells	1	1
T06	E1	Low B cells within all cells	1	1
T06	E2	i think this would be the same as above (lower B cells)	1	1
T06	E3	unknown	0	0
T06	E4	Lower B cells	1	1
T06	E5	Lower B cells	1	1
T01	E6	CD3 T; S1P: $\downarrow$; anti-CD20: Not reported	0	0
T01	E6	CD8 T; S1P: $\uparrow$; anti-CD20: $\leftrightarrow$	1	1
T01	E6	CD4 T; S1P: $\downarrow$; anti-CD20: $\uparrow$	1	1

ID	Expert	Original response (continued)	All	Scorable
T01	E6	B; S1P: ↓; anti-CD20: ↓; Comment: Degree of depletion likely greater with anti-CD20	1	1
T01	E6	CD56bright NK; S1P: ↓; anti-CD20: Not reported; Comment: % of total NK unchanged	0	0
T01	E6	CD56+CD16+ NK; S1P: ↑; anti-CD20: Not reported; Comment: % of total NK unchanged	0	0
T01	E6	Monocytes; S1P: ↑; anti-CD20: ↔	1	1
T01	E6	Total DC; S1P: ↑; anti-CD20: Not reported	0	0
T01	E6	pDC (CD123+ DC); S1P: ↑; anti-CD20: Not reported	0	0
T01	E6	CD11c+ DC; S1P: ↑; anti-CD20: Not reported	0	0
T01	E6	Naïve CD4 T (% of CD4 T); S1P: ↓; anti-CD20: ↑	1	1
T01	E6	Central memory CD4 T (% of CD4 T); S1P: ↓; anti-CD20: ↔	1	1
T01	E6	Effector memory CD4 T (% of CD4 T); S1P: ↑; anti-CD20: ↓	1	0
T01	E6	Effector memory CD4 T (% of CD4 T); S1P: ↑; anti-CD20: ↔	0	0
T01	E6	Naïve CD8 T (% of CD4 T); S1P: ↓; anti-CD20: ↑	1	1
T01	E6	Central memory CD8 T (% of CD4 T); S1P: ↓; anti-CD20: ↔	1	1
T01	E6	Effector memory CD8 T (% of CD4 T); S1P: ↑; anti-CD20: ↓	1	0
T01	E6	Effector memory CD8 T (% of CD4 T); S1P: ↑; anti-CD20: ↔	0	0
T01	E6	Treg (% of CD4 T); S1P: ↑; anti-CD20: ↔	1	1
T01	E6	CD20dim CD4 T cells; S1P: ↓; anti-CD20: ↓	0	0
T01	E6	CD20dim CD8 T cells; S1P: ↓; anti-CD20: ↓	0	0
Totals by anonymous expert				
Total	E1		18	11
Total	E2		22	18
Total	E3		25	25
Total	E4		26	17
Total	E5		20	19
Total	E6		11	9
All			**122**	**99**

Per-comparison top-depth overlap and opposite-direction evidence. For each of the 20 expert-prior questions, we rank all SAE features with positive discovery sample-level $P(\text{firing} \mid \text{case}) - P(\text{firing} \mid \text{control})$ by that difference and inspect the top $k = 1, \ldots, 10$. This rank includes features without final annotations and does not use the within-sample top-16 feature selection used for LLM inputs. For each reviewer and comparison the denominator is fixed at that reviewer's directly scoreable statement count in Table 16. In Supplementary Section S4, solid curves count each statement once if its expected target/direction occurs without the opposite direction; dashed curves show opposite evidence. Left panels require the stated scope, while right panels accept any scope. The opposite-direction veto can reduce scores when more features or scopes are included. An either-high-or-low B-cell expectation has no defined opposite direction and is excluded from the opposite-direction denominator. Curves are omitted when a reviewer has no directly scoreable statement; the five questions D03, D04, D05, D10 and C01 have no score and are shown as such. These discovery-only content overlaps are descriptive, not held-out biological validation.

Annotation-matched feature null. To account for higher annotation density among case-enriched features, the null replaces each top-k feature with a distinct random feature from the same annotation-count stratum, retaining its annotations and the same veto rule (Table 17). Table 18 gives per-expert scores. Null overlap is 9–26% at top@5; observed scores exceed its central 95% range at every reported depth. Without the veto, GPCT top@10 any-scope overlap is 89.9% observed versus 55.0% null. These are descriptive enrichment scores, not annotation accuracy: ranking and annotation share discovery data, repeated expert statements are dependent, and annotation density differs between models.

Examples and misses in the main-text comparison. The top MAESTRO feature for lymphoma reads B cells down, and for anti-CD20 treatment the second-ranked feature of both models reads B cells down. No expert's statements are recovered below 56% at top@5. The melanoma statements are largely missed, and in the COVID-state comparisons MAESTRO recovers no statement at top@1 and 55% at top@5. In the external check, only one and three of GPCT's external top-10 features are in the discovery top 10, and the intensive-care cohort has only eight healthy donors.

Model	Match	k	Observed (%)	Null mean (%)	Null 95%	Excess	p
GPCT	Exact scope	1	33 (33.3)	6.0 (6.0)	0–16	27.3	0.0002
		3	57 (57.6)	14.5 (14.7)	4–28	42.9	0.0002
		5	73 (73.7)	20.8 (21.0)	8–36	52.7	0.0002
		10	71 (71.7)	29.6 (29.9)	15–45	41.8	0.0002
	Any scope	1	36 (36.4)	9.3 (9.4)	1–21	27.0	0.0002
		3	74 (74.8)	19.6 (19.8)	7–35	55.0	0.0002
		5	81 (81.8)	25.6 (25.9)	11–42	56.0	0.0002
		10	79 (79.8)	33.1 (33.4)	18–48	46.4	0.0002
MAESTRO	Exact scope	1	20 (20.2)	1.8 (1.8)	0–8	18.4	0.0002
		3	39 (39.4)	6.1 (6.2)	0–16	33.2	0.0002
		5	42 (42.4)	8.4 (8.5)	1–20	33.9	0.0002
		10	53 (53.5)	13.1 (13.2)	3–26	40.3	0.0002
	Any scope	1	24 (24.2)	3.2 (3.3)	0–10	21.0	0.0002
		3	52 (52.5)	9.2 (9.3)	1–20	43.2	0.0002
		5	64 (64.7)	11.6 (11.8)	2–24	52.9	0.0002
		10	74 (74.8)	16.9 (17.0)	5–31	57.7	0.0002

Table 17: Annotation-matched feature null for the pooled expert-prior overlap. Counts pool the 99 directly scoreable reviewer statements over the 20 questions and use the opposite-direction veto of Table 18. In each of 5,000 draws per model, every top-k feature is replaced by a random distinct feature from the same final-annotation-count stratum (0, 1, 2, 3–4, 5–8, 9–16, >16). The replacement keeps its own annotations and is scored by the same rule. Top-10 positions carry on average 16.5 (GPCT) and 12.5 (MAESTRO) final annotations, against 4.0 and 1.8 over all SAE features. Null 95% is the central range of null counts. Excess is observed minus null mean in percentage points. p is the one-sided empirical $(1 + \#\{\text{null} \geq \text{observed}\})/(1 + \text{draws})$, whose floor is set by the number of draws. Repeated statements across reviewers are not independent.

Table 18: Expert-prior overlap (%) at top@k by question family and expert, with any-scope matching, including family subtotals and the overall row.

		GPCT, any scope (%)					MAESTRO, any scope (%)				
Expert	n	top@1	top@2	top@3	top@4	top@5	top@1	top@2	top@3	top@4	top@5
Diagnosis (11 comparisons)											
E1	3	66.7	66.7	100.0	100.0	100.0	0.0	0.0	33.3	33.3	33.3
E2	6	33.3	50.0	50.0	50.0	66.7	16.7	16.7	16.7	33.3	50.0
E3	11	36.4	54.5	63.6	72.7	81.8	9.1	18.2	45.5	45.5	54.5
E4	7	57.1	71.4	85.7	85.7	100.0	28.6	28.6	42.9	42.9	57.1
E5	9	44.4	55.6	66.7	66.7	66.7	11.1	11.1	22.2	33.3	33.3
Subtotal	**36**	44.4	58.3	69.4	72.2	80.6	13.9	16.7	33.3	38.9	47.2
Feature null		8.2	13.8	20.5	24.8	27.3	3.3	5.6	8.3	10.1	11.5
COVID state (3 comparisons)											
E1	2	0.0	0.0	0.0	0.0	50.0	0.0	0.0	50.0	50.0	50.0
E2	4	25.0	25.0	50.0	50.0	50.0	0.0	25.0	25.0	50.0	75.0
E3	8	25.0	37.5	62.5	62.5	87.5	0.0	0.0	25.0	25.0	37.5
E4	4	25.0	25.0	50.0	50.0	75.0	0.0	0.0	25.0	25.0	50.0
E5	4	25.0	25.0	50.0	50.0	50.0	0.0	25.0	25.0	50.0	75.0
Subtotal	**22**	22.7	27.3	50.0	50.0	68.2	0.0	9.1	27.3	36.4	54.5
Feature null		4.1	7.5	10.1	11.9	13.5	0.2	1.0	2.2	2.7	3.7
Treatment (6 comparisons)											
E1	6	33.3	100.0	100.0	100.0	100.0	50.0	83.3	83.3	83.3	83.3
E2	8	50.0	100.0	100.0	100.0	100.0	62.5	87.5	87.5	87.5	87.5
E3	6	33.3	100.0	100.0	100.0	100.0	33.3	83.3	100.0	100.0	100.0
E4	6	16.7	100.0	100.0	100.0	100.0	50.0	83.3	100.0	100.0	100.0
E5	6	33.3	100.0	100.0	100.0	100.0	66.7	100.0	100.0	100.0	100.0
E6	9	44.4	66.7	66.7	66.7	55.6	22.2	44.4	44.4	55.6	55.6
Subtotal	**41**	36.6	92.7	92.7	92.7	90.2	46.3	78.0	82.9	85.4	85.4
Feature null		13.2	20.3	24.4	28.2	31.2	4.9	11.2	14.0	15.6	16.3
Overall	**99**	36.4	65.7	74.7	75.8	81.8	24.2	40.4	52.5	57.6	64.6
Feature null		9.4	15.1	19.8	23.3	25.9	3.3	6.9	9.3	10.7	11.8

E.8 DISCOVERY DIAGNOSIS MAP FROM RELATIVE FIRING

For each of 11 diseases, use saved discovery sample-level SAE Top-k masks to compute $P(\text{firing} \mid \text{disease}) - P(\text{firing} \mid \text{Healthy})$ for each feature. The common Healthy reference contains 178 ordinary Healthy and 141 BatchControl samples. Among features with eligible final annotations, select the four largest positive differences and the most negative difference, with feature-index tie breaking. Heatmaps in Supplementary Section S3 show every selected feature against every disease: 36 unique GPCT and 32 unique MAESTRO features. Each model has two pages with identical disease rows; color encodes the signed difference, black outlines mark the four case-enriched features selected in that row, and teal outlines mark the control-enriched feature. The other colored cells show descriptive ON-frequency differences for the same selected-feature union.

Case counts range from 16 (multiple myeloma) to 258 (MS), against the same 319 Healthy samples. These discovery-only maps may reflect cohort structure and establish neither disease specificity nor held-out replication.

E.9 EXPLORATORY DECODER STEERING AND VISUALIZATION

For each of the 7,440 confident or partially confident MAESTRO annotations, we calibrated a feature-specific scale as the empirical 95th percentile of positive post-Top-k activation over all 1,380 discovery samples. We added the corresponding SAE decoder direction to the frozen sample embedding at signed multipliers $-4, -2, -1, +1, +2, +4$, decoded each intervened embedding through the MAESTRO decoder, and applied the fixed population gates to its 40,000 output cells. The outcome is the mean within-sample change in target A as a percentage of its enclosing scope S, relative to decoding the original embedding. A match requires this change to have the annotation's expected sign at a positive dose, and the opposite sign at a negative dose. This is a directional criterion; the annotations do not predict effect sizes.

All 7,440 claims were evaluated at all six signed doses on the 314 held-out samples. Of these, 6,935 (93.2%) matched at all six, 280 (3.8%) at one to five, and 225 (3.0%) at none (Figure 3). Identical deterministic measurements are reused for 375/1,167 features; recomputation agrees on 1,213 shared claims to $< 10^{-14}$ percentage points. These reused held-out samples test model compatibility, not biological validity. Steering can change other SAE activations and gate membership, and large multipliers extrapolate beyond the discovery scale.

The four examples use 24 held-out samples selected by a seeded hash before visualization. The first two T-cell claims were fixed before selection, the B-cell claim is a positive discovery annotation, and the fourth uses feature 2004's separate annotation for CD8 within CD3 T cells. One UMAP is fitted jointly to the baseline and eight signed extreme-dose conditions using the decoded 30-marker vectors directly, with 250 cells drawn uniformly per sample and condition. Gate colors are assigned before UMAP; the percentages use all 40,000 decoded cells in each sample. In the 24-sample subset, B-cell frequency rises from 1.53% at baseline to 12.51% at $+4$. For the CD8/CD3T claim, the within-CD3T CD8 fraction falls from 34.35% to 27.09%, even as CD8 frequency among all cells rises from 7.54% to 8.94%: the enclosing CD3T population expands more strongly. All 216 decoded sample–condition outputs pass the prespecified OOD guard. UMAP area and point density are qualitative, and decoder output positions are not tracked biological cells. The complete exploratory cases also include a feature 3664 LateNK/Mononuclear claim whose observed direction disagrees with its annotation; the selected panels do not replace the full denominator.

F DECODER-FREE STEERING READOUT

This analysis reads composition changes from neighbouring measured samples after moving an embedding. It uses a separate set of 7,267 MAESTRO annotations selected under a different rule (93.9% decoder agreement), so it is not matched to the 7,440 annotations of Section 5.5. Because the held-out samples are also used in other evaluations, this analysis is retrospective.

The 314 held-out samples use 40,000-cell baseline embeddings standardized as SAE inputs; gate compositions use all source cells. Steps follow each SAE decoder direction at $\pm 1, \pm 2, \pm 4$ times its 95th-percentile positive discovery activation, as in decoder steering.

The readout averages gate compositions over ten Euclidean-nearest held-out neighbours, excluding the same patient; each unlinked sample forms its own group. We average the after-minus-before change in A/S over samples with nonempty neighbour scope counts at both endpoints. Its expected sign is the annotation direction times the step sign and decoder–encoder gain (the inner product of the feature's decoder direction and encoder column), positive for all 793 annotated features. Directions, steps and annotations were fixed on discovery data.

Each feature is compared with 20 Gaussian combinations of stable decoder directions, matched to its norm and steps (Appendix C). Overall, 6,569/7,267 annotations (90.4%) match at all six doses, against 40.4% of annotation–random-direction pairs (Table 19). Of 793 features, 704 match at all doses on most annotations. Agreement is 86.9% for features firing in one to five held-out samples and 94.2% for those firing in at least 24.

Table 19: Decoder-free agreement for the separate 7,267-annotation set on 314 held-out samples. A step size counts as matched when both its positive and its negative step have the expected sign.

Step size	1	2	4
Own feature	91.6%	93.1%	93.5%
Random directions	44.1%	45.6%	46.2%
Samples with a changed neighbour set	46%	69%	88%

The sign of the held-out Spearman correlation between pre-activation and A/S agrees with 93.1% of annotations, without a significance test; 6,273 annotations agree with both this sign and the readout. Thus the readout largely reflects local covariation. Discovery agreement is 96.0%, using 527 samples from 378 patients outside the annotation measurement set as neighbours and 853 samples from the other patients as queries. Because Stage 3 selects for discovery covariation, this is not independent validation.

We also compare decoded and retrieved changes in all 53 gate shares for 120 directions: 40 annotated features, 50 unannotated features and 30 random directions. Decoding uses step $+2$ on up to 24 discovery samples; retrieval uses held-out samples. Per-unit-step decoded patterns rank their corresponding retrieved patterns by cosine similarity at medians 2, 2 and 7.5, respectively (chance: 60.5). Removing axes fitted to the random directions changes these medians to 1, 3 and 4.5. One-sided Mann–Whitney comparisons with random directions give $p = .124$ before and .012 after removal for annotated features, and .010 and .148 for unannotated features. Agreement also occurs for random decoder-span directions and does not itself establish semantic specificity.

G Exploratory GPCT discovery metadata steering atlas

This exploratory analysis steers case-enriched GPCT features in control embeddings for the 11 discovery comparisons of Appendix E.8, using a separate selection rule and the same 319 Healthy samples.

Selection and intervention. We use the Top-k feature firing indicators for 1,380 GPCT discovery samples. Within each comparison, we select every feature that fires in at least half of case samples and whose firing frequency exceeds the control frequency by at least 0.10. This discovery-only rule selects 2–37 features per disease (median seven; 130 disease–feature memberships across the 11 comparisons). It is a firing-frequency threshold, not a statistical-significance or final annotation filter. All selected features are steered simultaneously in *control* embeddings. Joint SAE-decoder coefficients are calibrated against the selected decoder–encoder interaction matrix so that every selected pre-activation has a positive target increment. The $+1$ target uses each feature's 95th percentile positive baseline pre-activation and decoder–encoder self-gain; $+2$ and $+4$ scale the same vector. The intervention changes a frozen GPCT sample embedding. It does not resample or re-embed the original cells, or decode a new cytometry specimen.

Reading the atlas. One UMAP is fitted to standardized discovery embeddings after 50-component PCA, and intervened controls are projected into that map. Figure 12 shows baseline and $+4$ for all 11 comparisons, including unfavorable movements; Supplementary Section S5 also shows $+1$ and

+2 with fixed axes, all control endpoints and 14 paired arrows per dose. These discovery-only diagnostics provide no independent validation. UMAP proximity establishes neither reduced 50-PC distance nor disease conversion.

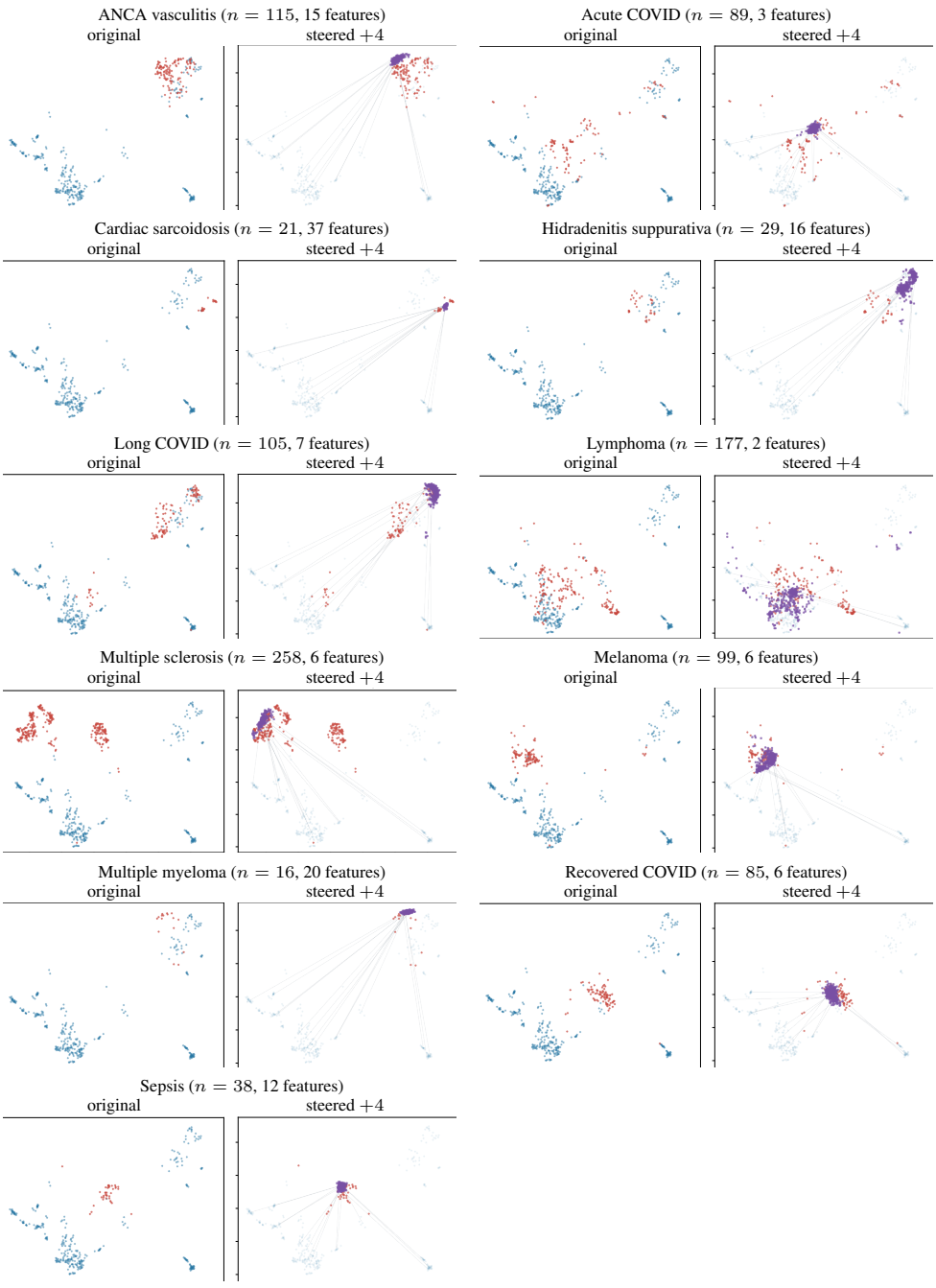

Figure 12: GPCT discovery steering, original versus +4 dose, for the 11 diagnosis versus Healthy comparisons. Each pair shows the fixed UMAP of the 1,380 discovery embeddings. Left: original case (red) and control (blue). Right: after adding the +4 joint decoder vector of the case-shared ON features to every control embedding, with steered controls in purple, original controls faint, and original cases red. Axis limits are fixed within each comparison. The full four-dose pages with axes and paired arrows are in Supplementary Section S5. Point proximity in two dimensions does not establish similarity of the full embedding or disease conversion.

H SHARED-FEATURE PROGRAMS: DETAILS

Case–control slope differences. For every feature–target–scope triple, Eq. 2 is fitted separately on the discovery samples of each group with the same sample intercepts and CR2 standard errors; groups are compared with a Welch test, and the family-wise error rate is controlled by the minimum-p distribution over 1,000 permutations of the group labels. A significant difference is kept when the group that drives it passes the annotation rule on its own: case-driven when the case group is consistent and $\operatorname{sign}\Delta\beta = \operatorname{sign}\beta_{\text{case}}$, control-driven when the control group is consistent and $\operatorname{sign}\Delta\beta = -\operatorname{sign}\beta_{\text{control}}$.

Graph and programs. Nodes are conditions (A, S) with a kept difference. Two conditions are linked by the features they share, with weight $\sum_F \operatorname{sign}(\tilde{d}_{aF}\tilde{d}_{bF})\sqrt{|\tilde{d}_{aF}\tilde{d}_{bF}|}$, where $\tilde{d}$ is $\Delta\beta$ divided by the rank-one scale of all discovery slopes of that condition and feature. Communities maximize signed modularity (Gómez et al., 2009). The program of a comparison is the community with the largest case-driven cohesion; rules and programs are frozen before any held-out sample is read.

Steering. The program direction is $u = \operatorname{normalize}(\mathrm{sd} \odot \sum_F q_{95,F} W_{\text{dec}}[F])$, and each held-out embedding z becomes $z + \alpha\, d\, u$, where d is the distance between the discovery case and control centroids; unlike the single-feature doses of Section 5.5, α is in units of this distance. The step is $\alpha = 2$: in a discovery-only dose check, decoded member shares change monotonically up to $\alpha = 2$ and the out-of-distribution guard fails for many files at $\alpha = 4$. The relative change of a member is the group mean change of A/S divided by the group mean unsteered A/S; for MAESTRO A/S is gated on decoded cells, for GPCT it is the decoder-free neighbour readout of Appendix F. The comparison value is the mean over members of the predicted sign times the case minus control relative change. Relative change was added after the absolute-change analysis showed that changes scale with baseline abundance. Table 20 lists every comparison.

Table 20: Case shared-feature programs at $\alpha = 2$. Files: held-out (or external) case/control files. F/m: program features and members with a predicted direction. Abs.: mean over members of the predicted sign times the mean change of A/S in percentage points; rel.: the same for the change divided by the group mean unsteered A/S, in %. Case minus control of the relative change is Figure 4D. MAESTRO reads decoded cells, GPCT the decoder-free neighbour readout; external cohorts use the acute-COVID program. – : no program.

		MAESTRO (decoder)					GPCT (decoder-free)				
Comparison	Files	F/m	abs. case	abs. ctrl	rel. case	rel. ctrl	F/m	abs. case	abs. ctrl	rel. case	rel. ctrl
ANCA Vasculitis	28/70	2/5	1.62	0.93	42.4	23.2	264/48	0.28	0.12	1.3	1.7
Acute COVID	21/70	1/2	2.58	2.76	130.7	85.7	1/2	4.36	2.44	33.2	11.1
Cardiac Sarcoidosis	5/70	–	–	–	–	–	14/5	−0.07	0.31	−13.9	1.7
Hidradenitis Suppurativa	10/70	–	–	–	–	–	3/6	0.18	0.04	0.9	−1.6
Long COVID	24/70	–	–	–	–	–	101/29	−0.12	−0.19	3.4	2.5
Lymphoma	42/70	3/5	4.32	2.60	78.9	73.6	386/48	3.22	2.48	32.4	22.6
MS	58/70	86/11	12.01	7.67	188.5	274.4	727/45	0.47	0.12	2.7	−2.0
Melanoma	22/70	1/2	2.41	1.42	111.0	64.5	191/23	0.30	0.06	29.5	26.8
Multiple Myeloma	4/70	–	–	–	–	–	–	–	–	–	–
Recovered COVID	18/70	5/7	1.51	1.44	21.9	16.0	193/32	−0.13	0.06	−1.2	−0.2
Sepsis	8/70	–	–	–	–	–	172/13	1.64	0.88	65.9	49.3
Long vs Recovered COVID	24/15	–	–	–	–	–	65/24	−0.48	−0.39	−9.5	−8.4
Acute vs Recovered COVID	21/15	–	–	–	–	–	3/5	−0.06	−1.33	−1.7	−10.2
Acute vs Long COVID	21/24	–	–	–	–	–	3/4	−0.32	−0.61	−5.6	−5.3
MS: anti-CD20 vs S1P	35/19	6/5	2.27	6.82	80.3	70.3	399/52	1.12	1.27	7.3	0.5
MS: anti-CD20 vs S1P/GA/Interferon	35/22	2/2	0.09	−0.02	9.8	−3.0	78/22	0.41	0.21	0.5	0.5
MS: S1P vs anti-CD20/GA/Interferon	19/38	1/2	1.13	−1.53	88.9	−20.6	898/27	−0.01	−0.24	12.8	4.5
Lymphoma: explicit CAR-T vs Untreated	27/5	1/2	4.54	−0.64	26.5	−4.2	–	–	–	–	–
Lymphoma: CAR-T vs Untreated	29/5	1/2	4.27	−0.62	25.8	−4.0	–	–	–	–	–
Lymphoma: CAR-T vs Untreated/BTKi	29/5	1/2	4.86	0.51	29.4	3.3	1/2	0.83	0.09	10.3	2.3
Hospitalized (external)	77/15	1/2	2.65	4.49	259.4	192.2	1/2	4.08	2.36	55.1	13.3
Intensive care (external)	116/8	1/2	3.16	3.79	128.6	94.3	1/2	4.55	3.71	53.8	20.9

I EXTENDED LIMITATIONS

Rare targets. Across 53 targets, annotations per target–scope pair correlate with median cell abundance at Spearman $\rho = .66$ for GPCT and .29 for MAESTRO (Figure 13). CD8TEM2 and ILC, the rarest targets (.003% and .004% of cells), receive 18 and 7 GPCT annotations but 1 and 0 MAESTRO annotations. For ILC, no MAESTRO response reaches the top 10% of sample-direction agreement; GPCT retains 7 concordant responses from 21 consistent ones. Exchanges affecting few cells can produce noisy per-sample slopes, so absent annotations do not establish model insensitivity. Plasmablasts are an exception: GPCT yields 109 annotations at .008% abundance.

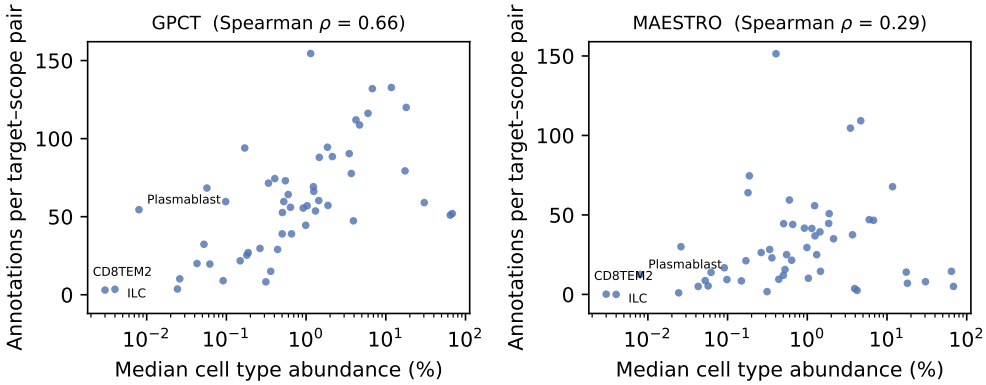

Figure 13: Target rarity and annotation count across 53 populations. Abundance is the median share of all cells over 1,380 discovery samples (log scale); annotation counts are normalized by the number of target–scope pairs.

Gates and supported inputs. All cohorts use discovery-defined univariate marker thresholds. Intensity shifts can change gate membership across studies, and these gates cannot resolve populations requiring multivariate or density-based definitions. Gate overlap and exact-threshold gaps also affect scope contrasts. The OOD guard and discovery dose ranges restrict evaluation to supported mixtures; they do not cover every biologically plausible input. Composition resampling is distinct from editing molecular profiles within cells (Theodoris et al., 2023; Cui et al., 2024).

Estimation and inference. Pooled slopes can hide nonlinear and sample-specific responses; pooled replication therefore does not imply uniform responses in every sample. Sample-level CR2/CR1 and permutation inference do not account for arbitrary donor dependence or study–disease confounding. Patient-disjoint splits prevent cross-partition overlap but do not establish within-partition independence; small firing groups further limit uncertainty calibration.

Selection ranks depend on the measured conditions, hierarchy and dose design. Cross-seed stability measures decoder-direction recurrence, not biological meaning; neither it nor global selection ranks provide a set-wide error guarantee beyond the stated per-exchange correction. Displayed features were selected for stability and annotation count, not randomly sampled.

Retrospective changes to held-out criteria are identified in Appendix E.1. Reanalysis does not make these endpoints prospective. Unsupported claims need not be false, and the post-hoc expected support calculation cannot attribute every failure to low power (Appendix E.4). Interval overlap is not effect equivalence. Perturbation replication, observational association, expert agreement and classification utility assess different properties; none establishes clinical usefulness or biological causality.