BALANCING INVARIANCE AND INFORMATION PRESERVATION IN SINGLE-CELL JEPA

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-supervised learning for single-cell genomics must balance invariance to nuisance variation against preservation of biological information. Existing single-cell Joint-Embedding Predictive Architecture (JEPA) adaptations change backbones, anti-collapse mechanisms, and auxiliary objectives at the same time, so it is unclear which choice actually drives transfer. We therefore fix the encoder, view construction, training budget, and evaluation protocol, and vary only two axes: the anti-collapse regularizer and the observation-space preservation target. Within this setting we introduce PCG-BCE, a per-cell-balanced and gene-entropy-weighted binary cross-entropy loss that supervises gene detection rather than expression magnitude, and DET-JEPA (detection-aware JEPA), which couples PCG-BCE with projector-space latent prediction. Three findings emerge. First, preventing collapse is not sufficient: variants trained with latent prediction alone, including one built on a provably collapse-free regularizer, fall below both a plain masked autoencoder and PCA on zero-shot cell-type annotation and on scIB biological conservation. Second, which observation-space target is preserved matters more than whether one is used: detection prediction outperforms magnitude reconstruction on every out-of-distribution cohort, consistent with a sampling analysis showing that detection labels are insensitive to positive-count magnitude. Third, PCG-BCE alone is already a strong representation learner; adding latent prediction improves 11 of 12 downstream metrics, but each gain is small and not individually significant over five seeds. Across zero-shot cell-type annotation, multi-batch integration, and protein-abundance prediction, DET-JEPA is the best matched JEPA variant, with VISReg as the preferred regularizer. Our study localizes where the gains in single-cell JEPAs come from and gives concrete guidance for designing preservation objectives.

1 INTRODUCTION

Single-cell RNA sequencing has matured from individual experiments to atlas-scale compendia spanning millions of cells across tissues, donors, and protocols. The resulting abundance of unlabeled data has made self-supervised learning (SSL) the engine of an ambitious agenda: learning transferable representations, and ultimately predictive models, of cellular state—a direction articulated by the "virtual cell" program (Bunne et al., 2024) and pursued by a growing family of single-cell foundation models (Yang et al., 2022; Theodoris et al., 2023; Cui et al., 2024; Hao et al., 2024; Fischer et al., 2024; Rosen et al., 2026).

What makes objective design hard in this domain is the measurement process itself. Observed counts reflect biological abundance convolved with library size, capture efficiency, and dropout, so an objective that regresses expression magnitudes must spend capacity on variation that is specific to how a cell was measured rather than to what the cell is. This argues against treating "reconstruct the input" as a neutral default. It does not, however, argue for abandoning observation-space supervision: latent-space objectives carry their own failure mode. Contrastive learning aligns augmented views of a cell (Chen et al., 2020; Han et al., 2022) but depends on heuristic augmentations whose biological validity is unclear (Richter et al., 2025), and joint-embedding predictive architectures (JEPAs) (Assran et al., 2023; Balestriero & LeCun, 2025) must be actively prevented from collapsing. Recent regularizers address collapse directly and with increasing rigor—VICReg constrains second-order moments (Bardes et al., 2022), SIGReg tests against an isotropic Gaussian and is

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.

proven collapse-free (Balestriero & LeCun, 2025), VISReg adds sketch-based shape matching (Wu et al., 2026)—which raises a question that the collapse literature does not answer: once collapse is no longer the binding constraint, is a latent objective alone enough?

Our answer, obtained under a controlled protocol, is no. Variants trained with latent prediction alone fall below both a plain masked autoencoder and even PCA on zero-shot cell-type transfer and on scIB biological conservation, even though they are trained with regularizers designed (and in one case proven) to avoid degenerate solutions. This reframes the design question. The relevant axis is not input space versus latent space, but which observation-space target to preserve, and how to weight it. We therefore keep the axis of "how much detail to retain" explicit: rather than reconstructing magnitudes, we supervise the binary detection pattern, which retains gene presence while discarding within-positive magnitude variation. We show analytically that detection labels are invariant to positive-count magnitude, that their flip probability under molecular downsampling decays as expected counts grow, and that the corresponding logit gradient is bounded independently of count magnitude—three properties that magnitude regression does not have.

Recent single-cell JEPAs (Litman et al., 2025; ElSheikh et al., 2026; Zhai et al., 2026) all pair latent prediction with some form of preservation, but each varies architecture, regularizer, and auxiliary loss simultaneously, so the contribution of any single choice is not identifiable. We hold the encoder, masking, optimizer, budget, and downstream protocol fixed and vary only the regularizer and the preservation target. Within this framework we introduce DET-JEPA (detection-aware JEPA), which combines projector-space latent prediction under VISReg with PCG-BCE, a detection loss that applies per-cell class balancing followed by online gene-entropy weighting to match the sparsity and the detection heterogeneity of single-cell data. Our results localize the gains: most of the improvement over latent-only JEPA comes from detection-aware preservation itself (e.g., HLCA macro-F1 $0.601 \rightarrow 0.748$; scIB Bio $0.562 \rightarrow 0.630$ at essentially unchanged batch mixing), while adding latent prediction on top of PCG-BCE yields consistent but small further gains (better on 11 of 12 metrics, none individually significant).

Contributions.

- **A controlled design space for single-cell JEPA.** We separate view construction, anti-collapse regularization, and observation-space preservation into three axes, and vary the latter two under a shared backbone, budget, and evaluation protocol while holding view construction fixed at a setting established in prior work (Richter et al., 2025; Ovcharenko et al., 2025).
- **Detection-aware preservation with PCG-BCE.** We motivate binary gene detection as a noise-robust preservation target through an explicit sampling analysis, and introduce a two-stage weighting scheme (per-cell class balancing and online gene-entropy weighting) whose stages we ablate separately.
- **Attribution of the gains, including where they do not appear.** Across three task families we show that collapse prevention alone is insufficient, that the preservation target matters more than its presence, that VISReg dominates SIGReg and VICReg by 0.13–0.15 mean OOD macro-F1, and that the latent-prediction branch contributes a small, directionally consistent, but statistically unresolved increment.

2 RELATED WORK

Joint-embedding predictive architectures and collapse prevention. Joint-embedding architectures learn from paired views without reconstructing inputs, but must prevent representational collapse. Contrastive methods use negative pairs (Chen et al., 2020; He et al., 2020); non-contrastive designs rely on engineered asymmetries (predictor branches, stop-gradient, exponential moving average (EMA)) (Grill et al., 2020; Chen & He, 2021; Caron et al., 2021). Joint-Embedding Predictive Architectures (JEPAs) recast this as latent-space prediction (Assran et al., 2023). The family has evolved: MC-JEPA disentangles motion and content (Bardes et al., 2023), V-JEPA scales feature prediction to video (Bardes et al., 2024), and V-JEPA 2 enables understanding, prediction, and planning at world-model scale (Assran et al., 2025). Early training relied on heuristic asymmetries (stop-gradient, EMA) to avoid collapse; recent regularized objectives simplify this: VICReg uses variance and covariance terms (Bardes et al., 2022), LeJEPA proves variance-matching regularization is collapse-free (Balestriero & LeCun, 2025), and VISReg extends this with sketch-based

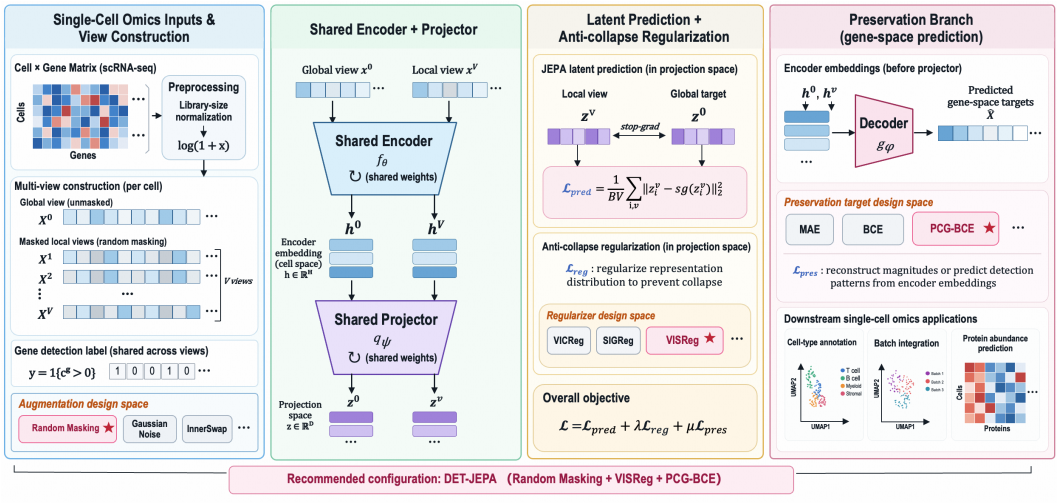


Figure 1: Framework overview and **design space** of single-cell JEPA and DET-JEPA. JEPA prediction and anti-collapse regularization operate after the shared projector, while the preservation decoder receives encoder embeddings before projection and reconstructs magnitudes or predicts detection patterns. The controlled variables are the regularizer and the preservation target. Highlighted choices indicate our recommended configuration.

regularization (Wu et al., 2026). These advances make JEPA simpler and more stable, shifting the question to which design choices matter for specific modalities.

Self-supervised learning for single-cell genomics. Single-cell SSL includes reconstruction and denoising (Eraslan et al., 2019; Lopez et al., 2018), contrastive objectives (Han et al., 2022), and large-scale masked modeling (Yang et al., 2022; Theodoris et al., 2023; Cui et al., 2024; Hao et al., 2024). These methods use different targets: for example, Geneformer predicts masked gene identities, whereas UCE predicts whether queried genes are expressed (Rosen et al., 2026). Binary detection provides a target focused on gene presence rather than expression magnitude. Recent single-cell JEPAs also combine representation learning with different preservation and collapse-prevention mechanisms: scJEPA pairs latent prediction with expression-space denoising and SIGReg (Zhai et al., 2026); Cell-JEPA combines expression reconstruction with cell-level latent prediction using an EMA teacher (ElSheikh et al., 2026); and GeneJEPA predicts latents of held-out gene sets with VICReg regularization (Litman et al., 2025). These studies motivate preserving input information alongside latent prediction, but their differing architectures and objectives leave the choice of preservation target and its interaction with regularization unresolved under a common protocol. Following the emphasis on controlled evaluation in Richter et al. (2025), we investigate those choices within a shared MLP framework. Their single-cell masked autoencoder (SCMAE), benchmarked against zero-shot Geneformer under the same protocol, establishes a competitive reference for comparing SSL pretraining strategies without architectural confounds.

3 METHOD

JEPAs aim to learn representations by predicting the representation of the original global view from partially observed local views, rather than by directly reconstructing the input. Figure 1 summarizes our design space and recommended configuration for single-cell JEPAs. The general training objective comprises two branches: a **latent prediction branch**, which combines a *predictive loss* that matches local and global representations with an *anti-collapse regularization loss* that keeps the projected representation informative, and an **information-preservation branch**, whose decoder predicts *gene-space targets* from encoder embeddings.

3.1 LATENT PREDICTION BRANCH

Local Views Construction. Let $\boldsymbol{x} \in \mathbb{R}^G$ denote a gene-expression vector obtained from raw counts c_g via library-size normalization followed by a $\log(1+x)$ transformation. For view augmentation, we adopt random masking, which has been shown to be competitive across multiple downstream tasks (Richter et al., 2025; Ovcharenko et al., 2025). Specifically, a random mask $\boldsymbol{m}^{(v)} \in \{0,1\}^G$ is drawn with $m_g^{(v)} \sim \text{Bernoulli}(\rho)$ independently across genes, and the masked entries are set to zero, yielding the augmented view $\boldsymbol{x}^{(v)} = \boldsymbol{x} \odot (\mathbf{1} - \boldsymbol{m}^{(v)})$. We use $\rho = 0.5$ throughout. The detection label $y_g = \mathbf{1}\{c_g > 0\}$ is computed from the raw counts prior to augmentation and thus remains identical across views.

Predictive Loss. Let $f_\theta : \mathbb{R}^G \to \mathbb{R}^H$ be the encoder and $q_\psi : \mathbb{R}^H \to \mathbb{R}^D$ a shared projector. For each cell i, we construct one unmasked global view $\boldsymbol{x}_i^{(0)}$ and V masked local views $\{\boldsymbol{x}_i^{(v)}\}_{v=1}^V$. We distinguish encoder embeddings $\boldsymbol{h}_{iv} = f_\theta(\boldsymbol{x}_i^{(v)})$ from projected representations $\boldsymbol{z}_{iv} = q_\psi(\boldsymbol{h}_{iv})$, for $v = 0, \ldots, V$. Let $\boldsymbol{Z}_v \in \mathbb{R}^{B \times D}$ collect the projected representations of view v in a batch. The predictive loss $\mathscr{L}_{\text{pred}}$ matches each masked view to the stop-gradient global target in the same projection space, averaged over a batch of B cells:

$$\mathscr{L}_{\text{pred}} = \frac{1}{BV} \sum_{i=1}^{B} \sum_{v=1}^{V} \|\boldsymbol{z}_{iv} - \text{sg}(\boldsymbol{z}_{i0})\|_2^2. \tag{1}$$

Regularization Loss. Inspired by recent regularized joint-embedding approaches, particularly LeJEPA (Balestriero & LeCun, 2025), we regularize the representation distribution to prevent collapse without an EMA teacher or an additional predictor. We apply the anti-collapse regularizer to every view and average over the global and local views:

$$\mathscr{L}_{\text{reg}} = \frac{1}{V+1} \left[\mathscr{R}(\text{sg}(\boldsymbol{Z}_0)) + \sum_{v=1}^{V} \mathscr{R}(\boldsymbol{Z}_v) \right]. \tag{2}$$

We compare three choices of $\mathscr{R}$ as controlled alternatives. VICReg uses hinged per-dimension variance together with an off-diagonal covariance penalty (Bardes et al., 2022); SIGReg sketches an Epps–Pulley distributional test toward an isotropic Gaussian (Balestriero & LeCun, 2025); VISReg uses a scale–shape–center decomposition with sliced-Wasserstein shape matching (Wu et al., 2026). Full definitions are deferred to Appendix A.1, and concrete implementations are given in Appendix A.2. We adopt VISReg as our default regularizer for collapse prevention.

For a batch of projected representations $\boldsymbol{Z} \in \mathbb{R}^{B \times D}$, let $\boldsymbol{\mu} = \frac{1}{B} \boldsymbol{Z}^\top \mathbf{1}$ be the batch mean, $\hat{\boldsymbol{Z}} = \boldsymbol{Z} - \mathbf{1}\boldsymbol{\mu}^\top$ the centered projections, and σ_j the standard deviation of dimension j. VISReg combines

$$\mathscr{R}_{\text{VISReg}} = \lambda_{\text{scale}} \mathscr{L}_{\text{scale}} + \lambda_{\text{shape}} \mathscr{L}_{\text{shape}} + \lambda_{\text{center}} \mathscr{L}_{\text{center}}, \quad \lambda_{\text{scale}} + \lambda_{\text{shape}} + \lambda_{\text{center}} = 3, \tag{3}$$

where the center and scale penalties are given by

$$\mathscr{L}_{\text{center}} = D^{-1} \|\boldsymbol{\mu}\|_2^2, \qquad \mathscr{L}_{\text{scale}} = \frac{1}{D} \sum_{j=1}^{D} (1 - \sigma_j)^2. \tag{4}$$

The scale penalty targets unit standard deviation. The shape term operates on normalized projections with a stop-gradient denominator, $\tilde{\boldsymbol{Z}} = \hat{\boldsymbol{Z}} / (\text{sg}(\sigma) + \varepsilon)$. For K random unit directions $\{\boldsymbol{a}_k\}_{k=1}^K$,

$$\mathscr{L}_{\text{shape}} = \frac{1}{BK} \sum_{k=1}^{K} \left\| \text{sort}(\tilde{\boldsymbol{Z}} \boldsymbol{a}_k) - \boldsymbol{q}_B \right\|_2^2, \tag{5}$$

where $\boldsymbol{q}_B$ contains the B standard-normal quantiles $\Phi^{-1}(1/(B+1)), \ldots, \Phi^{-1}(B/(B+1))$. The stop-gradient blocks derivatives through the normalization denominator, allowing shape matching to operate on standardized projections alongside explicit scale calibration.

The base objective of the latent prediction branch combines the predictive and regularization losses,

$$\mathscr{L}_{\text{JEPA}} = \mathscr{L}_{\text{pred}} + \lambda_{\mathscr{R}} \mathscr{L}_{\text{reg}}. \tag{6}$$

3.2 INFORMATION-PRESERVATION BRANCH

Why an auxiliary branch? In computer vision, the Context Autoencoder provides a representative dual-space design that combines latent-space representation learning with input-space reconstruction (Chen et al., 2024). As visual augmentation and view construction pipelines have matured, more recent methods have shown that latent-space objectives alone can outperform such dual-space learning (Assran et al., 2023; Balestriero & LeCun, 2025).

Single-cell transcriptomics, however, remains in a different regime: current augmentation operators are limited proxies for biologically valid views (Richter et al., 2025). We therefore posit that, *at the current stage* of augmentation design, aligning single-cell views only in latent space does not retain sufficient biological signal, a premise that we evaluate in Section 4. This motivates an explicit information-preservation branch alongside the latent objective.

Detection-aware objectives. Rather than reconstructing noisy expression magnitudes with a masked autoencoder (MAE), our preservation decoder g_ϕ predicts the original binary detection pattern $y_{ig} = \mathbf{1}\{c_{ig} > 0\}$ from each clean or masked pre-projector embedding $\boldsymbol{h}_{iv}$ (Rosen et al., 2026). This target retains gene presence while removing variation among positive count magnitudes. Appendix D motivates this choice and analyzes detection probabilities, positive-magnitude invariance, sampling sensitivity, and loss gradients.

Starting from ordinary binary cross-entropy (BCE), we incorporate two priors of single-cell data: (i) gene detection is sparse, and the number of detected genes varies across cells; (ii) genes differ in their information content. For the first prior, let $P_i = \{g : y_{ig} = 1\}$ and $N_i = \{g : y_{ig} = 0\}$, and assign $a_{ig} = |N_i| / \max(|P_i|, 1)$ to detected genes and $a_{ig} = 1$ otherwise. For the second, we estimate each gene's detection frequency online. At batch t, let $\bar{p}_g^{(t)} = B^{-1} \sum_i y_{ig}^{(t)}$ and update

$$p_g^{(t)} = \begin{cases} \bar{p}_g^{(t)}, & t = 1, \\ \beta p_g^{(t-1)} + (1-\beta)\bar{p}_g^{(t)}, & t > 1, \end{cases} \tag{7}$$

with $\beta = 0.99$. We compute the Bernoulli entropy $H_g = -p_g \log p_g - (1-p_g)\log(1-p_g)$, with $0 \log 0 := 0$, and a gene weight normalized across genes with a lower floor:

$$\eta_g = \max\left(\eta_{\min}, \frac{(H_g + \varepsilon)^\alpha}{G^{-1} \sum_{j=1}^{G} (H_j + \varepsilon)^\alpha}\right). \tag{8}$$

For each view, let $s_{ig}^{(v)} = [g_\phi(\boldsymbol{h}_{iv})]_g$ and $\hat{y}_{ig}^{(v)} = \mathrm{sigmoid}(s_{ig}^{(v)})$. Averaging the weighted binary cross-entropy over the G genes, the B cells, and the global and local views defines the preservation loss

$$\mathscr{L}_{\text{pres}} = -\frac{1}{(V+1)BG} \sum_{v=0}^{V} \sum_{i=1}^{B} \sum_{g=1}^{G} a_{ig} \eta_g \left[y_{ig} \log \hat{y}_{ig}^{(v)} + (1 - y_{ig}) \log(1 - \hat{y}_{ig}^{(v)}) \right]. \tag{9}$$

BCE sets $a_{ig} = \eta_g = 1$; PC-BCE applies only the per-cell factor a_{ig}; and PCG-BCE applies both a_{ig} and η_g. Thus, we progressively refine the preservation objective from BCE to PC-BCE and finally PCG-BCE. The entropy weight emphasizes genes with variable detection patterns, while the EMA provides a stable estimate of gene detection frequency across training batches.

3.3 TRAINING OBJECTIVE AND INFERENCE

For the selected preservation objective, the decoder maps every global and local pre-projector embedding to the original, unmasked target. The full objective is the weighted sum of the predictive, regularization, and preservation losses:

$$\mathscr{L} = \mathscr{L}_{\text{pred}} + \lambda_{\mathscr{R}} \mathscr{L}_{\text{reg}} + \lambda_{\text{pres}} \mathscr{L}_{\text{pres}}. \tag{10}$$

Pure JEPA sets $\lambda_{\text{pres}} = 0$, whereas DET-JEPA uses VISReg and PCG-BCE. At inference, we discard the projector and preservation decoder and use $\boldsymbol{h} = f_\theta(\boldsymbol{x})$ for downstream tasks. Coefficient settings are listed in Appendix A.2.

4 EXPERIMENTS

4.1 RESEARCH QUESTIONS

We organize the experiments around three questions that mirror the design space:

RQ1: Regularization: Which anti-collapse regularizer is most effective for single-cell JEPA under matched downstream evaluations?

RQ2: Objective trade-offs: How does input-space preservation change the balance between invariance and biological signal?

RQ3: Preservation targets and weighting: Does detection-aware preservation improve over magnitude reconstruction, and how does the PCG-BCE design affect downstream performance?

4.2 EXPERIMENTAL PROTOCOL

We follow the established single-cell SSL and scIB protocols (Zhai et al., 2026; Richter et al., 2025; Luecken et al., 2022). Models are pretrained on donor-disjoint scTab partitions (approximately 15.2M/3.5M/3.4M cells) after library-size normalization and $\log(1+x)$ transformation (Fischer et al., 2024). We evaluate (i) zero-shot 5-nearest-neighbor (5-NN) macro-F1 transfer on in-distribution (IID) and out-of-distribution (OOD) cohorts, (ii) lung integration with scIB Batch, Bio, and Total scores, and (iii) fine-tuned CITE-seq protein prediction from 2,000 highly variable genes (HVGs), reporting Pearson, Spearman, and mean squared error (MSE). Dataset-specific protocols are given in Appendix B.7. Classification and scIB use the shared 19,331-gene scTab vocabulary, with missing genes zero-padded after filtering. For classification, the pretrained encoder is frozen and embeds both the 3.5M-cell scTab validation reference pool and each test cohort; a 5-NN classifier is then applied without updating the encoder. The scIB evaluation uses 56,771 lung cells spanning 40 cell types and 9 batches, and reports batch removal separately from biological conservation before forming the weighted Total score. For multiomics, pretraining and downstream fine-tuning use the same 2,000 Scanpy-selected HVGs under the published random 80/10/10 split, with all 134 protein markers predicted from the RNA input.

All controlled variants share the encoder, masking, optimizer, training budget, and downstream protocol; only the regularizer and auxiliary preservation loss change. Each hyperparameter is selected by an independent grid search under the same search budget. SCMAE (standalone MAE) and standalone PCG-BCE serve as external references on all three benchmarks. A raw-gene PCA reference is included for cell-type classification and scIB, and a NoSSL reference—a network trained directly on the downstream supervised task without SSL pretraining—for multiomics. The SIGReg+MAE variant, which reproduces the setup of Zhai et al. (2026), is included as a controlled comparison. SCMAE is the competitive random masking baseline established by Richter et al. (2025) under the same evaluation protocol, so we do not introduce further external benchmarks. Results are means $\pm$ standard deviations over five independent pretraining seeds. Selected hyperparameters and implementation details are provided in Appendix A.2.

4.3 CONTROLLED EXPERIMENT RESULTS

Cell-type transfer under IID and OOD shifts. We evaluate zero-shot cell-type transfer under both IID and OOD shifts to test whether the auxiliary branch improves transfer and whether the gain survives distribution shift, comparing regularizers and preservation objectives.

The IID benchmarks measure transfer to donor-disjoint reference datasets drawn from the same broad data ecosystem as pretraining. Pure VISReg JEPA reaches 0.6013, 0.4242, and 0.2454 macro-F1 on HLCA, PBMC, and Tabula Sapiens, while adding PCG-BCE raises these scores to 0.7484, 0.5424, and 0.3046. Detection preservation also improves the SIGReg and VICReg variants, showing that the benefit is not specific to one anti-collapse objective, although the VISReg combination is strongest across all three IID datasets.

The held-out HiT, NN, and Great Apes cohorts impose stronger dataset, protocol, and species shifts while retaining the same frozen-encoder 5-NN protocol and shared gene vocabulary. Here DET-JEPA improves over pure VISReg JEPA by 0.2554, 0.0644, and 0.1367 macro-F1, respectively, and has the highest mean among the matched JEPA variants on every OOD cohort. The paired

Table 1: Cell-type transfer performance (Macro-F1). Left: IID settings; Right: OOD settings. Darker blue-gray shading indicates a better column-wise rank (best = darkest); shading applies to the evaluated JEPA variants; external baselines are unranked.

Method	IID			OOD		
	HLCA	PBMC	Tabula	HiT	NN	Great Apes
VISReg ONLY	0.6013 ± 0.0094	0.4242 ± 0.0185	0.2454 ± 0.0078	0.4971 ± 0.0747	0.1283 ± 0.0031	0.1163 ± 0.0289
VISReg + MAE	0.6080 ± 0.0095	0.4502 ± 0.0138	0.2562 ± 0.0093	0.5406 ± 0.0968	0.1280 ± 0.0060	0.1391 ± 0.0417
DET-JEPA (Ours)	**0.7484 ± 0.0075**	**0.5424 ± 0.0062**	**0.3046 ± 0.0099**	**0.7525 ± 0.0961**	**0.1927 ± 0.0237**	**0.2530 ± 0.0187**
SIGReg ONLY	0.6080 ± 0.0260	0.4202 ± 0.0271	0.2354 ± 0.0075	0.4841 ± 0.0180	0.0954 ± 0.0076	0.0609 ± 0.0099
SIGReg + MAE	0.6145 ± 0.0215	0.4318 ± 0.0195	0.2483 ± 0.0084	0.4925 ± 0.0412	0.1018 ± 0.0085	0.0824 ± 0.0186
SIGReg + PCG-BCE	0.6322 ± 0.0170	0.4389 ± 0.0268	0.2562 ± 0.0058	0.5110 ± 0.0849	0.1104 ± 0.0076	0.1164 ± 0.0467
VICReg ONLY	0.5333 ± 0.0370	0.3587 ± 0.0102	0.2161 ± 0.0121	0.3190 ± 0.0296	0.0944 ± 0.0071	0.0549 ± 0.0066
VICReg + MAE	0.5482 ± 0.0289	0.3812 ± 0.0163	0.2294 ± 0.0110	0.3541 ± 0.0385	0.1026 ± 0.0092	0.0768 ± 0.0140
VICReg + PCG-BCE	0.7060 ± 0.0218	0.5059 ± 0.0206	0.2825 ± 0.0115	0.5248 ± 0.0838	0.1427 ± 0.0236	0.1371 ± 0.0266
SCMAE	0.7370 ± 0.0084	0.4888 ± 0.0200	0.2957 ± 0.0113	0.5953 ± 0.0636	0.1536 ± 0.0113	0.2412 ± 0.0214
PCG-BCE ONLY	0.7463 ± 0.0052	0.5397 ± 0.0067	0.3017 ± 0.0085	0.7149 ± 0.0861	0.1908 ± 0.0159	0.2853 ± 0.0321
PCA	0.7257 ± 0.0000	0.5342 ± 0.0000	0.2964 ± 0.0000	0.6906 ± 0.0000	0.2167 ± 0.0000	0.1817 ± 0.0000

rerun analysis supports these three improvements (Appendix B.8 and Figure 3). Together, the IID and OOD results demonstrate the value of detection-aware preservation for cell-type transfer across distribution shifts.

Multi-batch integration with scIB. Following the scIB framework of Luecken et al. (2022), we assess batch-effect removal separately from conservation of biological variation and combine them with the framework's 40/60 weighting into a Total score. The lung benchmark therefore assesses how the learned embeddings balance batch removal with biological organization.

Among the controlled variants, DET-JEPA attains the highest Bio and Total (0.6299, 0.5207), while pure VICReg attains the highest Batch but the lowest Bio and Total. Relative to VISReg ONLY, DET-JEPA raises Bio by $+0.068$ and Total by $+0.039$ with negligible Batch change (-0.004). Detection-aware preservation thus improves biological conservation while maintaining a similar level of batch removal. scVI is reported as an external reference retrained on this benchmark with batch labels, which attains a higher Batch score (0.4620) but lower Bio (0.5961) than DET-JEPA, illustrating that supervised batch correction and biological conservation can diverge.

A qualitative view of this effect is given in Figure 5 (Appendix C.2): under the same visualization procedure, VISReg-only embeddings form elongated, overlapping structures, whereas DET-JEPA produces more compact cell-type regions. The individual scIB metrics are reported in Appendix C.2.

Multiomics protein-abundance prediction. Finally, we evaluate transfer to continuous protein-abundance prediction. Figure 2b reports Pearson and Spearman correlations together with MSE, measuring linear correlation, rank agreement, and prediction error. The margins on this benchmark are small: even the NoSSL baseline already reaches 0.8932 Pearson, and every method falls within a band of about 0.003, so fine-grained protein regression may offer limited discriminative power for separating pretraining objectives. However, across the controlled JEPA variants DET-JEPA still attains the highest mean Pearson and Spearman correlations and the lowest MSE, leading on all three metrics simultaneously, and the ordering is directionally consistent with the cell-type and scIB results, demonstrating that the gains are stable rather than an artifact of a single favorable measure.

4.4 ABLATION STUDY

We examine three design dimensions under the shared evaluation protocol: the anti-collapse regularizer, the preservation branch, and the design of the detection-aware objective. The testing protocol is described in Appendix B.8; results are shown in Figure 3.

Anti-collapse regularizer. Fixing the preservation branch to PCG-BCE, VISReg significantly outperforms both SIGReg and VICReg on all three OOD cohorts (Table 1). The mean OOD macro-F1 improves by 0.1535 over SIGReg and by 0.1312 over VICReg. On multiomics prediction, VISReg is also the best on all three metrics (Figure 2b). Paired effect sizes are large across all comparisons (Figure 3a), supporting VISReg as the anti-collapse regularizer of choice in DET-JEPA.

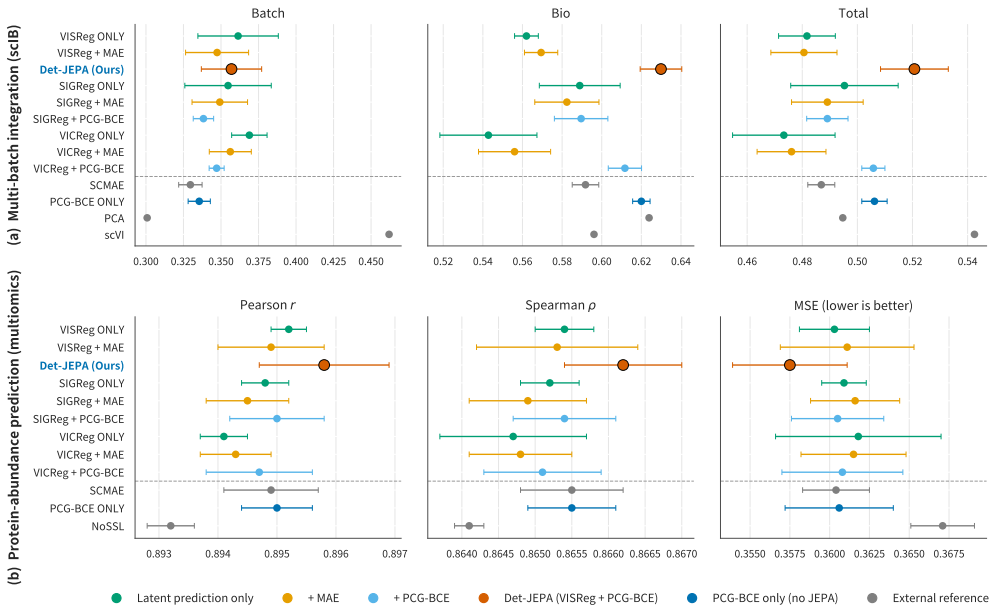


Figure 2: **Detection-aware preservation improves biological conservation and transfers across task families.** (a) Multi-batch integration on the scIB lung benchmark (Batch, Bio, and Total composite scores). (b) Protein-abundance prediction on the NeurIPS multiomics benchmark (Pearson r, Spearman ρ, and MSE; for MSE lower is better). Points show the mean over five independent pretraining runs; whiskers show ± 1 standard deviation. Rows are ordered identically across panels: controlled JEPA variants above the dashed line, external references below it; DET-JEPA is emphasized. Exact values are reported in Tables 3 and 4 (Appendix C).

Preservation complements collapse prevention. Adding detection-aware preservation (PCG-BCE) to a pure VISReg JEPA yields significant gains on every OOD cohort (Table 1) and raises scIB Bio from 0.5620 to 0.6299 (Figure 2a), confirming that collapse prevention alone is insufficient to retain biologically useful signal. The preservation target matters: predicting binary gene-detection patterns outperforms both magnitude reconstruction (VISReg + MAE) and reconstruction-free collapse prevention (VISReg only) on every OOD cohort and most multiomics metrics. The two-stage weighting in PCG-BCE further strengthens the objective: per-cell balancing accounts for a mean OOD gain of $+0.076$ over ordinary BCE, and gene-entropy weighting adds another $+0.023$, outperforming PC-BCE significantly on Great Apes and all three protein metrics (Figure 3b). A complete comparison of the three detection-objective variants (BCE, PC-BCE, and PCG-BCE) under VISReg is provided in Appendix C.3 (Table 5).

Contribution of the JEPA branch. Surprisingly, the PCG-BCE single-branch model already constitutes a strong baseline, demonstrating that detection-aware preservation alone is an effective standalone objective. Coupling it with latent prediction further improves the mean scIB Total from 0.5062 to 0.5207. Across the twelve metric columns of Table 1 and Figure 2, the dual-branch model is numerically better in eleven, with Great Apes the sole exception. Although these paired differences do not reach individual significance (Figure 3c), the consistency of this pattern together with no additional inference cost supports the dual-branch configuration as the preferred operating point in this design space. More challenging downstream tasks, such as perturbation response or temporal dynamics, may offer greater discriminative power for studying how the latent-prediction branch and the information-preservation branch complement each other, as discussed in Section 5.

5 DISCUSSION

Field context and our contribution. Single-cell foundation models and the virtual-cell agenda are moving the field toward transferable representations of cellular state. Recent single-cell JEPA vari-

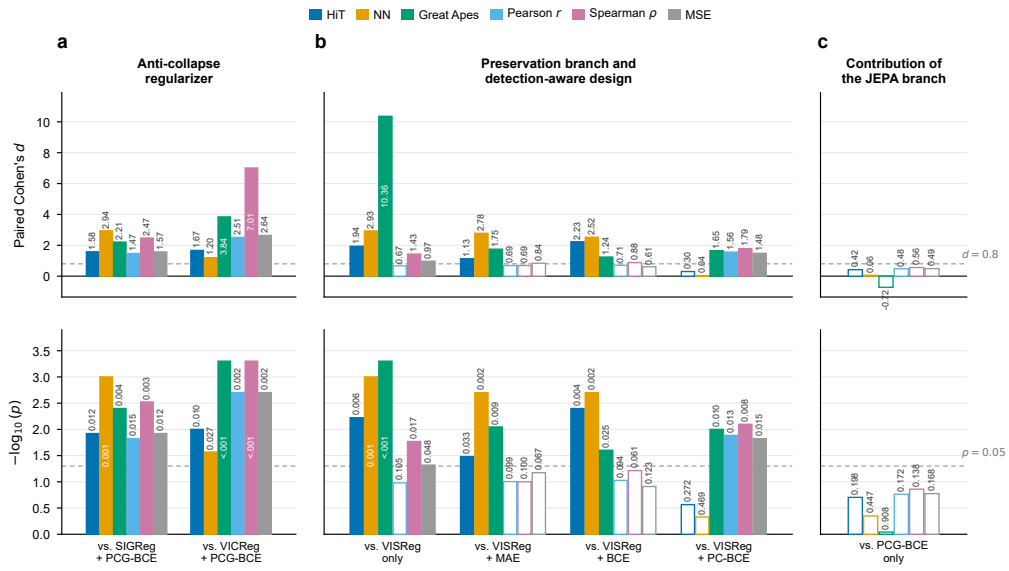


Figure 3: **Statistical evidence for all ablation claims.** Each bar group compares DET-JEPA against one alternative on six downstream metrics (HiT, NN, and Great Apes OOD macro-F1; multiomics Pearson r, Spearman ρ, and MSE), and every bar is annotated with its exact value. Top row: paired Cohen's d over the five pretraining seeds (**positive favors DET-JEPA**); the dashed line marks $d = 0.8$, the conventional large-effect threshold. Bottom row: $-\log_{10}(p)$ of the one-sided paired t-test; the dashed line marks $p = 0.05$, and bars with $p < 0.001$ are drawn at $p = 5 \times 10^{-4}$. Filled bars mark $p < 0.05$, hollow bars are not significant. (a) Anti-collapse regularizer; (b) preservation branch and detection-aware design; (c) contribution of the JEPA branch.

ants make latent prediction attractive, but leave open how to balance invariance with the preservation of biological information. Our controlled study organizes this design space around view construction, anti-collapse regularization, and information-preservation branches. It shows that avoiding collapse is not sufficient for downstream transfer, and that DET-JEPA improves representation quality by combining latent prediction with detection-aware preservation. The PCG-BCE branch adapts supervision to sparse single-cell measurements using per-cell balancing and online gene-entropy weighting. Its ablations indicate that the choice of preserved target and its weighting scheme matter alongside the anti-collapse regularizer. More broadly, these results suggest that single-cell SSL should ***exploit domain knowledge*** when designing losses, rather than directly importing objectives developed for computer vision or natural language processing.

Limitations and future work. This study holds data augmentation and view construction fixed, uses strategies previously shown to be effective in single-cell learning, and does not extensively explore encoder architectures. A central direction for future work is to incorporate biological priors across all three dimensions of the design space. Such priors can guide (i) view augmentations that perturb technical variation while preserving cell identity, (ii) anti-collapse regularizers that respect structured heterogeneity among cell states, and (iii) information-preservation targets that capture gene programs or pathway activity beyond individual gene detection. Jointly designing these components is important because view construction specifies the desired invariances, regularization shapes the latent geometry, and the preservation target determines which biological signals must remain accessible. These choices should also be evaluated with higher-capacity encoders.

Extending single-cell JEPA from static snapshots to temporally resolved representations is another open direction. In vision, JEPA progressed from I-JEPA on static images to V-JEPA on temporal video and V-JEPA 2 at world-model scale. Analogously, longitudinal and developmental measurements could support representations of differentiation, disease progression, and perturbation response over time. We leave these as future work.

Reproducibility Statement

All benchmarks are public: scTab/CELLxGENE for pretraining and IID splits, the published adult brain and comparative cohorts for OOD transfer, the scIB lung benchmark for integration, and the NeurIPS multiomics benchmark for RNA-to-protein prediction. Preprocessing, splits, leakage controls, architectures, masks, objective coefficients, seeds, and evaluation protocols are specified in Section 4 and Appendices A, A.2, B, and C.

Ethics Statement

This work aims to advance self-supervised learning methods for single-cell transcriptomics. All data used in this study come from publicly available, de-identified datasets, and all evaluation benchmarks and protocols are public and transparent. No human subjects were recruited, no private or identifiable data were collected, and the study raises no ethical concerns at this stage.

AI Usage Statement

Large language models were used during the preparation of this work to polish the writing and improve readability, to assist in the derivation and verification of mathematical formulations, and to aid in literature survey and the analysis of related work. The authors retain full responsibility for the content, accuracy, and integrity of the manuscript.

References

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025. URL https://arxiv.org/abs/2506.09985.

Randall Balestriero and Yann LeCun. Lejepa: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025. URL https://arxiv.org/abs/2511.08544.

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

Adrien Bardes, Jean Ponce, and Yann LeCun. MC-JEPA: A joint-embedding predictive architecture for self-supervised learning of motion and content features. *arXiv preprint arXiv:2307.12698*, 2023. URL https://arxiv.org/abs/2307.12698.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024. URL https://arxiv.org/abs/2404.08471.

Charlotte Bunne, Yusuf Roohani, Yanay Rosen, Ankit Gupta, Xikun Zhang, Marcel Roed, Theo Alexandrov, Mohammed AlQuraishi, Patricia Brennan, Daniel B Burkhardt, et al. How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell*, 187(25):7045–7063, 2024.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9630–9640, 2021. doi: 10.1109/ICCV48922.2021.00951.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *PMLR*, pp. 1597–1607, 2020.

Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *International Journal of Computer Vision*, 132:208–223, 2024. doi: 10.1007/s11263-023-01852-4.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15745–15753, 2021. doi: 10.1109/CVPR46437.2021.01549.

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.

Ali ElSheikh, Rui-Xi Wang, Weimin Wu, Yibo Wen, Payam Dibaeinia, Jennifer Yuntong Zhang, Jerry Yao-Chieh Hu, Mei Knudson, Sudarshan Babu, Shao-Hua Sun, Aly A. Khan, and Han Liu. Cell-jepa: Latent representation learning for single-cell transcriptomics. *arXiv preprint arXiv:2602.02093*, 2026. doi: 10.48550/arXiv.2602.02093. URL `https://arxiv.org/abs/2602.02093`.

Gökcen Eraslan, Lukas M. Simon, Maria Mircea, Nikola S. Mueller, and Fabian J. Theis. Single-cell rna-seq denoising using a deep count autoencoder. *Nature Communications*, 10(1):390, 2019. doi: 10.1038/s41467-018-07931-2.

Felix Fischer, David S. Fischer, Roman Mukhin, Andrey Isaev, Evan Biederstedt, Alexandra-Chloé Villani, and Fabian J. Theis. sctab: scaling cross-tissue single-cell annotation models. *Nature Communications*, 15(1):6611, 2024. doi: 10.1038/s41467-024-51059-5.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Wenkai Han, Yuqi Cheng, Jiayang Chen, Huawen Zhong, Zhihang Hu, Siyuan Chen, Licheng Zong, Liang Hong, Ting-Fung Chan, Irwin King, Xin Gao, and Yu Li. Self-supervised contrastive learning for integrative single cell RNA-seq data analysis. *Briefings in Bioinformatics*, 23(5): bbac377, 2022. doi: 10.1093/bib/bbac377.

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

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9726–9735, 2020. doi: 10.1109/CVPR42600.2020.00975.

Nikolas L. Jorstad, Janet H. T. Song, David Exposito-Alonso, Hamsini Suresh, Nathan Castro-Pacheco, Fenna M. Krienen, Anna Marie Yanny, Jennie Close, Emily Gelfand, Brian Long, et al. Comparative transcriptomics reveals human-specific cortical features. *Science*, 382(6667): eade9516, 2023. doi: 10.1126/science.ade9516.

Elon Litman, Tyler Myers, Vinayak Agarwal, Ekansh Mittal, Orion Li, Ashwin Gopinath, and Timothy Kassis. Genejepa: A predictive world model of the transcriptome. *bioRxiv*, 2025. doi: 10.1101/2025.10.14.682378. URL `https://doi.org/10.1101/2025.10.14.682378`.

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

Malte D Luecken, Daniel Bernard Burkhardt, Robrecht Cannoodt, Christopher Lance, Aditi Agrawal, Hananeh Aliee, Ann T Chen, Louise Deconinck, Angela M Detweiler, Alejandro A Granados, Shelly Huynh, Laura Isacco, Yang Joon Kim, Dominik Klein, BONY DE KUMAR, Sunil Kuppasani, Heiko Lickert, Aaron McGeever, Honey Mekonen, Joaquin Caceres Melgarejo, Maurizio Morri, Michaela Müller, Norma Neff, Sheryl Paul, Bastian Rieck, Kaylie Schneider, Scott Steelman, Michael Sterr, Daniel J. Treacy, Alexander Tong, Alexandra-Chloe Villani, Guilin Wang, Jia Yan, Ce Zhang, Angela Oliveira Pisco, Smita Krishnaswamy, Fabian J Theis, and Jonathan M. Bloom. A sandbox for prediction and integration of DNA, RNA, and proteins in single cells. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL https://openreview.net/forum?id=gN35BGa1Rt.

Malte D. Luecken, M. Büttner, K. Chaichoompu, A. Danese, M. Interlandi, M. F. Müller, D. C. Strobl, L. Zappia, M. Dugas, M. Colomé-Tatché, and Fabian J. Theis. Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods*, 19(1):41–50, 2022. doi: 10.1038/s41592-021-01336-8.

Olga Ovcharenko, Florian Barkmann, Philip Toma, Imant Daunhawer, Julia E Vogt, Sebastian Schelter, and Valentina Boeva. scSSL-bench: Benchmarking self-supervised learning for single-cell data. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 47416–47442. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/ovcharenko25a.html.

Pia Rautenstrauch and Uwe Ohler. Shortcomings of silhouette in single-cell integration benchmarking. *Nature Biotechnology*, 44(6):954–958, 2026. doi: 10.1038/s41587-025-02743-4.

Till Richter, Mojtaba Bahrami, Yufan Xia, David S. Fischer, and Fabian J. Theis. Delineating the effective use of self-supervised learning in single-cell genomics. *Nature Machine Intelligence*, 7 (1):68–78, 2025. doi: 10.1038/s42256-024-00934-3.

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

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

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

Haiyu Wu, Randall Balestriero, and Morgan Levine. VISReg: Variance-invariance-sketching regularization for JEPA training. *arXiv preprint arXiv:2606.02572*, 2026. doi: 10.48550/arXiv.2606.02572. URL https://arxiv.org/abs/2606.02572.

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

Yuyao Zhai, Weixu Wang, Till Richter, and Fabian J. Theis. scJEPA: Dual-space self-supervised learning for single-cell representations. In *LMRL Workshop at ICLR 2026*, 2026.

A Detailed Methods and Implementation

A.1 Regularizer objectives

All three anti-collapse regularizers act on batches of projector outputs $\boldsymbol{Z} \in \mathbb{R}^{B\times D}$, rather than the encoder embeddings in $\mathbb{R}^H$. They are averaged across views as in Eq. equation 2 and combined with the same prediction loss; only $\mathscr{R}(\boldsymbol{Z})$ changes.

VICReg. For two view batches $\boldsymbol{Z}$ and $\boldsymbol{Z}'$, the original VICReg objective is $\ell(\boldsymbol{Z},\boldsymbol{Z}') = \lambda_{\text{inv}}\,\mathrm{s}(\boldsymbol{Z},\boldsymbol{Z}') + \lambda_{\text{var}}[\mathrm{v}(\boldsymbol{Z}) + \mathrm{v}(\boldsymbol{Z}')] + \lambda_{\text{cov}}[\mathrm{c}(\boldsymbol{Z}) + \mathrm{c}(\boldsymbol{Z}')]$, where s is the mean squared embedding distance, $\mathrm{v}(\boldsymbol{Z}) = \frac{1}{D}\sum_{j=1}^{D} \max(0, \gamma - \sqrt{\mathrm{Var}(\boldsymbol{Z}_{\cdot j}) + \varepsilon})$ is a hinged standard-deviation penalty with target $\gamma = 1$, and $\mathrm{c}(\boldsymbol{Z}) = \frac{1}{D}\sum_{i\neq j}[\mathrm{Cov}(\boldsymbol{Z})]_{ij}^2$ penalizes off-diagonal covariance. In our controlled objective, the invariance role of s is played by Eq. equation 1; to avoid double counting, $\mathscr{R}_{\text{VICReg}}$ contains only the variance and covariance penalties. It constrains second-order moments without embedding normalization, but its covariance term costs $\mathscr{O}(BD^2)$.

SIGReg. SIGReg sketches a distributional test toward $\mathscr{N}(\mathbf{0}, \boldsymbol{I}_D)$ using random unit directions $\{\boldsymbol{a}_m\}_{m=1}^{M}$. For each direction, the scalar projections $X_m = \{\boldsymbol{a}_m^\top \boldsymbol{z}_n\}_{n=1}^{B}$ are compared with the standard normal characteristic function $\varphi(t) = e^{-t^2/2}$ by an Epps–Pulley statistic,

$$\mathrm{EP}(X_m) = B\int_{-\infty}^{\infty} |\hat{\varphi}_{X_m}(t) - \varphi(t)|^2\, w(t)\, dt, \quad \hat{\varphi}_{X_m}(t) = \frac{1}{B}\sum_{n=1}^{B} e^{itX_{mn}}, \tag{11}$$

with Gaussian window $w(t) = e^{-t^2/\varsigma^2}$; the regularizer averages $\mathrm{EP}(X_m)$ over directions. The Gaussian window downweights large $|t|$ in this discrepancy.

VISReg implementation. VISReg (Wu et al., 2026) is defined in Eqs. equation 3–equation 5. We use K random slices per batch; the sliced-Wasserstein shape term costs $\mathscr{O}(BDK) + \mathscr{O}(KB\log B)$ and avoids forming a $D\times D$ covariance matrix. The stop-gradient denominator in $\tilde{\boldsymbol{Z}}$ blocks gradients through scale normalization. The formulation uses neither an EMA teacher nor negative samples.

A.2 Hyperparameters and training details

Masking strategy. We use random masking throughout: each gene is independently selected as a target with probability 0.5. Target entries are set to zero in the encoder input. The unmasked entries provide context for predicting the representation of the original global view. JEPA and MAE-style comparisons share the same mask generator, and preservation losses use the original full-gene targets.

Architecture and regularizer implementation details. In the notation of Section 3, classification/scIB uses encoder dimension $H = 64$ and projector dimension $D = 128$; multiomics uses $H = 40$ and $D = 64$. Both JEPA loss terms act in $\mathbb{R}^D$, while preservation decoders take inputs in $\mathbb{R}^H$. The classification/scIB encoder is an MLP with layer widths $19{,}331 \to 512 \to 512 \to 256 \to 256 \to 64$; the embedding dimension is 64. The projector q_ψ is a 3-layer MLP $64 \to 128 \to 128 \to 128$ (hidden size 128). The preservation decoder g_ϕ used for detection-aware preservation is $64 \to 256 \to 256 \to 512 \to 512 \to 19{,}331$, producing gene-wise logits. The MAE magnitude decoder has the same widths but predicts normalized expression values. All layers use ReLU activation. Other architectural choices follow the scJEPA setup of Zhai et al. (2026) and the pretraining protocol of Richter et al. (2025).

VISReg (Wu et al., 2026) is implemented with $K = 1024$ random unit directions per batch; the internal scale, shape, and center weights are set to $\lambda_{\text{scale}} = \lambda_{\text{shape}} = \lambda_{\text{center}} = 1.0$, and the centering epsilon is 10^{-4}. SIGReg uses $M = 1024$ random unit directions. For each direction the empirical characteristic function is compared with the standard-normal CF $\varphi(t) = e^{-t^2/2}$ on a trapezoidal grid $t \in \{-5.0, -4.375, \ldots, 5.0\}$ with 17 points; the Epps–Pulley statistic is integrated over this grid with Gaussian window $w(t) = e^{-t^2/\varsigma^2}$, $\varsigma^2 = 2$.

The multiomics variant uses the 2,000 highly variable genes selected by Scanpy as input. Its encoder is $2{,}000 \to 256 \to 256 \to 40$ (embedding dimension 40), and its projector is a 3-layer MLP $40 \to 64 \to 64 \to 64$ (hidden size 64).

Training objectives. All JEPA variants share the same encoder, masking, and projector; they differ only in the anti-collapse regularizer applied after the projector (Appendix A.1) and, optionally, the auxiliary preservation branch. The preservation branch is one of *None* (pure JEPA), *MAE* (magnitude reconstruction in input space), *BCE* (unweighted binary gene-detection prediction), *PC-BCE* (per-cell balanced binary cross-entropy), and *PCG-BCE* (PC-BCE further weighted by per-gene detection entropy). PCG-BCE first rebalances positives within each cell and then multiplies by gene-wise entropy weights computed from an online EMA ($\beta = 0.99$). We use $\alpha = 0.5$, $\varepsilon = 10^{-6}$, and $\eta_{\min} = 0.1$.

Full-gene detection supervision. The detection-aware branch computes BCE over all G genes. Visible entries anchor cell-specific detection structure, while masked entries provide predictive supervision. Detection labels are used during pretraining; downstream evaluation uses encoder embeddings. Both detection and magnitude preservation supervise the full gene vector under the same encoder, masks, and training budget.

For every regularizer–preservation setting, we performed an independent grid search using the same search budget and validation-based selection procedure. Table 2 lists the selected task-specific view counts, regularizer coefficients, auxiliary-branch coefficients, and batch sizes.

Table 2: Validation-selected task-specific hyperparameters for JEPA variants. Each regularizer–preservation setting was tuned independently by grid search under the same search budget and selection protocol.

	Classification / scIB	Multiomics
Local views	8	1
VISReg coefficient	0.8	2.0
SIGReg coefficient	0.02	0.05
VICReg coefficient	1.0	1.0
MAE branch coefficient	0.2	0.2
Detection branch coefficient	1.5	1.0
Batch size	4,096	8,192

Optimization. Classification and scIB pretraining use AdamW, whereas multiomics pretraining uses Adam. Both use a learning rate of 10^{-3}, weight decay 0.05, and one epoch over the task-specific pretraining set. Results aggregate five independent pretraining seeds.

Computational cost. All pretraining runs were conducted on a single NVIDIA A100 GPU. Under our experimental setting, training is not GPU-bound: because the pretraining corpus is large, the pipeline is dominated by CPU-side data loading, and once a data slice has been read, the GPU completes the corresponding computation quickly. Each pretraining run (one epoch over the task-specific pretraining set) takes approximately two days of wall-clock time.

B Datasets and Experimental Protocol

B.1 Pretraining data: scTab

The pretraining dataset is scTab (Fischer et al., 2024), derived from the CELLxGENE census (version 2023-05-15). It encompasses approximately 22.1 million human single-cell RNA-seq profiles spanning 164 cell types, 5,052 donors, and 56 tissues. The data are restricted to primary 10x-based sequencing protocols. Donor-disjoint partitions comprise approximately 15.2M cells for training, 3.5M for validation, and 3.4M for testing, ensuring that no donor appears in more than one split.

B.2 IN-DISTRIBUTION BENCHMARKS

The IID benchmarks are CELLxGENE census subsets tailored to specific biological contexts (Richter et al., 2025). Their train/validation/test splits are subsets of the corresponding scTab partitions, so they share the same donor-disjoint structure.

HLCA (Human Lung Cell Atlas). 775,790 cells after filtering, 51 cell types, focused on lung tissue and providing a dense sampling of respiratory cell types.

PBMC (peripheral blood mononuclear cells). 78,354 cells after filtering, 30 cell types, derived from circulating immune cells after SARS-CoV-2 infection, vaccination, and healthy controls.

Tabula Sapiens. 335,861 cells after filtering, 161 cell types. A multi-tissue atlas with the greatest cell-type diversity among the IID benchmarks.

B.3 OUT-OF-DISTRIBUTION BENCHMARKS

The OOD benchmarks are datasets published after the CELLxGENE census version used to construct scTab and are therefore fully disjoint from all pretraining data (Richter et al., 2025). Each dataset is filtered to the shared 19,331-gene vocabulary; missing genes are zero-padded. All datasets are normalized to 10,000 counts per cell and $\log(1+x)$ transformed. The full dataset is used as the test split.

HiT (Human Brain Atlas, hippocampal subregion). Cells from the dissection of the tail of the hippocampus (caudal hippocampus, CA4–DGC) in the adult human brain atlas (Siletti et al., 2023). This anatomically specific subset targets excitatory and inhibitory neuron subtypes as well as local glial populations, testing fine-grained neuronal transfer in a novel anatomical context.

NN (Human Brain Atlas, non-neuronal cells). All non-neuronal cells from the adult human brain atlas (Siletti et al., 2023), including astrocytes, oligodendrocytes, oligodendrocyte precursor cells, microglia, and vascular cell types. This benchmark probes transfer to glial and vascular lineages that are under-represented in the neuron-dominated pretraining data.

Great Apes (comparative brain transcriptomics). Comparative single-cell transcriptomics of cortical cells from human and great ape species (Jorstad et al., 2023). This introduces a species-level distribution shift: the pretraining data are exclusively human, so the model must rely on evolutionarily conserved transcriptomic signatures for transfer.

B.4 MULTIOMICS BENCHMARK

The NeurIPS multiomics dataset (Luecken et al., 2021) is a multimodal bone marrow collection integrating gene-expression counts with CITE-seq proteomics data. It underwent the same library-size normalization and $\log(1+x)$ transformation as the other datasets. We selected 2,000 highly variable genes using Scanpy; these features are used as input during both multiomics pretraining and fine-tuning. We used a random 80/10/10 train/validation/test split, following the established evaluation protocol (Richter et al., 2025).

B.5 PREPROCESSING

All datasets undergo the same preprocessing pipeline (Richter et al., 2025): library-size normalization to 10,000 counts per cell followed by $\log(1+x)$ transformation. The gene vocabulary consists of the 19,331 human protein-coding genes defined by scTab (Fischer et al., 2024). For IID and OOD classification benchmarks, each dataset is filtered to this shared vocabulary; missing genes are zero-padded. For the multiomics benchmark, 2,000 highly variable genes are selected from the vocabulary using Scanpy prior to training. Multiomics pretraining uses this 2,000-gene input directly (rather than the 19,331-gene input used for classification and scIB), and the same selected features are used during fine-tuning, following the established evaluation protocol (Richter et al., 2025).

B.6 scIB EVALUATION FRAMEWORK

We evaluate multi-batch integration using the scIB benchmarking framework (Luecken et al., 2022), which assesses integration quality along two orthogonal axes: batch-effect removal and biological conservation. Each axis is composed from multiple complementary metrics so that no single measure dominates the composite score.

Batch-effect removal metrics.

- **kBET** (k-nearest-neighbor batch effect test): Tests whether the local batch composition around each cell matches the global batch proportions. A higher acceptance rate indicates better mixing.
- **Graph connectivity**: Measures the fraction of cells in the largest connected component per cell-type label in the kNN graph, assessing whether cells of the same type form a single connected subgraph across batches.
- **PCR batch** (PCA regression): Fits a linear regression of the principal components against batch labels and reports $1 - R^2$, so higher values indicate that batch explains less variance in the embedding.
- **iLISI** (integration LISI): Measures local batch diversity; higher values indicate better batch mixing within neighbors.
- **BRAS** (batch-removal-adapted silhouette): A silhouette-based batch-mixing score computed within each cell type; unlike standard batch ASW, it compares each cell against all other batches rather than only the nearest one, making it robust to nested batch effects (Rautenstrauch & Ohler, 2026).

Biological conservation metrics.

- **NMI** (normalized mutual information): Compares Leiden clusters against cell-type labels; higher values indicate that clusters correspond to annotated types.
- **ARI** (adjusted Rand index): A chance-adjusted measure of agreement between clusters and labels.
- **Silhouette** (average silhouette width): Quantifies how well each cell sits within its annotated type relative to other types, ranging from -1 to 1.
- **cLISI** (cell-type local inverse Simpson's index): Measures local neighborhood purity with respect to cell-type labels; a score near 1 indicates that each neighborhood contains a single cell type.

These nine reported metrics are a subset of the broader scIB metric suite and are aggregated into a **Batch** composite and a **Bio** composite following the weighting scheme of Luecken et al. (2022). In particular, iLISI and graph connectivity are treated as batch-effect removal metrics, while cLISI and the label-based metrics are treated as biological-conservation metrics. The composites are then combined into a **Total** score ($0.4 \times \text{Batch} + 0.6 \times \text{Bio}$). The metric-level breakdown for all evaluated settings is visualized in Figure 4 (Appendix C.2).

B.7 EVALUATION PROTOCOL DETAILS

Protocol provenance. Unless noted otherwise, we use the original settings of Richter et al. (2025): scTab is derived from the CELLxGENE census 2023-05-15 release and is partitioned by donor; the IID atlas subsets inherit their corresponding scTab allocations; the unseen cohorts are used only for testing after filtering to the scTab gene vocabulary, zero-padding missing genes, and applying library-size normalization to 10,000 counts followed by $\log(1+x)$; and the NeurIPS multiomics task uses the published random 80/10/10 split with 2,000 Scanpy-selected highly variable genes. Our controlled JEPA runs do not use batch labels. The scVI row is an external reference retrained on the lung benchmark with its available batch labels, as described in the main text.

Cell-type classification (IID and OOD). All classification benchmarks use a zero-shot 5-NN classifier. The pretrained encoder extracts embeddings for both the reference pool and the test set; no encoder parameters are updated during evaluation. For IID benchmarks (HLCA, PBMC, Tabula Sapiens), the scTab validation set (3.5M cells) serves as the reference pool, following Richter et al. (2025). For OOD benchmarks (HiT, NN, Great Apes), the same scTab validation pool is used so that all methods are evaluated against an identical reference. The primary metric is macro-F1, which treats all cell types equally regardless of abundance. Embeddings are computed on the full gene vocabulary (19,331 genes) for classification; no HVG subselection is applied.

Table 3: Multi-batch integration under scIB. Batch, Bio, and Total denote batch-effect removal, biological conservation, and their aggregate. Darker blue-gray shading indicates a better column-wise rank (best = darkest); shading applies to the evaluated JEPA variants; external baselines are unranked.

Method	Batch	Bio	Total
VISReg ONLY	0.3614 ± 0.0268	0.5620 ± 0.0060	0.4817 ± 0.0103
VISReg + MAE	0.3474 ± 0.0210	0.5694 ± 0.0084	0.4806 ± 0.0120
DET-JEPA (Ours)	0.3570 ± 0.0201	$\mathbf{0.6299 \pm 0.0105}$	$\mathbf{0.5207 \pm 0.0123}$
SIGReg ONLY	0.3547 ± 0.0288	0.5889 ± 0.0204	0.4953 ± 0.0195
SIGReg + MAE	0.3492 ± 0.0185	0.5824 ± 0.0162	0.4891 ± 0.0130
SIGReg + PCG-BCE	0.3383 ± 0.0068	0.5896 ± 0.0135	0.4891 ± 0.0075
VICReg ONLY	$\mathbf{0.3689 \pm 0.0118}$	0.5428 ± 0.0245	0.4733 ± 0.0186
VICReg + MAE	0.3562 ± 0.0140	0.5560 ± 0.0182	0.4761 ± 0.0125
VICReg + PCG-BCE	0.3471 ± 0.0050	0.6117 ± 0.0084	0.5058 ± 0.0042
SCMAE	0.3296 ± 0.0078	0.5918 ± 0.0067	0.4869 ± 0.0049
PCG-BCE ONLY	0.3355 ± 0.0074	0.6200 ± 0.0044	0.5062 ± 0.0046
PCA	0.3008 ± 0.0000	0.6239 ± 0.0000	0.4947 ± 0.0000
scVI	0.4620 ± 0.0000	0.5961 ± 0.0000	0.5425 ± 0.0000

Multi-batch integration (scIB). We follow the scIB benchmark (Luecken et al., 2022) on the lung subset of the scTab test split, using dataset IDs 131, 59, and 205 from the CELLxGENE census (56,771 cells, 40 cell types, 9 batches), consistent with the published evaluation protocol (Richter et al., 2025). The scIB pipeline computes batch correction scores (including kBET, graph connectivity, iLISI, BRAS, and PCA-based metrics) and biological conservation scores (NMI, ARI, silhouette, and cLISI). These are aggregated into Batch, Bio, and Total composite scores via weighted means, with higher values indicating better integration. The PCA reference uses raw-gene PCA embeddings.

Multiomics protein-abundance prediction. We follow the paired CITE-seq protocol from the NeurIPS multiomics benchmark (Luecken et al., 2021). The dataset is a multimodal bone marrow collection with coupled RNA and protein measurements. We select 2,000 highly variable genes using Scanpy and predict all 134 surface protein markers. The dataset is split into training, validation, and test sets using a random 80/10/10 partition. Pretrained encoders are fine-tuned on paired RNA–protein cells for up to 2,000 epochs. We report test-set Pearson r, Spearman ρ, and MSE. The NoSSL reference is a network with the same encoder architecture trained directly on this supervised task, without SSL pretraining.

B.8 Statistical testing protocol for the ablation claims

We assess the design comparisons discussed in Section 4.4 with paired tests over the five independent pretraining reruns. Reruns of two compared settings share the same random seeds and are therefore treated as paired samples. We use a one-sided paired t-test as the primary statistic and report Cohen's d as an effect-size measure ($d > 0$ favors DET-JEPA; for MSE, d is computed so that positive values indicate lower error). All reported p-values are uncorrected. The complete results are consolidated in Figure 3.

C Additional Experimental Results

C.1 Result tables for the scIB and multiomics benchmarks

Tables 3 and 4 report the exact mean and standard-deviation values behind Figure 2 in the main text.

C.2 scIB metric-level comparison

Table 3 reports the composite Batch, Bio, and Total scores. To provide a more granular view, Figure 4 displays all nine individual scIB metrics together with the three composite scores for five representative settings. The evaluation framework and metric definitions are described in Appendix B.6.

Table 4: Multiomics protein prediction. Darker blue-gray shading indicates a better column-wise rank (best = darkest); shading applies to the evaluated JEPA variants; external baselines are unranked.

Method	Pearson r	Spearman ρ	MSE
VISReg ONLY	0.8952 ± 0.0003	0.8654 ± 0.0004	0.3603 ± 0.0022
VISReg + MAE	0.8949 ± 0.0009	0.8653 ± 0.0011	0.3611 ± 0.0042
DET-JEPA (Ours)	$\mathbf{0.8958 \pm 0.0011}$	$\mathbf{0.8662 \pm 0.0008}$	$\mathbf{0.3575 \pm 0.0036}$
SIGReg ONLY	0.8948 ± 0.0004	0.8652 ± 0.0004	0.3609 ± 0.0014
SIGReg + MAE	0.8945 ± 0.0007	0.8649 ± 0.0008	0.3616 ± 0.0028
SIGReg + PCG-BCE	0.8950 ± 0.0008	0.8654 ± 0.0007	0.3605 ± 0.0029
VICReg ONLY	0.8941 ± 0.0004	0.8647 ± 0.0010	0.3618 ± 0.0052
VICReg + MAE	0.8943 ± 0.0006	0.8648 ± 0.0007	0.3615 ± 0.0033
VICReg + PCG-BCE	0.8947 ± 0.0009	0.8651 ± 0.0008	0.3608 ± 0.0038
SCMAE	0.8949 ± 0.0008	0.8655 ± 0.0007	0.3604 ± 0.0021
PCG-BCE ONLY	0.8950 ± 0.0006	0.8655 ± 0.0006	0.3606 ± 0.0034
NoSSL	0.8932 ± 0.0004	0.8641 ± 0.0002	0.3671 ± 0.0020

The metric-level view reveals several patterns that are not visible from the composite scores alone. VISReg-only JEPA achieves the highest PCR batch score, while VICReg has the best iLISI among the reported regularizer-only settings; VISReg-only JEPA also scores lowest on the biological-conservation metrics NMI, ARI, Silhouette, and cLISI as well as on the batch-removal metric BRAS, illustrating the invariance–preservation trade-off at the individual-metric level. The detection-aware branch substantially improves biological conservation metrics while slightly reducing batch-mixing scores, consistent with the composite-level finding in Table 3. MAE occupies an intermediate position on most metrics, while the PCA reference attains the highest NMI and BRAS but the lowest PCR and Batch scores.

Figure 5 provides a qualitative view of the same effect at the embedding level, comparing VISReg-only JEPA with DET-JEPA on identical cells.

C.3 DETECTION OBJECTIVE COMPARISON

Table 5 reports the per-run results for the three detection-objective variants under VISReg: ordinary BCE, PC-BCE (per-cell balanced), and PCG-BCE (per-cell balanced plus gene-entropy weighting), the last of which is the preservation branch used by DET-JEPA. For both OOD cell-type transfer and multiomics protein prediction, PCG-BCE achieves the best or near-best performance across all metrics, confirming that the two-stage weighting improves over plain BCE and the intermediate PC-BCE.

Table 5: Detection-objective comparison under VISReg. Left: OOD macro-F1 (higher is better). Right: Multiomics protein prediction (Pearson r and Spearman ρ higher is better; MSE lower is better).

Method	**OOD macro-F1**			**Multiomics**		
	HiT	NN	Great Apes	Pearson r	Spearman ρ	MSE
DET-JEPA (Ours)	0.7525 ± 0.0961	0.1927 ± 0.0237	0.2530 ± 0.0187	0.8958 ± 0.0011	0.8662 ± 0.0008	0.3575 ± 0.0036
VISReg + PC-BCE	0.7277 ± 0.0736	0.1915 ± 0.0242	0.2090 ± 0.0186	0.8941 ± 0.0016	0.8647 ± 0.0012	0.3621 ± 0.0047
VISReg + BCE	0.5586 ± 0.0702	0.1322 ± 0.0065	0.2106 ± 0.0183	0.8947 ± 0.0016	0.8651 ± 0.0013	0.3615 ± 0.0060

D AUXILIARY PRESERVATION BRANCHES AND DETECTION-AWARE OBJECTIVES

Magnitude reconstruction (MAE). For one view, a decoder r_ω maps the pre-projector encoder embedding $\boldsymbol{h}_{iv} = f_\theta(\boldsymbol{x}_i^{(v)})$ to expression magnitudes over the full gene vector:

$$\mathscr{L}_{\text{MAE}} = \frac{1}{BG} \sum_{i=1}^{B} \sum_{g=1}^{G} (x_{ig} - \hat{x}_{ig})^2, \qquad \hat{x}_{ig} = [r_\omega(\boldsymbol{h}_{iv})]_g. \tag{12}$$



Figure 4: Metric-level scIB comparison for five representative settings on the lung test-set benchmark (56,771 cells, 40 cell types, 9 batches). Each panel shows one of the nine individual metrics or one of the three composite scores (Batch, Bio, Total). Markers indicate the mean across five independent runs; error bars show ± 1 standard deviation. The dashed gray line marks the PCA reference. The shaded region highlights the JEPA variant group. See Appendix B.6 for metric definitions.

The target is the original unmasked normalized expression. This branch plays the role of the preservation loss $\mathscr{L}_{\text{pres}}$ under the same gene, cell, and view averaging as Eq. equation 9, entering the full objective through the $\lambda_{\text{pres}}\mathscr{L}_{\text{pres}}$ term in Eq. equation 10; MAE here denotes the masked-autoencoder-style branch, whose reconstruction criterion is mean squared error.

Detection weighting. The BCE objective family is given in Eq. equation 9, with the gene-entropy weights defined in Eq. equation 8. For $0 < |P_i| < G$, PC-BCE assigns equal total coefficient mass $|N_i|$ to the positive and negative classes within each cell. The guard in the main-text definition yields negative-only BCE if $|P_i| = 0$ and zero loss if $|N_i| = 0$.

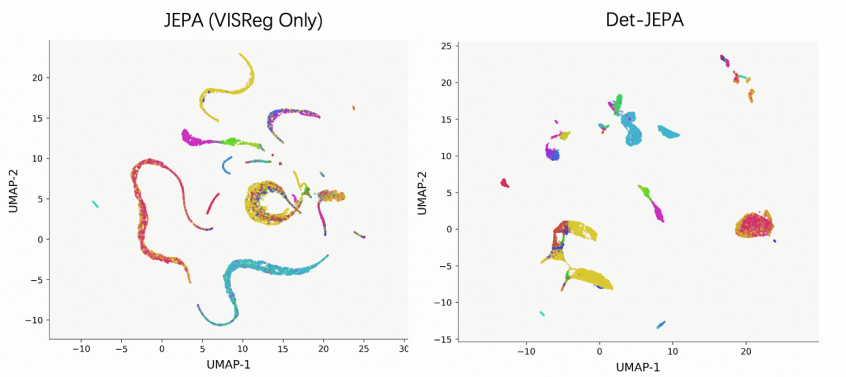


Figure 5: Effect of the information-preservation branch on the scIB lung embedding. UMAPs show the same 15,000 cells (40 cell types, 9 batches) for VISReg-only JEPA and DET-JEPA, colored by cell type. Adding detection-aware preservation yields more compact and visually separable cell-type regions, qualitatively complementing the Bio improvement in Figure 2a.

Observation model and objective properties. We characterize detection probabilities, invariance to positive-count magnitudes, and weighted loss gradients under the following observation model.

Count sampling and detection probability. Let $\mu_{ig} \geq 0$ denote biological expression, $d_i \geq 0$ a cell-level sampling/capture factor, and $\gamma_g \geq 0$ a gene-specific capture factor. Assume

$$c_{ig} \mid \delta_{ig} \sim \text{Poisson}(\delta_{ig}\lambda_{ig}), \qquad \lambda_{ig} = d_i\gamma_g\mu_{ig}. \tag{13}$$

Here $\delta_{ig} \in \{0,1\}$ is a Bernoulli capture indicator with $\pi_{ig} = \Pr(\delta_{ig} = 1)$ denoting the capture success probability; $\delta_{ig} = 0$ represents complete capture failure (the count is forced to zero), and $\pi_{ig} = 1$ recovers an ordinary Poisson sampling model. Conditional on the model parameters, the detection and non-detection probabilities are

$$\Pr(y_{ig} = 1) = \pi_{ig}(1 - e^{-\lambda_{ig}}), \qquad \Pr(y_{ig} = 0) = 1 - \pi_{ig} + \pi_{ig}e^{-\lambda_{ig}}. \tag{14}$$

For fixed π_{ig}, non-detection decreases with λ_{ig} but approaches the floor $1 - \pi_{ig}$. Thus a detection label can be stable at high expected counts, with sampling and capture sensitivity governed by λ_{ig} and π_{ig}.

Invariance to positive-count magnitudes. For cells with positive library size, normalization gives

$$x_{ig} = \log\left(1 + \frac{10^4 c_{ig}}{\sum_h c_{ih}}\right), \qquad y_{ig} = \mathbf{1}\{c_{ig} > 0\}. \tag{15}$$

If two observations both have positive counts for gene g, their detection labels are both one regardless of their positive magnitudes or normalization denominators. The detection target therefore removes within-positive magnitude variation, while the encoder continues to receive expression-valued inputs.

For a concrete sampling perturbation, independently retain each molecule with probability $r \in [0,1]$, so that $c'_{ig} \mid c_{ig} \sim \text{Binomial}(c_{ig}, r)$ and $y'_{ig} = \mathbf{1}\{c'_{ig} > 0\}$. Poisson thinning under Eq. equation 13 gives the exact label-flip probability

$$\Pr(y_{ig} \neq y'_{ig}) = \pi_{ig}\left(e^{-r\lambda_{ig}} - e^{-\lambda_{ig}}\right). \tag{16}$$

This expression quantifies detection-label sensitivity to molecule sampling. For fixed $r > 0$, the flip probability tends to zero as $\lambda_{ig} \to \infty$.

Weighted logit gradients. For one cell and one view, let $b_{ig} = a_{ig}$ for PC-BCE and $b_{ig} = a_{ig}\eta_g$ for PCG-BCE. Treat these target-derived weights as constants during backpropagation. For $\ell_b(i) = G^{-1}\sum_g b_{ig}\text{BCE}(s_{ig}, y_{ig})$,

$$\frac{\partial \ell_b(i)}{\partial s_{ig}} = \frac{b_{ig}}{G}(\text{sigmoid}(s_{ig}) - y_{ig}), \qquad \left|\frac{\partial \ell_b(i)}{\partial s_{ig}}\right| \leq \frac{b_{ig}}{G}. \tag{17}$$

Batch and view averaging add the corresponding averaging factors. The bound scales with the class and gene weights. In comparison, the derivative of the per-cell magnitude loss with respect to its output is

$$\frac{\partial \ell_{\mathrm{MAE}}(i)}{\partial \hat{x}_{ig}} = \frac{2}{G}(\hat{x}_{ig} - x_{ig}). \tag{18}$$

The BCE logit residual is bounded independently of the positive count magnitude, whereas the squared-error output gradient scales with its residual. The normalized magnitude targets satisfy $0 \leq x_{ig} \leq \log(10001)$.