PHYSICS-GUIDED GRAPH NETWORKS FOR GENE EXPRESSION RECOVERY IN SPATIALLY OVERLAPPING CELLS*

Anonymous authors
Paper under double-blind review

ABSTRACT

Overlapping segmented cell footprints can mix gene expression and obscure source profiles in Stereo-cell sequencing. We form paired mixtures by aligning two measured PBMC chips and translating one chip map relative to the other; directional polygon overlap sets donor contributions, while true overlap edges and source profiles remain hidden at inference. We present PhyGRec (Physics-Guided Graph Recovery), a four-block network with a learned circle-overlap surrogate, residual backprojection, adaptive increment modulation, and graph-context correction. Identity-disjoint evaluation uses three technically matched chips. On the fixed test split, PhyGRec achieves the lowest Count-MAE, Log-MAE, and Rel. L_1 among evaluated methods; its Log-MAE is 0.00369 versus 0.00888 for GATv2 (58.4% lower, based on three-seed means). Under strict receiver leave-one-chip-out evaluation within this chip series, it again has the lowest three errors, with Log-MAE of 0.00801 versus 0.01305 for GraphSAGE. Component removals support all three modules within the evaluated stack. These results support source recovery under controlled geometric crosstalk.

1 INTRODUCTION

Stereo-seq's high-density DNA-nanoball arrays couple expression measurements with positional information (Chen et al., 2022; Wei et al., 2022). Stereo-cell profiles dissociated-cell suspensions, including peripheral blood mononuclear cells (PBMCs), on these arrays and reports local lateral RNA diffusion (Liao et al., 2025). Overlapping segmented cell footprints can mix RNA signals and obscure cell-specific counts and marker profiles. We construct paired mixtures by translating one measured chip's footprint map relative to another, retaining single-chip profiles as Reference.

Existing tools address related sources of expression error: SAVER and DCA denoise single-cell profiles (Huang et al., 2018; Eraslan et al., 2019); SoupX and DecontX estimate ambient RNA contamination (Young & Behjati, 2020; Yang et al., 2020); and SpotClean corrects spatial spot swapping (Ni et al., 2022). Our task instead recovers cell-specific sources from directional mixing between segmented footprints.

Figure 1 contrasts physics-based inversion (Ongie et al., 2020; Monga et al., 2021), data-driven prediction, and physics-guided recovery. Source expression and transfer strengths are unknown at inference. Geometry limits plausible transfer but not its magnitude. We therefore build a directional forward operator, back-project its residual, and refine the gene-shared update with graph context.

PhyGRec (Physics-Guided Graph Recovery) stacks four independently parameterized Physics-Guided Recovery (PGR) Blocks. Each Block uses candidate distances and validity masks to construct a circle-overlap surrogate for directional mixing, back-projects a robust reconstruction residual, modulates the resulting increment, and adds a two-layer GATv2 correction (Brody et al., 2022). A nonnegative state update and positive output scaling produce the recovered expression.

To evaluate source recovery against known targets, we use measured single-chip PBMC profiles from three technically matched chips as Reference sources. Overlaps between translated segmented footprints generate directional mixtures whose overlap coefficients remain hidden, while independently

*Code is available at `https://anonymous.4open.science/r/phygrec/`.

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

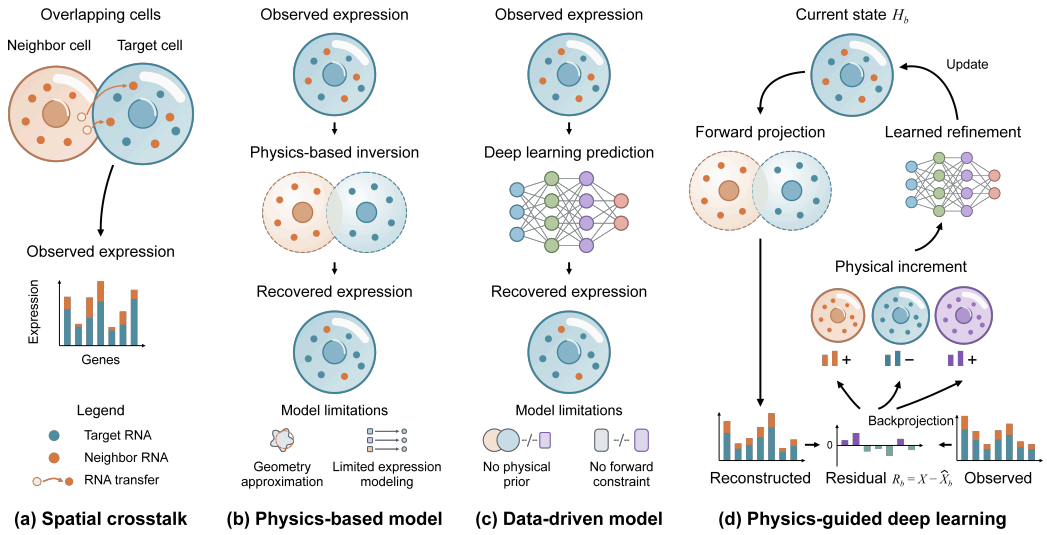

Figure 1: Spatial crosstalk and recovery strategies. **a**, RNA transfer between overlapping cell footprints mixes observed expression. **b**, physics-based inversion uses an explicit observation model. **c**, data-driven prediction learns expression recovery directly. **d**, physics-guided recovery combines forward reconstruction, residual backprojection, and learned refinement. All panels are schematic.

constructed $K = 8$ coordinate graphs provide model-visible neighborhoods. We split cell identities before scene generation so that each cell and all its translated appearances remain in the same split.

On the fixed test split and under strict receiver leave-one-chip-out (receiver-LOCO), PhyGRec attains the lowest three primary expression errors among the evaluated methods. Receiver-LOCO excludes the held-out chip from training, validation, and highly variable gene selection. Removing any of the three update modules raises error in the evaluated stack. Together, these results support residual-guided geometric recovery with graph refinement under controlled crosstalk, including transfer to held-out receiver chips.

2 Related Work

Expression recovery and contamination correction. Expression mixing between overlapping segmented cells calls for a receiver-specific account of contaminating counts. Expression recovery methods model profiles through probabilistic, autoencoding, or manifold priors (SAVER, DCA, and MAGIC) (Huang et al., 2018; Eraslan et al., 2019; van Dijk et al., 2018), while BayesSpace shares spatial information for subspot enhancement (Zhao et al., 2021). SoupX, DecontX, and CellBender address ambient or background contamination through distinct approaches (Young & Behjati, 2020; Yang et al., 2020; Fleming et al., 2023); SpotClean models spatial spot swapping (Ni et al., 2022). Stereo-cell reports local lateral diffusion at segmented-cell resolution (Liao et al., 2025). We therefore place directional transfer over candidate neighbors in the observation model, rather than using the neighborhood only as an expression prior.

Coordinate-aware transcriptomics. Spatial context can guide alignment and gene imputation (Tangram and SpaGE) (Biancalani et al., 2021; Abdelaal et al., 2020), cell-type composition estimates (Stereoscope, RCTD, and cell2location) (Andersson et al., 2020; Cable et al., 2022; Kleshchevnikov et al., 2022), or learned tissue representations (SpaGCN, GraphST, and SEDR) (Hu et al., 2021; Long et al., 2023; Xu et al., 2024); see also Bressan et al. (2023). In Stereo-cell suspensions, coordinates instead describe a technical deposition layout. We use them to define candidate assay interactions in a directional forward model, then use message passing to refine the inverse recovery.

Algorithm unrolling. Algorithm unrolling retains measurement structure within trainable updates (Ongie et al., 2020; Monga et al., 2021). LISTA learns sparse-coding updates (Gregor &

LeCun, 2010), ISTA-Net learns proximal-gradient transforms (Zhang & Ghanem, 2018), and learned primal–dual networks couple updates in tomography (Adler & Öktem, 2018). In our setting, the constrained trainable observation operator approximates directional transfer from chip geometry, while each PGR Block combines its residual backprojection with graph-context refinement and nonnegative recovery.

3 METHOD

3.1 PROBLEM FORMULATION

A mixing scene contains N cell nodes and G genes. Indices m, n denote cells, k a candidate-donor slot, g a gene, and b a recovery block. Let $S \in \mathbb{R}_+^{N\times G}$ denote source expression, represented by measured single-chip Reference profiles in our benchmark, and $X \in \mathbb{R}_+^{N\times G}$ the mixed observation. For receiver cell m,

$$X_{m:} = S_{m:} + \sum_{n\in\mathcal{D}^\star(m)} \beta^\star_{mn} S_{n:}, \qquad \beta^\star_{mn} \geq 0, \tag{1}$$

where $\mathcal{D}^\star(m)$ denotes the hidden set of overlapping donors and $\beta^\star_{mn}$ denotes a gene-shared directional transfer coefficient. At inference, a coordinate-derived graph $\mathcal{G} = (\mathcal{C}, \mathcal{E})$ supplies at most $K = 8$ candidate donors d_{mk}, distances δ_{mk}, and validity masks $c_{mk} \in \{0, 1\}$, with $\mathcal{C} = \{d_{mk}, \delta_{mk}, c_{mk}\}$ and message-passing edges $\mathcal{E}$. Candidates are the nearest other cells in joint translated coordinates before symmetrization; GCC uses both directions, so node degrees can exceed eight. The network maps $(X, \mathcal{G})$ to $\widehat{S}$; Reference profiles are training targets, not donor inputs.

Receiver and donor are roles in directed transfer relations. For example, two overlapping cells can have $X_{1:} = S_{1:} + \beta^\star_{12} S_{2:}$ and $X_{2:} = S_{2:} + \beta^\star_{21} S_{1:}$. Both input rows are mixed; both output rows are supervised when $\Omega = \{1, 2\}$. Candidate nodes without a receiver observation are initialized by the gene-wise median of observed node profiles and updated jointly, but are excluded from source-recovery supervision. Repeated observations of a node are combined by a gene-wise minimum to form the node-aligned input X.

The states H_b, output $\widehat{S}$, displacement V_b, and update increments have shape $N \times G$; increments may have either sign. The physical operator has one row per observed receiver and one column per scene node; D_b is node-space diagonal and W_b is receiver-space diagonal. Let P_Ω select observed receiver rows. Then $A_b \in \mathbb{R}^{|\Omega|\times N}$, $D_b \in \mathbb{R}^{N\times N}$, and $W_b \in \mathbb{R}^{|\Omega|\times|\Omega|}$. We use $[\cdot]_+$ for elementwise nonnegative clipping, $\odot$ for elementwise multiplication, and $\|$ for feature concatenation.

3.2 PHYGREC ARCHITECTURE

PhyGRec sequentially stacks $L = 4$ Physics-Guided Recovery (PGR) Blocks. All four Blocks perform the same computation, but their physical and neural parameters are independent. Throughout, Block indices are one-based: Block b receives H_{b-1} and produces H_b. The stack is

$$H_0 = X, \qquad H_b = \mathcal{B}_b(H_{b-1}, V_{b-1}, X, \mathcal{G}), \quad b \in \{1, \ldots, L\}. \tag{2}$$

We initialize $V_0 = 0$ and update $V_b = H_b - H_{b-1}$ after each projected block update. The displacement is carried into the next block as momentum; S and $\widehat{S}$ denote the source target and final recovery.

Figure 2 expands each PGR Block into Residual Backprojection (RB), Adaptive Increment Modulation (AIM), and Graph-Context Correction (GCC). These modules construct successive increments before a single projected state update.

3.3 BLOCK-SPECIFIC CIRCLE OPERATOR

Block b approximates the generator's segmented-polygon overlap using fixed X, candidate distances δ_{mk}, and validity masks c_{mk}; it receives no true footprints and does not recompute geometry from H_{b-1}. Expression totals from fixed X set the surrogate capture radii, so intermediate recovery states do not change the operator geometry. Let $a_n = \sum_g X_{ng}$, $\bar{a}_n = \max(a_n, 1)$, and let a_{ref} be the detached median of positive receiver totals and valid donor-slot totals, lower-bounded by one and

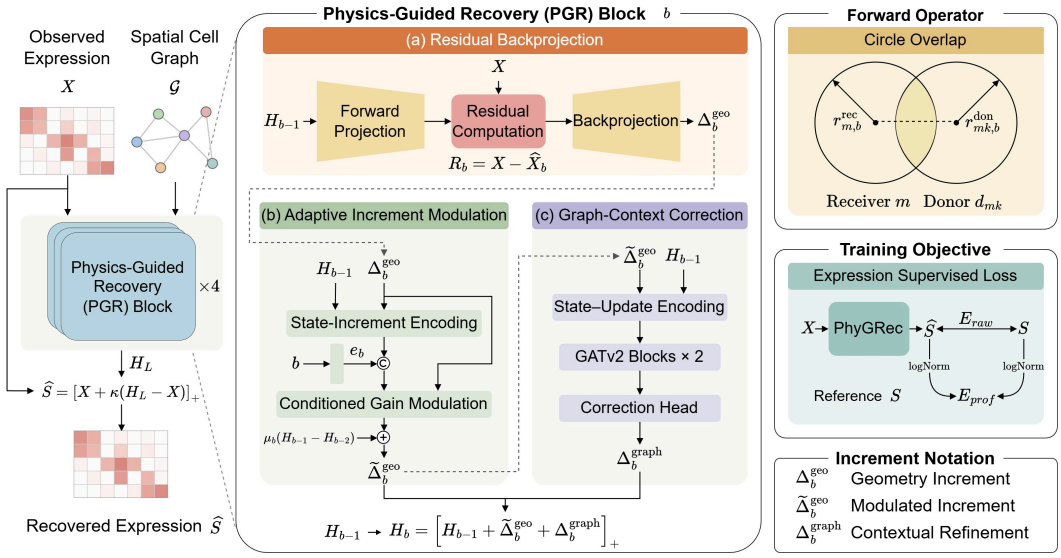

Figure 2: PhyGRec architecture. Each independently parameterized Physics-Guided Recovery (PGR) Block combines (a) Residual Backprojection (RB), (b) Adaptive Increment Modulation (AIM), and (c) Graph-Context Correction (GCC), followed by a nonnegative state update. The training objective is $E_{\text{prof}} + E_{\text{raw}}$. The residual panel abbreviates $P_\Omega X$ as X; Reference targets enter only the losses.

computed separately for each scene. The receiver and donor radii are

$$r_{m,b}^{\text{rec}} = \rho_b \eta_b \left(\frac{\bar{a}_m}{a_{\text{ref}}}\right)^{\gamma_b}, \qquad r_{mk,b}^{\text{don}} = \rho_b \left(\frac{\bar{a}_{d_{mk}}}{a_{\text{ref}}}\right)^{\gamma_b}, \tag{3}$$

where ρ_b is the base radius in native pixels, $\gamma_b > 0$ is the dimensionless expression-to-radius exponent, and $\eta_b > 0$ is the dimensionless receiver-to-donor radius multiplier. Radii and distances use the same pixel units. The dimensionless efficiency $\tau_b \in (0, 1)$ scales the overlapping fraction of donor area into transfer strength. Before candidate masking,

$$\widetilde{\alpha}_{mk}^{(b)} = \tau_b \frac{\text{Lens}(r_{m,b}^{\text{rec}}, r_{mk,b}^{\text{don}}, \delta_{mk})}{\pi (r_{mk,b}^{\text{don}})^2}, \qquad 0 < \tau_b < 1, \tag{4}$$

where Lens is the disk-intersection area. Candidate masking gives $\alpha_{mk}^{(b)} = c_{mk} \widetilde{\alpha}_{mk}^{(b)}$, without a row-mass cap. The resulting operator is

$$[\mathcal{A}_b(Z)]_{m:} = Z_{m:} + \sum_{k=1}^{K} \alpha_{mk}^{(b)} Z_{d_{mk}:} = [A_b Z]_{m:}, \qquad A_b = A_{\phi_b}(X, \mathcal{G}), \tag{5}$$

with $\phi_b = (\rho_b, \gamma_b, \eta_b, \tau_b)$. Appendix C.2 gives the overlap formula and numerical safeguards.

3.4 PGR Block

(a) Residual Backprojection (RB). The Forward Projection stage reconstructs $\widehat{X}_b = A_b H_{b-1}$ from the input state. Residual Computation then forms $R_b = P_\Omega X - \widehat{X}_b$, following the sign convention in Figure 2. Let W_b robustly downweight receiver rows poorly explained by the circle approximation, and let D_b be the lower-bounded diagonal of $A_b^\top A_b$. The back-projected increment is

$$\Delta_b^{\text{geo}} = \lambda_b \ell_b D_b^{-1} A_b^\top W_b R_b, \tag{6}$$

where $\lambda_b > 0$ and $\ell_b > 0$ are learned scalars. The adjoint distributes receiver residuals to candidate source nodes; diagonal normalization accounts for unequal participation in mixing relations. Robust weights reduce the influence of poorly fitted receiver rows, while the learned scalars control update magnitude.

(b) Adaptive Increment Modulation (AIM). State–Increment Encoding combines the current state and the RB increment. Each independently parameterized Block selects an eight-dimensional trainable vector $q_b = Q_b[b]$ that is constant across cells and scenes. Because it enters an affine map, q_b acts as a learned block-specific bias. Conditioned Gain Modulation, represented by m_{θ_b}, uses H_{b-1}, Δ_b^{geo}, and q_b to predict positive coordinate-wise gains,

$$M_b = m_{\theta_b}(H_{b-1}, \Delta_b^{\mathrm{geo}}, q_b) \in \mathbb{R}_{>0}^{N\times G}. \tag{7}$$

With the preceding displacement $V_{b-1} = H_{b-1} - H_{b-2}$ for $b > 1$ and the initialization $V_0 = 0$, the modulated geometry increment is

$$\widetilde{\Delta}_b^{\mathrm{geo}} = M_b \odot \Delta_b^{\mathrm{geo}} + \mu_b V_{b-1}, \qquad 0 < \mu_b < 1. \tag{8}$$

Positive gain rescales each coordinate of the RB increment without reversing its sign; momentum then adds the preceding displacement. Thus AIM comprises both learned gain modulation and momentum: removing the learned gain sets $M_b = \mathbf{1}$ while retaining the momentum term. Appendix C.3 specifies the positive gain mapping.

(c) Graph-Context Correction (GCC). The gene-separable circle operator approximates segmented footprints. State–Update Encoding forms node features from the current state and the AIM output. Two GATv2 layers (Brody et al., 2022) propagate these features over the observable graph, and a Correction Head produces the complementary increment:

$$\Delta_b^{\mathrm{graph}} = \mathcal{R}_{\psi_b}\left(H_{b-1}, \widetilde{\Delta}_b^{\mathrm{geo}}, \mathcal{G}\right). \tag{9}$$

The parameters ψ_b are Block-specific. Unlike the gene-shared geometric coefficients, GCC combines cross-gene features with neighboring states and proposed increments to learn profile-dependent corrections. All three modules read the same state H_{b-1}; only their increments pass successively between modules.

Proposal update. The Block makes exactly one explicit update of the recovery state:

$$H_b = \left[H_{b-1} + \widetilde{\Delta}_b^{\mathrm{geo}} + \Delta_b^{\mathrm{graph}}\right]_+. \tag{10}$$

Projection follows the combined physical and graph increments in every Block.

3.5 Stack Output and Training Objectives

After four Blocks, positive output scaling calibrates the total displacement from the observation:

$$\widehat{S} = [X + \kappa(H_L - X)]_+, \qquad \kappa > 0. \tag{11}$$

Source losses use the receiver rows selected by P_Ω. The Expression Loss combines E_{prof}, which compares separately library-normalized log profiles, and E_{raw}, which divides the absolute count error by the reference cell total (floored at one). The latter does not normalize prediction and reference separately. The complete objective is

$$\mathcal{L} = E_{\mathrm{prof}}(\widehat{S}, S; \Omega) + E_{\mathrm{raw}}(\widehat{S}, S; \Omega). \tag{12}$$

The four Block-specific circle operators have independent parameters. The evaluated configuration uses domain-only maps for raw coordinates $u_b \in \mathbb{R}^4$:

$$\rho_b = \mathrm{softplus}(u_{b,1}), \quad \gamma_b = \mathrm{softplus}(u_{b,2}),$$
$$\eta_b = \mathrm{softplus}(u_{b,3}), \quad \tau_b = \mathrm{sigmoid}(u_{b,4}).$$

Thus ρ_b, γ_b, and η_b are positive without finite upper bounds, while $\tau_b \in (0, 1)$. Section 4.5 reports the effect of removing each module from this stack. Appendix C.4 defines the expression errors, and Appendix D.1 gives training and inference details.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Baselines. We compare GATv2 (Brody et al., 2022), GraphSAGE (Hamilton et al., 2017), GAT (Veličković et al., 2018), GCN (Kipf & Welling, 2017), MPNN (Gilmer et al., 2017), VAE (Kingma & Welling, 2014), and Physical PGD with a train-fitted, frozen circle operator. Learned baselines predict nonnegative residual outputs from X; graph baselines also receive adjacency, while PhyGRec and PGD additionally use candidate distances and masks. Each neural baseline uses a 16-configuration validation search and three training seeds; PGD selects its step size and iteration count on validation. Inputs, capacities, and search settings are detailed in Appendix D.2.

Metrics. Count-MAE is mean absolute raw-count error. Log-MAE follows nonnegative clipping, library normalization to 10^4, and log1p. Rel. L_1 sums absolute count errors and divides by Reference library size (Appendix B); all three are lower-is-better. Errors use all cell–scene instances. Log-RMSE is additionally reported in the appendix. We report mean and sample SD over three seeds. LOCO averages fold metrics equally within each seed before seed aggregation. Checkpoints are selected using validation Log-MAE and Rel. L_1 (Appendix D).

Data and split. Single-chip A9/L7/NA PBMC profiles define Reference. Each chip pair uses an 11×11 translation grid; positive cross-chip overlaps add expression from up to eight donors, weighted by intersection over donor area. Receiver–scene rows require overlap. Models see a separate joint $K = 8$ coordinate graph, not true overlap edges or coefficients. The panel has 232 prespecified markers and 768 training-selected highly variable genes.

Before augmentation, approximately 64-cell coordinate components are assigned $60/20/20$ to train/validation/test; receiver cells and all candidate donor cells must belong to the same split. The split has 24,256/8,227/8,170 cells and 290,136/32,904/33,896 cell–scene instances (Appendix A.1).

4.2 SOURCE-EXPRESSION RECOVERY

Table 1 shows that PhyGRec attains the lowest mean error on all three expression metrics among the compared methods; Figure 3 resolves the improvements at the cell level. PhyGRec reduces Count-MAE from GATv2's 0.00343 to 0.00133, Log-MAE from 0.00888 to 0.00369, and Rel. L_1 from 0.10023 to 0.03968. The Log-MAE reduction relative to GATv2 is 58.4%.

The expression-error advantage holds across all three chip pairings (Appendix Table 9). Physical PGD, which uses a fitted geometric inverse without learned modulation or graph correction, has higher error on all three metrics.

4.3 PBMC REFERENCE AGREEMENT

On 6,394 matched cells, PhyGRec increases agreement with Reference-derived PBMC annotations and clusters, raising Macro F1 from GATv2's 0.888 to 0.947 and ARI from 0.686 to 0.760 (Table 1). Scores use fixed Reference-derived annotations and independent per-seed evaluation (Appendix B.2). Figure 4 pairs this matched-cell visualization with a separate exploratory diagnostic on an unseen chip; its protocols and marker evidence are in Appendix F.2.

Gene-level recovery. Figure 5 shows donor- and receiver-enriched marker errors in examples selected by the PhyGRec/baseline error ratio. Figure 3 reports full-cohort error distributions.

4.4 TRANSFER TO HELD-OUT RECEIVER CHIPS

Receiver-LOCO evaluates transfer to each held-out receiver chip within the matched Y40360 series. Each fold trains on the other two chips, excludes the held-out chip from validation and highly variable gene selection, and evaluates directions with that chip as receiver. Configurations transfer from main-split validation without LOCO retuning; fold validation selects checkpoints. Appendix A.3 gives the protocol.

Table 1: Test expression reconstruction (33,896 cell–scene instances) and annotation/clustering agreement (6,394 matched cells). Annotation labels are marker-derived from Reference expression. Entries show mean ± sample SD over three training seeds; best/second-best means are bold/underlined. Bal. acc.: macro recall; MCC: multiclass Matthews correlation coefficient. Metric definitions and aggregation are given in Section 4 and Appendix B.2.

Method	**Expression reconstruction**			**Annotation agreement**			**Clustering structure**	
	Count-MAE ↓	**Log-MAE** ↓	**Rel.** L_1 ↓	**Macro F1** ↑	**Bal. acc.** ↑	**MCC** ↑	**ARI** ↑	**NMI** ↑
GATv2	<u>0.00343</u> ±0.00015	<u>0.00888</u> ±0.00040	<u>0.10023</u> ±0.00133	<u>0.888</u> ±0.010	<u>0.889</u> ±0.009	<u>0.869</u> ±0.012	<u>0.686</u> ±0.021	<u>0.635</u> ±0.006
GraphSAGE	0.00394 ±0.00003	0.00952 ±0.00020	0.12215 ±0.00107	0.881 ±0.004	0.880 ±0.004	0.858 ±0.005	0.655 ±0.005	0.618 ±0.011
GAT	0.00547 ±0.00019	0.01729 ±0.00095	0.19381 ±0.00653	0.784 ±0.009	0.785 ±0.008	0.741 ±0.010	0.483 ±0.006	0.459 ±0.004
MPNN	0.00637 ±0.00022	0.01554 ±0.00104	0.22564 ±0.00521	0.806 ±0.020	0.805 ±0.021	0.766 ±0.026	0.494 ±0.049	0.477 ±0.035
Physical PGD	0.00393 ±0.00001	0.03056 ±0.00064	0.16925 ±0.00186	0.680 ±0.005	0.689 ±0.005	0.614 ±0.006	0.238 ±0.024	0.261 ±0.011
GCN	0.00664 ±0.00020	0.02143 ±0.00037	0.23407 ±0.00544	0.745 ±0.003	0.746 ±0.003	0.692 ±0.004	0.411 ±0.019	0.399 ±0.012
VAE	0.00649 ±0.00004	0.02460 ±0.00043	0.25340 ±0.00075	0.710 ±0.004	0.714 ±0.003	0.647 ±0.005	0.306 ±0.012	0.312 ±0.001
PhyGRec	**0.00133** ±0.00005	**0.00369** ±0.00021	**0.03968** ±0.00167	**0.947** ±0.004	**0.947** ±0.003	**0.937** ±0.004	**0.760** ±0.022	**0.719** ±0.009

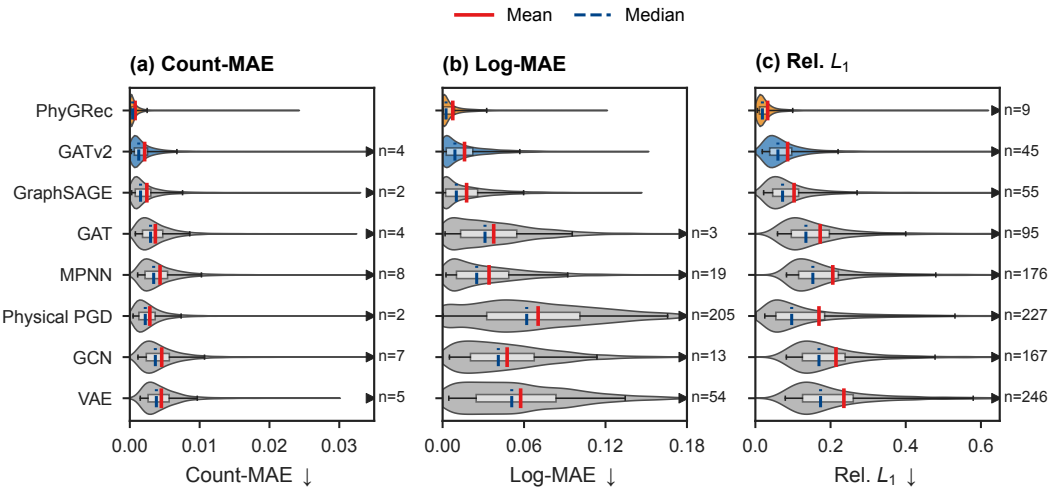

Figure 3: Cell-level reconstruction errors on the A9/L7/NA test split: (a) Count-MAE, (b) Log-MAE, and (c) Rel. L_1; lower is better. Errors for 6,402 cells are averaged across three seeds after within-seed raw-count prediction averaging by cell identity. Violins show densities within the displayed range; boxes and 5th–95th percentile whiskers summarize all cells. Red solid/blue dashed segments mark means/medians; right-edge triangles and counts indicate off-scale cells. PhyGRec is orange, GATv2 blue, and other baselines gray.

Table 2 reports fold-equal receiver-LOCO results. Relative to the next-best method for each expression metric, PhyGRec reduces Count-MAE from Physical PGD's 0.00396 to 0.00220, Log-MAE from GraphSAGE's 0.01305 to 0.00801, and Rel. L_1 from GraphSAGE's 0.16003 to 0.07055. Reference-agreement ARI rises from GraphSAGE's 0.561 to 0.654. Within this controlled three-chip benchmark, these improvements persist for held-out receiver chips without LOCO-specific hyperparameter tuning.

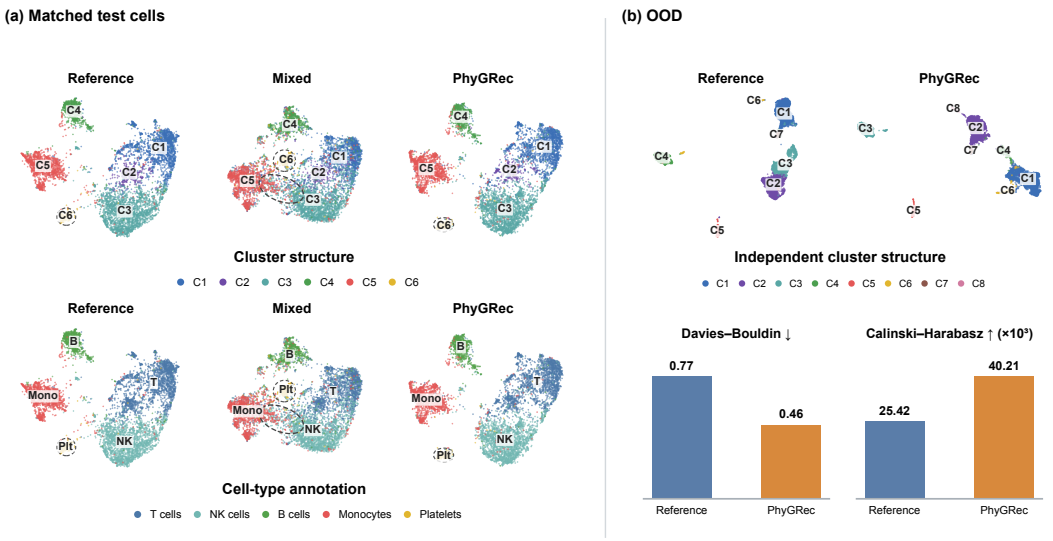

Figure 4: Matched-cell and out-of-distribution (OOD) PBMC structure. (a) The same 6,394 A9/L7/NA test cells in Reference, Mixed, and three-seed mean PhyGRec expression. Cluster structure is above cell-type annotation; colors and labels are fixed from Reference. Each state is embedded independently, then rigidly aligned to Reference with shared axis limits. Dashed outlines are visual guides to platelet cores and the mixed C5/C3 region. (b) Direct inference on the unseen Y40176KE chip without synthetic overlap (8,821 plotted cells). Reference and PhyGRec have independently assigned Leiden clusters; cluster IDs are state-specific. The bars compare cluster separation in the displayed two-dimensional UMAP: Davies–Bouldin index decreases from 0.770 to 0.463, and Calinski–Harabasz index increases from 25.42 to 40.21×10^3. Silhouette changes from 0.524 to 0.512; these unsupervised scores do not establish cell-type accuracy. Protocols: Appendix F.2.

Table 2: Strict receiver-LOCO over A9/L7/NA. Entries show mean ± sample SD across three seeds after equal-weight fold averaging; bold/underline mark best/second-best means. Metrics follow Table 1, with fold-specific Reference labels. Full results: Appendix E.2; protocol: Appendix B.3.

Method	Expression reconstruction			Annotation agreement			Clustering structure	
	Count-MAE ↓	Log-MAE ↓	Rel. L_1 ↓	Macro F1 ↑	Bal. acc. ↑	MCC ↑	ARI ↑	NMI ↑
GraphSAGE	0.00503 ±0.00004	<u>0.01305</u> ±0.00012	<u>0.16003</u> ±0.00070	<u>0.845</u> ±0.003	<u>0.846</u> ±0.003	<u>0.816</u> ±0.003	<u>0.561</u> ±0.016	<u>0.542</u> ±0.009
GATv2	0.00549 ±0.00017	0.01711 ±0.00061	0.16823 ±0.00640	0.811 ±0.005	0.811 ±0.004	0.775 ±0.005	0.518 ±0.016	0.494 ±0.008
Physical PGD	<u>0.00396</u> ±0.00002	0.03032 ±0.00093	0.16848 ±0.00258	0.685 ±0.006	0.694 ±0.005	0.620 ±0.007	0.200 ±0.011	0.252 ±0.011
GAT	0.00646 ±0.00033	0.02140 ±0.00104	0.23098 ±0.00636	0.753 ±0.014	0.755 ±0.015	0.702 ±0.019	0.372 ±0.052	0.379 ±0.032
MPNN	0.00747 ±0.00023	0.01935 ±0.00162	0.26407 ±0.01143	0.792 ±0.002	0.797 ±0.002	0.752 ±0.003	0.403 ±0.028	0.408 ±0.021
GCN	0.00715 ±0.00021	0.02404 ±0.00064	0.25371 ±0.00343	0.721 ±0.005	0.724 ±0.005	0.663 ±0.006	0.346 ±0.022	0.350 ±0.013
VAE	0.00659 ±0.00011	0.02484 ±0.00059	0.25466 ±0.00244	0.704 ±0.003	0.707 ±0.003	0.640 ±0.003	0.250 ±0.043	0.277 ±0.023
PhyGRec	**0.00220** ±0.00023	**0.00801** ±0.00065	**0.07055** ±0.00739	**0.900** ±0.007	**0.899** ±0.008	**0.882** ±0.008	**0.654** ±0.014	**0.635** ±0.010

4.5 Component Analysis

Table 3 compares validation-selected checkpoints after retraining without RB, AIM, or GCC. Each removal worsens all three main expression errors and Reference-agreement ARI and NMI; omitting GCC raises Log-MAE from 0.00369 to 0.01075. Complete metrics are in Appendix E.3.

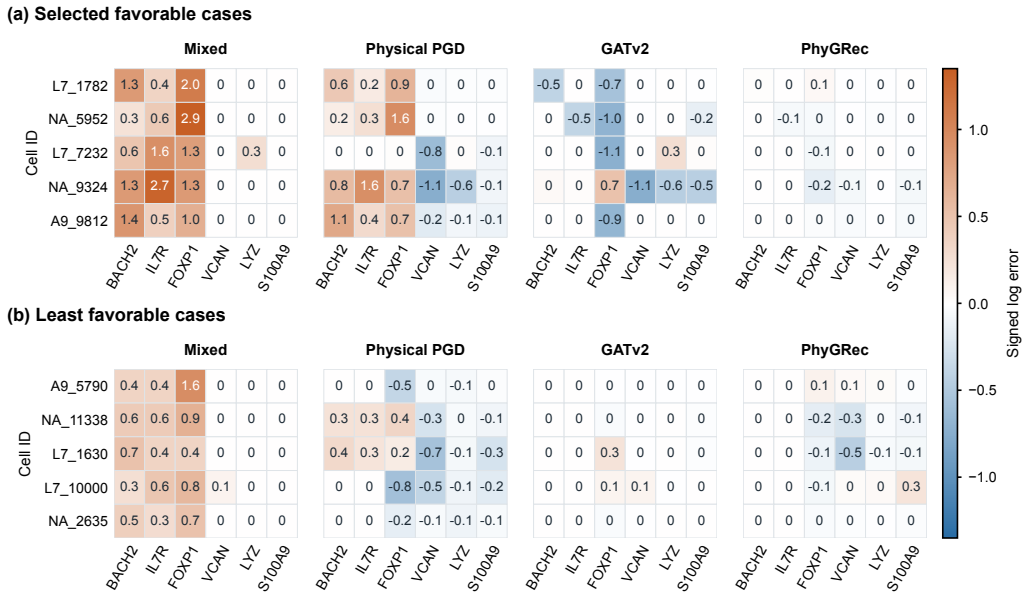

Figure 5: Gene-level errors in ten post-hoc selected cell–scene examples: (a) five favorable cases and (b) five high PhyGRec/baseline error-ratio cases. Numbers show raw-count errors $\Delta = \widehat{S} - S$ relative to Reference; colors show $\mathrm{sign}(\Delta)\log(1+|\Delta|)$ on a shared, zero-centered scale. Selection rules and Reference expression: Appendix F.1.

Table 3: Fixed-split component removals, using Table 1 metrics and evaluation cells. Params are active trainable parameter counts in millions, rounded to three decimals. Metric cells show mean above ± sample SD across three seeds; best/second-best means are bold/underlined. Full metrics: Appendix E.3.

Variant	Params (M)	Expression reconstruction			Annotation agreement			Clustering structure	
		Count-MAE ↓	Log-MAE ↓	Rel. L_1 ↓	Macro F1 ↑	Bal. acc. ↑	MCC ↑	ARI ↑	NMI ↑
w/o RB	2.833	0.00226 ±0.00034	0.00657 ±0.00148	0.06582 ±0.01092	0.915 ±0.019	0.917 ±0.019	0.901 ±0.023	0.712 ±0.020	0.671 ±0.019
w/o AIM	2.620	<u>0.00163</u> ±0.00006	<u>0.00453</u> ±0.00007	<u>0.04637</u> ±0.00136	<u>0.945</u> ±0.004	<u>0.945</u> ±0.004	<u>0.934</u> ±0.005	<u>0.748</u> ±0.028	<u>0.709</u> ±0.014
w/o GCC	0.212	0.00261 ±0.00013	0.01075 ±0.00013	0.09121 ±0.00421	0.875 ±0.005	0.873 ±0.005	0.853 ±0.005	0.613 ±0.049	0.595 ±0.024
PhyGRec	2.833	**0.00133** ±0.00005	**0.00369** ±0.00021	**0.03968** ±0.00167	**0.947** ±0.004	**0.947** ±0.003	**0.937** ±0.004	**0.760** ±0.022	**0.719** ±0.009

Removing GCC produces the largest observed error increase while reducing active parameters from 2.833M to 0.212M. Removing RB retains model size, whereas removing AIM reduces it to 2.620M. All nine paired-seed Log-MAE differences favor the complete model (Appendix E.3). These comparisons measure each complete branch's contribution within the evaluated stack.

5 Conclusion

Under controlled geometric crosstalk, PhyGRec combines a circle-overlap forward surrogate with residual backprojection, adaptive increment modulation, and graph-context correction. Across three Stereo-cell PBMC chips, it achieves the lowest Count-MAE, Log-MAE, and Rel. L_1 among evaluated methods on both the identity-disjoint fixed split and strict receiver-LOCO. Its outputs better preserve Reference-derived annotations and clusters; removing any of the three components increases expression error.

AI USE STATEMENT

AI assistance was used in drafting this paper. All analyses, editorial judgments, and conclusions are the authors' own.

REPRODUCIBILITY STATEMENT

The benchmark construction and data splits, evaluation procedures, model implementation, training and baseline protocols, and detailed numerical results are documented in Appendices A–E. An anonymous code repository is linked on the first page.

REFERENCES

Tamim Abdelaal, Soufiane Mourragui, Ahmed Mahfouz, and Marcel J. T. Reinders. SpaGE: Spatial gene enhancement using scRNA-seq. *Nucleic Acids Research*, 48(18):e107, 2020. doi: 10.1093/nar/gkaa740.

Jonas Adler and Ozan Öktem. Learned primal-dual reconstruction. *IEEE Transactions on Medical Imaging*, 37(6):1322–1332, 2018. doi: 10.1109/TMI.2018.2799231.

Alma Andersson, Joseph Bergenstråhle, Michaela Asp, Ludvig Bergenstråhle, Aleksandra Jurek, José Fernández Navarro, and Joakim Lundeberg. Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. *Communications Biology*, 3:565, 2020. doi: 10.1038/s42003-020-01247-y.

Jonathan T. Barron. A general and adaptive robust loss function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4331–4339, 2019. doi: 10.1109/CVPR.2019.00446.

Dimitri P. Bertsekas. On the goldstein–levitin–polyak gradient projection method. *IEEE Transactions on Automatic Control*, 21(2):174–184, 1976. doi: 10.1109/TAC.1976.1101194.

Tommaso Biancalani, Gabriele Scalia, Lorenzo Buffoni, Raghav Avasthi, Ziqing Lu, Aman Sanger, Neriman Tokcan, Charles R. Vanderburg, Åsa Segerstolpe, Meng Zhang, et al. Deep learning and alignment of spatially resolved single-cell transcriptomes with tangram. *Nature Methods*, 18(11):1352–1362, 2021. doi: 10.1038/s41592-021-01264-7.

Dario Bressan, Giorgia Battistoni, and Gregory J. Hannon. The dawn of spatial omics. *Science*, 381(6657):eabq4964, 2023. doi: 10.1126/science.abq4964.

Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? In *International Conference on Learning Representations (ICLR)*, 2022.

Dylan M. Cable, Evan Murray, Luli S. Zou, Aleksandrina Goeva, Evan Z. Macosko, Fei Chen, and Rafael A. Irizarry. Robust decomposition of cell type mixtures in spatial transcriptomics. *Nature Biotechnology*, 40(4):517–526, 2022. doi: 10.1038/s41587-021-00830-w.

Ao Chen, Sha Liao, Mengnan Cheng, Kailong Ma, Liang Wu, Yiwei Lai, Xiaojie Qiu, Jin Yang, Jiangshan Xu, Shijie Hao, et al. Spatiotemporal transcriptomic atlas of mouse organogenesis using dna nanoball-patterned arrays. *Cell*, 185(10):1777–1792.e21, 2022. doi: 10.1016/j.cell.2022.04.003.

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*, 10:390, 2019. doi: 10.1038/s41467-018-07931-2.

Stephen J. Fleming, Mark D. Chaffin, Alessandro Arduini, Amer-Denis Akkad, Eric Banks, John C. Marioni, Anthony A. Philippakis, Patrick T. Ellinor, and Mehrtash Babadi. Unsupervised removal of systematic background noise from droplet-based single-cell experiments using CellBender. *Nature Methods*, 20(9):1323–1335, 2023. doi: 10.1038/s41592-023-01943-7.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1263–1272, 2017.

Karol Gregor and Yann LeCun. Learning fast approximations of sparse coding. In *Proceedings of the 27th International Conference on Machine Learning*, pp. 399–406, 2010.

William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Jian Hu, Xiangjie Li, Kyle Coleman, Amelia Schroeder, Nan Ma, David J. Irwin, Edward B. Lee, Russell T. Shinohara, and Mingyao Li. SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nature Methods*, 18(11):1342–1351, 2021. doi: 10.1038/s41592-021-01255-8.

Mo Huang, Jingshu Wang, Eduardo Torre, Hannah Dueck, Sydney Shaffer, Roberto Bonasio, John I. Murray, Arjun Raj, Mingyao Li, and Nancy R. Zhang. SAVER: gene expression recovery for single-cell RNA sequencing. *Nature Methods*, 15(7):539–542, 2018. doi: 10.1038/s41592-018-0033-z.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

Vitalii Kleshchevnikov, Artem Shmatko, Emma Dann, Alexander Aivazidis, Hamish W. King, Tong Li, Rasa Elmentaite, Artem Lomakin, Veronika Kedlian, Adam Gayoso, et al. Cell2location maps fine-grained cell types in spatial transcriptomics. *Nature Biotechnology*, 40(5):661–671, 2022. doi: 10.1038/s41587-021-01139-4.

Sha Liao, Xiaoxi Zhou, Chuanyu Liu, Chang Liu, Shijie Hao, Hongyu Luo, Huan Hou, Qian Liu, Zhe Zhang, Liyun Xiao, et al. Stereo-cell: Spatial enhanced-resolution single-cell sequencing with high-density DNA nanoball-patterned arrays. *Science*, 389(6762):eadr0475, 2025. doi: 10.1126/science.adr0475.

Yahui Long, Kok Siong Ang, Mengwei Li, Kian Long Kelvin Chong, Raman Sethi, Chengwei Zhong, Hang Xu, Zhiwei Ong, Karishma Sachaphibulkij, Ao Chen, et al. Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with GraphST. *Nature Communications*, 14:1155, 2023. doi: 10.1038/s41467-023-36796-3.

Vishal Monga, Yuelong Li, and Yonina C. Eldar. Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing. *IEEE Signal Processing Magazine*, 38(2):18–44, 2021. doi: 10.1109/MSP.2020.3016905.

Zijian Ni, Aman Prasad, Shuyang Chen, Richard B. Halberg, Lisa M. Arkin, Beth A. Drolet, Michael A. Newton, and Christina Kendziorski. Spotclean adjusts for spot swapping in spatial transcriptomics data. *Nature Communications*, 13(1):2971, 2022. doi: 10.1038/s41467-022-30587-y.

Gregory Ongie, Ajil Jalal, Christopher A. Metzler, Richard G. Baraniuk, Alexandros G. Dimakis, and Rebecca Willett. Deep learning techniques for inverse problems in imaging. *IEEE Journal on Selected Areas in Information Theory*, 1(1):39–56, 2020. doi: 10.1109/JSAIT.2020.2991563.

David van Dijk, Roshan Sharma, Juozas Nainys, Kristina Yim, Pooja Kathail, Ambrose J. Carr, Cassandra Burdziak, Kevin R. Moon, Christine L. Chaffer, Diwakar Pattabiraman, et al. Recovering gene interactions from single-cell data using data diffusion. *Cell*, 174(3):716–729.e27, 2018. doi: 10.1016/j.cell.2018.05.061.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Xiaoyu Wei, Sulei Fu, Hanbo Li, Yang Liu, Shuai Wang, Weimin Feng, Yunzhi Yang, Xiawei Liu, Yan-Yun Zeng, Mengnan Cheng, et al. Single-cell Stereo-seq reveals induced progenitor cells involved in axolotl brain regeneration. *Science*, 377(6610), 2022. doi: 10.1126/science.abp9444.

Hang Xu, Huazhu Fu, Yahui Long, Kok Siong Ang, Raman Sethi, Kelvin Chong, Mengwei Li, Rom Uddamvathanak, Hong Kai Lee, Jingjing Ling, et al. Unsupervised spatially embedded deep representation of spatial transcriptomics. *Genome Medicine*, 16:12, 2024. doi: 10.1186/s13073-024-01283-x.

Shiyi Yang, Sean E. Corbett, Yusuke Koga, Zhe Wang, W. Evan Johnson, Masanao Yajima, and Joshua D. Campbell. Decontamination of ambient RNA in single-cell RNA-seq with DecontX. *Genome Biology*, 21(1):57, 2020. doi: 10.1186/s13059-020-1950-6.

Matthew D. Young and Sam Behjati. SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *GigaScience*, 9(12):giaa151, 2020. doi: 10.1093/gigascience/giaa151.

Jian Zhang and Bernard Ghanem. Ista-net: Interpretable optimization-inspired deep network for image compressive sensing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1828–1837, 2018. doi: 10.1109/CVPR.2018.00196.

Edward Zhao, Matthew R. Stone, Xing Ren, Jamie Guenthoer, Kimberly S. Smythe, Thomas Pulliam, Stephen R. Williams, Cedric R. Uytingco, Sarah E. B. Taylor, Paul Nghiem, Jason H. Bielas, and Raphael Gottardo. Spatial transcriptomics at subspot resolution with BayesSpace. *Nature Biotechnology*, 39(11):1375–1384, 2021. doi: 10.1038/s41587-021-00935-2.

CONTENTS

A Benchmark Construction and Data Splits **14**

A.1 Experimental Setup and Benchmark Construction 14

A.2 Information Available to Each Stage 15

A.3 Strict Receiver-LOCO Protocol 15

B Evaluation Units, Metrics, and Reference Labels **16**

B.1 Evaluation Units 16

B.2 Annotation and Clustering Protocol 16

B.3 Receiver-LOCO and Component Scoring 17

C Model Implementation and Inference **18**

C.1 Notation and Tensor Dimensions 18

C.2 Circle-Overlap Operator 18

C.3 PGR Block Details 18

C.4 Training-Loss Definitions 19

C.5 Inference and Computational Complexity 19

D Training and Comparison Protocol **20**

D.1 Training Configuration and Ablation Protocol 20

D.2 Baselines 20

D.3 Training and Validation Dynamics 21

E Expression and Structure Results **22**

E.1 Fixed-Split Annotation and Clustering 24

E.2 Held-Out Receiver Results 24

E.3 Component Contributions 27

F Cell-Level Examples and PBMC Visualizations **27**

F.1 Qualitative Recovery Examples 27

F.2 PBMC Illustrations 28

Table 4: Technical properties of the three Stereo-cell PBMC chips. Genes and UMI denote the number of detected genes and total UMI counts per chip, respectively; per-cell values are medians.

Chip	Cells	Genes	UMI	Median Genes/Cell	Median UMI/Cell
Y40360A9	13,137	24,447	16,826,949	709	953
Y40360L7	13,870	24,025	15,049,856	623	816
Y40360NA	13,646	24,886	19,264,203	738	1,000

A Benchmark Construction and Data Splits

A.1 Experimental Setup and Benchmark Construction

Stereo-cell PBMC assay context. Stereo-cell profiles dissociated cells on an open DNB capture surface, where limited lateral RNA diffusion can cross segmented cell regions (Liao et al., 2025). The benchmark models expression mixing through controlled footprint overlap.

The three benchmark chips were selected from 19 candidates using six technical features: cell count, detected-gene count, median UMI per cell, median detected genes per cell, spatial bounding-box area, and aspect ratio. Positive features were log-transformed and robustly standardized by the 19-chip median and interquartile range, and the triple with minimum mean pairwise distance was selected. Table 4 reports the principal statistics of the resulting A9/L7/NA set. All three belong to the Y40360 series and contain PBMC profiles. They are technically matched assay sources, and their measured single-chip profiles define the benchmark reference.

Controlled overlap-mixture generator. Each mixing scene aligns a chip pair and translates one chip's segmented footprints relative to the other. A receiver's observed expression includes contributions only from opposite-chip donors whose translated polygons intersect its footprint; receiver and donor are directional roles. A separately constructed coordinate graph supplies the model-visible neighborhood.

Let P_n be the segmented chip footprint of source node n after applying a mixing scene translation. For each receiver cell m and overlapping donor cell n, the generator uses the directional donor coefficient

$$\beta^{\star}_{mn} = \frac{|P_m \cap P_n|}{\max(|P_n|, 1)}, \qquad X_{m:} = S_{m:} + \sum_{n \in \mathcal{D}^{\star}(m)} \beta^{\star}_{mn} S_{n:}, \tag{13}$$

where $|\cdot|$ denotes segmented area in pixels and $\mathcal{D}^{\star}(m)$ contains at most the eight donor cells with largest directional overlap. Only receiver–scene rows with at least one positive polygon-overlap donor are retained. The mixture is a deterministic continuous combination of measured benchmark-reference profiles; no molecular resampling is applied. Hidden overlap edges and coefficients generate X but are removed from model-visible inputs.

The foreground bounding-box centers are aligned before applying the Cartesian translation grid $\{-400, -320, \ldots, 400\}^2$ in native pixel coordinates, yielding 121 placements per pairing. Each placement remains a separate graph. After each translation, the observable graph is reconstructed as a symmetrized joint spatial $K = 8$ nearest-neighbor graph over both chips. Same-chip and cross-chip neighbors follow the same rule, so the graph does not reveal which chip supplied a simulated donor. The candidate set recalls 99.987% of hidden overlap edges in training and 100.000% in validation and test. This high recall makes coefficient estimation and source recovery, rather than candidate enumeration, the principal modeled challenge.

Cells are assigned to training, validation, and test splits before chip translation. For each chip, a symmetric spatial $K = 8$ graph is partitioned into contiguous components of approximately 64 cells and assigned $60/20/20$ with seed 20260816. Generated relations retain receiver and donor cells from the same split. Table 5 counts the distinct cells that appear as receivers in these experiments, with each cell counted once across all mixing scenes.

Each split contains 121 translations per chip pairing and 363 mixing-scene graphs. Repeated appearances are separate prediction instances linked by cell identity for cell-level aggregation.

Table 5: Cell counts by chip and split. Each cell is counted once across all mixing scenes.

Split	Chip cells			Total cells
	A9	L7	NA	
Training	7,776	8,125	8,080	23,981
Validation	2,197	1,980	1,877	6,054
Test	2,033	2,128	2,241	6,402

After symbol normalization and intersection, 21,810 true gene names are shared across the three chips. The panel contains all 232 available genes from the curated PBMC marker list and 768 nonmarker genes selected from the 24,256 training cells. Training-only means and unbiased variances feed a Cell Ranger-style normalized-dispersion ranking; gene-name order resolves score ties. Marker genes are deduplicated before the ranked nonmarker fill, and the selected 1,000 genes retain their ranking order. Validation and test expression do not enter this ranking. On test, uncorrected Mixed has Log-MAE 0.05117 and Rel. L_1 0.30119 against the Reference receiver profiles.

A.2 Information Available to Each Stage

The single-chip measurements define the Reference target, and the generator adds known geometric contamination. Recovery is scored against the measured single-chip state; its original measurement noise and background are part of the Reference. Reference expression supplies training targets and validation checkpoint selection, while test Reference profiles and their derived annotations and clusters are reserved for scoring; none are prediction inputs. Hidden overlap edges and coefficients define injected transfer but are unavailable to the predictor; they enter only generator diagnostics and the stated case-selection rules.

All methods receive mixed expression X. Graph methods and Physical PGD also receive the coordinate-derived adjacency rebuilt for each mixing scene; PhyGRec and Physical PGD additionally use observable candidate distances and masks. The mixed input supports input-error comparisons, while the rebuilt graph supports candidate-edge diagnostics. Reference-derived labels are used only for annotation and cluster agreement.

The five marker programs used for scoring (Appendix B.2.1) are distinct from the larger 232-marker list used to construct the 1,000-gene model panel.

A.3 Strict Receiver-LOCO Protocol

Receiver-LOCO separates cross-chip transfer from the fixed identity split. In each fold, the held-out chip contributes no training or validation receiver, no node in the training combination, and no expression value to highly variable gene selection. The same external marker list is used in all folds, and the remaining genes in each 1,000-gene panel are selected from the two training chips only. Consequently, the three fold-specific panels need not contain identical highly variable genes. Architectures, learning rates, loss weights, batch settings, and training length transfer from fixed-split validation selection; only the checkpoint epoch is selected on the fold's in-distribution validation set.

Each fold contains 121 training and 121 validation mixing-scene graphs from the single training-chip pairing. The primary score uses the 242 scenes in the two directions for which the held-out chip is the receiver. A9 held out trains on L7–NA and evaluates A9←L7 and A9←NA; L7 held out trains on A9–NA and evaluates L7←A9 and L7←NA; NA held out trains on A9–L7 and evaluates NA←A9 and NA←L7.

The primary tests contain 11,607, 11,376, and 10,913 cell–scene instances and 2,033, 2,128, and 2,241 unique receiver cells for A9, L7, and NA, respectively. Reference identities and all translated receiver and donor appearances remain within one split.

B Evaluation Units, Metrics, and Reference Labels

The tables distinguish receiver–scene expression errors from physical-cell annotation and clustering scores. The nonlinear transforms and the order of scene, fold, and seed aggregation are fixed below.

B.1 Evaluation Units

The main-table expression errors use all 33,896 test receiver–scene instances and genes. Cell-error distributions average raw predictions by identity for 6,402 receiver cells before scoring; annotation and clustering score the 6,394 cells with nonzero Reference expression. The eight zero-Reference cells remain in expression-error scoring. Cell-consensus and instance-level errors can differ because scene averaging precedes nonlinear normalization and scoring. Receiver-LOCO scores each held-out-chip fold on its own gene panel, then averages folds equally within each training seed. Except for descriptive figure summaries, learned-method metrics are computed per seed before reporting mean and sample SD.

Count-MAE averages absolute count errors over instances and genes in stored expression units. Log-MAE equals E_{prof} in Equation 25, and Rel. L_1 equals E_{raw} in Equation 26. Count-RMSE takes the square root after averaging squared raw-count errors. Log-RMSE analogously takes the square root after averaging squared differences between transformed profiles $T(\widehat{S})$ and $T(S)$. Rel. L_2 averages $\|\widehat{S}_i - S_i\|_2 / \max(\|S_i\|_2, 1)$ over receiver–scene instances. RMSE is calculated within each split or fold before seed and, for LOCO, fold averaging. All expression errors are lower-is-better.

B.2 Annotation and Clustering Protocol

Annotation and clustering use the same 6,394 matched cell identities for Reference, Mixed, PhyGRec, and the baselines. For each trained method and seed, cell–scene predictions are averaged by identity before scoring. The PBMC illustrations in Appendix F.2 instead take a scene median within each seed, then average the three seed profiles for visualization.

Table 10 measures per-cell annotation preservation relative to Reference. We first apply a fixed marker annotator A to Reference expression alone. Each cell is library-normalized to 10^4 and log-transformed, after which gene-wise means and standard deviations are fitted once on Reference. Standardized values are clipped to the fixed interval $[-10, 10]$. For each of five fixed canonical programs (T cells, NK cells, B cells, monocytes, and platelets), the class score is the mean standardized expression over its marker set $\mathcal{M}_k$,

$$s_{ik}^{\mathrm{Reference}} = \frac{1}{|\mathcal{M}_k|} \sum_{g \in \mathcal{M}_k} z_{ig}^{\mathrm{Reference}}, \qquad y_i^{\mathrm{Reference}} = \arg\max_k s_{ik}^{\mathrm{Reference}}. \tag{14}$$

The resulting Reference per-cell annotations are frozen as the sole reference. Mixed and every method–seed output are independently processed with the same normalization, Reference-fitted scaler, clipping interval, marker sets, score definition, and argmax rule; no state-specific refitting, thresholding, or marker selection is permitted. The confusion matrix for expression state q is

$$C_{ab}^{(q)} = \sum_{i=1}^{6394} \mathbf{1}\left[y_i^{\mathrm{Reference}} = a \ \wedge \ A(X_i^{(q)}) = b \right], \tag{15}$$

so rows are frozen Reference per-cell annotations and columns are annotations inferred from state q for the same cells. Reference therefore has unit agreement for all hard-label metrics by definition.

We report Agreement Accuracy, Macro Precision, Balanced Accuracy, Macro F1, multiclass MCC, Cohen's κ, and Macro AP. Balanced Accuracy is the macro-average of class recall. Macro AP uses the five continuous marker scores after softmax normalization: the frozen Reference annotations are one-hot encoded, one-versus-rest average precision is computed for each class, and the five values are averaged. Reference Macro AP is not reported because the Reference scores themselves define the frozen labels. Mixed is deterministic and has one point estimate.

Fixed-split clustering. Reference cluster labels are frozen on the shared 6,394-cell cohort. Each expression state independently fits a standardized 20-PC representation and a 30-neighbor cosine

Table 6: Marker programs used for Reference-derived annotation scoring. Programs are fixed across methods and fully present in each LOCO panel; FCGR3A intentionally belongs to both NK-cell and monocyte programs.

Program	Genes
T cells	CD3D, CD4, CD8A, CD8B, IL7R, CCR7, TCF7, LEF1
NK cells	NKG7, GNLY, KLRD1, NCR1, KLRF1, FCGR3A, XCL1
B cells	CD79A, CD19, MS4A1, CD74, CD27, TCL1A
Monocytes	LYZ, FCN1, CTSD, CD14, S100A8, S100A9, S100A12, VCAN, FCGR3A
Platelets	PPBP, PF4, GP9, TUBB1

graph. Leiden resolution is 0.35, with initializations 20260816, 20260817, and 20260818. ARI and NMI average agreement with the frozen Reference partition over these initializations; stability averages the three pairwise ARIs. Frozen-label silhouette is measured in the recovered PC space using at most 4,000 cells selected with a fixed initialization. These initialization averages are formed within each training seed.

B.2.1 Fixed Marker Programs

B.3 Receiver-LOCO and Component Scoring

Tables 2 and 3 use the same marker programs and clustering hyperparameters as Table 1. The fixed-split component comparison uses the same 6,394 matched cells, gene panel, and frozen Reference partition as Table 1.

For receiver-LOCO, each fold retains its own training-selected 1,000-gene panel and only held-out receiver identities. Cells with zero Reference library size are excluded from structure scoring by a Reference-only rule: six A9 cells, one L7 cell, and one NA cell, leaving 2,027, 2,127, and 2,240 cells, respectively. All methods in a fold use exactly these identities. The complete five marker programs remain available in every fold (8 T-cell, 7 NK-cell, 6 B-cell, 9 monocyte, and 4 platelet markers). A gene-wise scaler and the five-class marker annotations are fitted once to each fold's Reference expression and frozen for scoring all methods and seeds. Held-out Reference profiles enter the scoring procedure after prediction; training and checkpoint selection use the two training chips.

Each LOCO fold freezes a Reference cluster partition using the fixed-split pipeline (20 PCs, 30 cosine neighbors, Leiden resolution 0.35). Predicted states use the same pipeline with three Leiden initializations (20260816–20260818); scores use the fold's Reference labels and scaler.

Fold and seed summary. For a metric q, let $q_{f,s}$ be the fold-specific metric computed on its evaluation unit for fold $f \in \{\mathrm{A9}, \mathrm{L7}, \mathrm{NA}\}$ and seed s. We compute $\bar{q}_s = \frac{1}{3}\sum_f q_{f,s}$ and report the mean and sample standard deviation of $\{\bar{q}_s\}_{s=1}^{3}$. This aggregation gives each held-out chip equal weight while preserving the per-fold definition of each error metric. The error bars describe algorithmic-seed variation after fold averaging.

Scoring environment and scope. Scoring used Python 3.13.14, PyTorch 2.6.0+cu124, NumPy 2.4.6, SciPy 1.18.0, scikit-learn 1.9.0, Scanpy 1.12.3, and leidenalg 0.12.0. The benchmark evaluates controlled geometric crosstalk in measured PBMC profiles from one technical series; fixed-split and receiver-LOCO findings concern within-series source recovery. Recovery on naturally overlapping cells without paired Reference profiles is outside this evaluation. Reported standard deviations describe training-seed or fitted-operator variation, rather than variation across independent biological samples. Learned circle parameters are predictive surrogates, not calibrated molecular-transfer estimates.

Table 7: Mixing-scene notation. Evaluation concatenates receiver rows over mixing scenes; repeated cell identities remain distinct cell–scene instances until explicitly aggregated.

Symbol	Meaning
N, G, L	Scene nodes, panel genes ($G = 1{,}000$), and Blocks ($L = 4$)
X, S, H_b	$N \times G$ mixed, Reference, and recovered-state matrices
Ω, P_Ω	Supervised/evaluated receiver nodes and their row selector
A_b, D_b, W_b	$\lvert\Omega\rvert \times N$ mixing operator; $N \times N$ normalization; $\lvert\Omega\rvert \times \lvert\Omega\rvert$ robust weights
$d_{mk}, c_{mk}, \delta_{mk}$	Candidate index, validity mask, and endpoint distance
$\mathcal{E}, E_g$	Bidirectional message-passing edges and their count
E_c	Number of active sparse candidate-operator edges
M	Number of cell–scene instances in a scored split or fold

C Model Implementation and Inference

C.1 Notation and Tensor Dimensions

C.2 Circle-Overlap Operator

Disk overlap and candidate masking. For disks with radii $r_1, r_2 > 0$ and center distance d, the overlap area is zero when $d \geq r_1 + r_2$ and equals $\pi \min(r_1, r_2)^2$ when $d \leq |r_1 - r_2|$. In the partial-overlap regime,

$$\begin{aligned}\mathrm{Lens}(r_1, r_2, d) = r_1^2 \cos^{-1}\left(\frac{d^2 + r_1^2 - r_2^2}{2dr_1}\right) + r_2^2 \cos^{-1}\left(\frac{d^2 + r_2^2 - r_1^2}{2dr_2}\right) \\ - \frac{1}{2}\sqrt{(-d + r_1 + r_2)(d + r_1 - r_2)(d - r_1 + r_2)(d + r_1 + r_2)}. \end{aligned} \tag{16}$$

Equation 4 divides this area by the donor area before multiplying by transfer efficiency. Candidate masks give

$$\alpha_{mk}^{(b)} = c_{mk} \widetilde{\alpha}_{mk}^{(b)}. \tag{17}$$

Each transfer coefficient is nonnegative; the evaluated configuration does not cap their sum over donor candidates.

C.3 PGR Block Details

Residual Backprojection (RB): robust adjoint normalization. For residual $R_b = P_\Omega X - A_b H_{b-1}$, define the relative receiver error $r_{m,b} = \|R_{b,m:}\|_1 / \max(\|X_{m:}\|_1, 1)$. We use the Barron IRLS weight normalized by its zero-residual value $1/\sigma_b^2$ (Barron, 2019):

$$w_{m,b} = \left(1 + \frac{(r_{m,b}/\sigma_b)^2}{|\nu_b - 2|}\right)^{\nu_b/2-1}, \qquad \sigma_b > 0, \quad \nu_b < 2, \tag{18}$$

with $W_b = \mathrm{diag}(w_{1,b}, \ldots, w_{|\Omega|,b})$. Unequal node incidence is normalized by

$$(D_b)_{nn} = \max\left(\sum_{m \in \Omega} (A_b)_{mn}^2, 10^{-6}\right). \tag{19}$$

Adaptive Increment Modulation (AIM). Each Block owns a trainable table $Q_b \in \mathbb{R}^{L \times 8}$ and selects its one-based row $q_b = Q_b[b]$. State and geometry increment are RMS-normalized independently and summed,

$$U_b = \mathrm{RMSNorm}_{H,b}(H_{b-1}) + \mathrm{RMSNorm}_{\Delta,b}(\Delta_b^{\mathrm{geo}}). \tag{20}$$

An affine map sends U_b to 32 channels. A 16-channel transform of q_b is concatenated with this representation, followed by a joint affine projection. Splitting the result into equal value and gate parts gives

$$h_b = \mathrm{SiLU}(h_{b,\mathrm{val}}) \odot h_{b,\mathrm{gate}}. \tag{21}$$

A zero-initialized output head gives raw coordinates Z_b, and $M_b = \mathrm{softplus}(Z_b + \mathrm{softplus}^{-1}(1))$. The Block therefore starts with unit gain and learns positive values without a finite upper bound. The learned q_b is fixed within its Block; scene-dependent gain comes from U_b.

Graph-Context Correction (GCC). For node n, let $s_{n,b} = \max(\mathrm{sg}[G^{-1}\sum_g (H_{b-1})_{ng}], 10^{-6})$ be the detached mean-count scale, where sg is stop-gradient. The node feature concatenates a log-compressed state and a signed, scale-normalized modulated increment:

$$f_{n,b} = \mathrm{LN}_{H,b}(\log(1 + [(H_{b-1})_{n:}]_+)) \parallel \mathrm{LN}_{\Delta,b}\left(\mathrm{asinh}\frac{\widetilde{\Delta}^{\mathrm{geo}}_{b,n:}}{s_{n,b}}\right). \tag{22}$$

The logarithm clips features only, leaving the stored state unchanged for the next residual calculation and AIM. The mean-count scale is floored at 10^{-6}, and the signed update uses asinh. An affine input projection followed by two width-128 residual GATv2 layers produces $Z_b^{(2)}$ over the bidirectional edge set $\mathcal{E}$. Each layer uses four attention heads, zero dropout, self-loops, LayerNorm, and a GELU residual update. Equation 9 is instantiated as

$$\Delta^{\mathrm{graph}}_{b,n:} = \chi_b s_{n,b}\,\mathrm{Aff}_{\mathrm{out},b}(Z^{(2)}_{b,n}), \qquad \chi_b > 0. \tag{23}$$

C.4 Training-Loss Definitions

For a row vector y, define the nonnegative library-normalized transform

$$T(y) = \log\left(1 + 10^4 \frac{[y]_+}{\max(\|[y]_+\|_1, 10^{-6})}\right). \tag{24}$$

The Expression Loss comprises the following profile and reference-total-normalized count errors on supervised receiver nodes Ω:

$$E_{\mathrm{prof}}(Y, Z; \Omega) = \frac{1}{|\Omega| G} \sum_{n\in\Omega}\sum_{g=1}^{G} |T(Y_{n:})_g - T(Z_{n:})_g|\,, \tag{25}$$

$$E_{\mathrm{raw}}(Y, Z; \Omega) = \frac{1}{|\Omega|} \sum_{n\in\Omega} \frac{\|Y_{n:} - Z_{n:}\|_1}{\max(\|Z_{n:}\|_1, 1)}. \tag{26}$$

For $r \in \{\mathrm{prof}, \mathrm{raw}\}$, when every row is included we omit Ω and write $E_r(Y, Z)$. In E_{raw}, the count difference is computed before division by the reference total; prediction and reference are not independently normalized. Equation 12 applies both errors to node-aligned recovery.

C.5 Inference and Computational Complexity

Inference. Inference requires mixed observations, candidate donor cells, candidate-edge distances, and the coordinate-derived graph. The mixed observation is the first Block input; each Block constructs its circle operator from X and $\mathcal{G}$, and Equation 11 produces the recovered receiver-aligned matrix.

Algorithm 1: Four-Block PhyGRec inference

Input: Mixed expression matrix X; candidate donor cells and distances; observable graph $\mathcal{G}$; all trained parameters Θ.

Output: Recovered source estimate $\widehat{S}$.

$L \leftarrow 4$; $H_0 \leftarrow X$; $V_0 \leftarrow 0$;
for $b = 1, \ldots, L$ **do**
  Construct $A_b = A_{\phi_b}(X, \mathcal{G})$ and D_b from Equation 19;
  Compute W_b and $\Delta_b^{\mathrm{geo}} \leftarrow -\lambda_b \ell_b D_b^{-1} A_b^\top W_b (A_b H_{b-1} - P_\Omega X)$;
  Select q_b and compute M_b; $\widetilde{\Delta}_b^{\mathrm{geo}} \leftarrow M_b \odot \Delta_b^{\mathrm{geo}} + \mu_b V_{b-1}$;
  $\Delta_b^{\mathrm{graph}} \leftarrow \mathcal{R}_{\psi_b}(H_{b-1}, \widetilde{\Delta}_b^{\mathrm{geo}}, \mathcal{G})$;
  $H_b \leftarrow [H_{b-1} + \widetilde{\Delta}_b^{\mathrm{geo}} + \Delta_b^{\mathrm{graph}}]_+$;
  $V_b \leftarrow H_b - H_{b-1}$;
$\widehat{S} \leftarrow [X + \kappa(H_L - X)]_+$; return $\widehat{S}$;

Computational complexity. Let E_c be the number of sparse candidate-operator edges, E_g the number of bidirectional graph edges, $H = 128$ the graph width, $h = 16$ the adaptive bottleneck, and $L = 4$ the number of PGR Blocks. Sparse forward and adjoint operations cost $O((N + E_c)G)$ per Block; adaptive update modulation costs $O(NGh)$; and graph-context correction costs $O(NGH + E_gH + NH^2)$. Full inference therefore costs

$$O\big(L\left[(N + E_c)G + NGh + NGH + E_gH + NH^2\right]\big), \tag{27}$$

with working memory $O(NG + E_c + E_g + NH)$. The implementation applies A_b and $A_b^\top$ through sparse edge operations and never materializes a dense mixing matrix.

D Training and Comparison Protocol

Fixed-split and receiver-LOCO experiments use the same model and comparison settings; Appendix B defines their scoring units.

D.1 Training Configuration and Ablation Protocol

Setup. The full PhyGRec model has four independently parameterized Blocks, each making one projected proposal, and 2,832,685 active trainable parameters, including four circle parameters per Block. Circle transfer uses candidate masking without a row-mass cap; adaptive update modulation uses a width-16 gated mapping and an eight-dimensional learned block vector; graph-context correction uses two residual width-128 GATv2 layers with four heads per Block. All circle, solver, adaptive-modulation, graph-refinement, and output-scaling parameters are optimized jointly from the start. The two learned output heads start at zero, giving unit adaptive gain and zero graph increment. In every Block, circle scalars start from $\rho_b = 50$, $\gamma_b = \mathrm{softplus}(-2) = 0.1269$, $\eta_b = 1$, and $\tau_b = \mathrm{sigmoid}(2.2) = 0.9002$. A training item is a complete mixing-scene graph; three graphs form a batch and two-batch gradient accumulation gives an effective batch of six graphs.

Scalar parameters are constrained to their stated domains. Step size $\lambda_b = \mathrm{softplus}(v_b)$ starts at 0.2, LayerScale $\ell_b = \mathrm{softplus}(p_b)$ at 1, and momentum $\mu_b = \mathrm{sigmoid}(m_b) \in (0, 1)$ at 0.0237. Graph strength and the post-GCC data-consistency step use softplus and start at 0.1 each. The output scale is $\kappa = \exp(r)$ and starts at 1. The Barron scale is positive through softplus and starts at 1, while its shape uses $\nu_b = 2 - \mathrm{softplus}(z_b) < 2$ and starts at 1.

Training and model selection. We train for 100 epochs with profile and raw loss weights of 1.0 each. The solver, circle operators, adaptive modulation, and graph refinement use learning rates 0.05, 0.01, 0.003, and 0.002, respectively. Training uses AdamW with weight decay 10^{-4}, global gradient clipping at 1.0, and EMA decay 0.99. Checkpoints are selected using validation Log-MAE and Rel. L_1, evaluated every ten epochs; ties favor the earliest epoch. Validation and test predictions use the EMA parameters saved at the selected checkpoint; optimization updates the live parameters.

Ablation protocol. Every direct ablation uses the same identities, profile and raw weights of 1.0, 100-epoch budget, three seeds, ten-epoch validation cadence, and selection rule as the full model. Nonremoved hyperparameters are inherited without per-variant retuning. The three variants disable Residual Backprojection, Adaptive Increment Modulation, and Graph-Context Correction respectively. Active parameter counts (full, w/o RB, w/o AIM, w/o GCC) are 2.833, 2.833, 2.620, and 0.212 million.

D.2 Baselines

Graph neural networks. GAT, GATv2, GCN, GraphSAGE, and MPNN receive the mixed expression X and bidirectional candidate adjacency. Their message-passing layers use expression features and adjacency, without candidate distance or coordinate edge features. Each network predicts a residual relative to X and applies a nonnegative output projection. PhyGRec and Physical PGD additionally use candidate indices, masks, and distances in their observation operators. Table 8 reports the learned baselines' model sizes and validation-selected configurations.

Table 8: Validation-selected baseline configurations. Architecture gives hidden width/layers for graph networks and hidden width/latent dimension for the VAE; GAT and GATv2 use four attention heads. Selected epoch identifies the checkpoint from a common 100-epoch training budget.

Method	Params (M)	Architecture	Learning rate	Selected epoch
GATv2	2.837	$256/2$	10^{-3}	90
GraphSAGE	2.064	$512/2$	3×10^{-3}	70
GAT	7.827	$512/3$	10^{-4}	50
MPNN	6.777	$256/2$	10^{-3}	20
GCN	1.554	$512/3$	10^{-3}	65
VAE	1.605	$512/32$	10^{-3}	75

Variational autoencoder. The VAE control encodes the mixed expression X into a stochastic latent representation and decodes a residual around X. It uses the same supervised profile/raw objective and nonnegative output constraint as the neural graph baselines, testing whether latent expression reconstruction alone can resolve the paired inverse.

Physical PGD. We construct Physical PGD as a physics-only optimization control using projected gradient descent (Bertsekas, 1976) with a train-fitted circle-overlap operator. It applies 100 nonnegative projected-gradient iterations with step size 0.1. For each run, its four circle-operator parameters come from the corresponding train-fitted operator instance; the operator is frozen during inversion. Each inverse step divides the physical gradient by a fixed diagonal normalization computed from the mixing weights before applying the nonnegative projection; it has no learned update modulation or graph-context correction. Validation selects the step size and iteration count. This control therefore measures recovery from a fitted observation model with a long fixed inverse.

Selection protocol. The graph-network search varies learning rate over $\{10^{-4}, 3\times10^{-4}, 10^{-3}, 3\times10^{-3}\}$, hidden width over $\{256, 512\}$, and depth over $\{2, 3\}$. Attention networks use four heads. The VAE uses the same learning-rate and width choices with latent dimension in $\{32, 64\}$ and fixed KL weight 0.02. All neural baselines use profile/raw weights $1.0/1.0$, train for 100 epochs without early stopping, evaluate validation every five epochs, and select checkpoints using validation Log-MAE and Rel. L_1. GAT and GATv2 use mixing scene-graph batch size two; the other neural baselines use three. Global gradient clipping is 5.0. Physical PGD searches step size $\{0.05, 0.1, 0.2, 0.4\}$ and iteration count $\{10, 25, 50, 100\}$ on validation, selecting 0.1 and 100. Learned models use three common training seeds; Physical PGD uses three corresponding train-fitted operator instances.

PhyGRec has $2{,}832{,}685$ trainable parameters, close to GATv2's $2{,}836{,}968$. Selected architectures and hyperparameters transfer to receiver-LOCO, where fold validation selects checkpoint epochs.

D.3 Training and Validation Dynamics

Across the sampled epochs, PhyGRec has lower validation Count-MAE, Log-MAE, and Rel. L_1 than GATv2 and GraphSAGE (Figure 6). The validation panels score the same 32,904 receiver–scene entries for every method; training panels show within-method trajectories under method-specific aggregation weights.

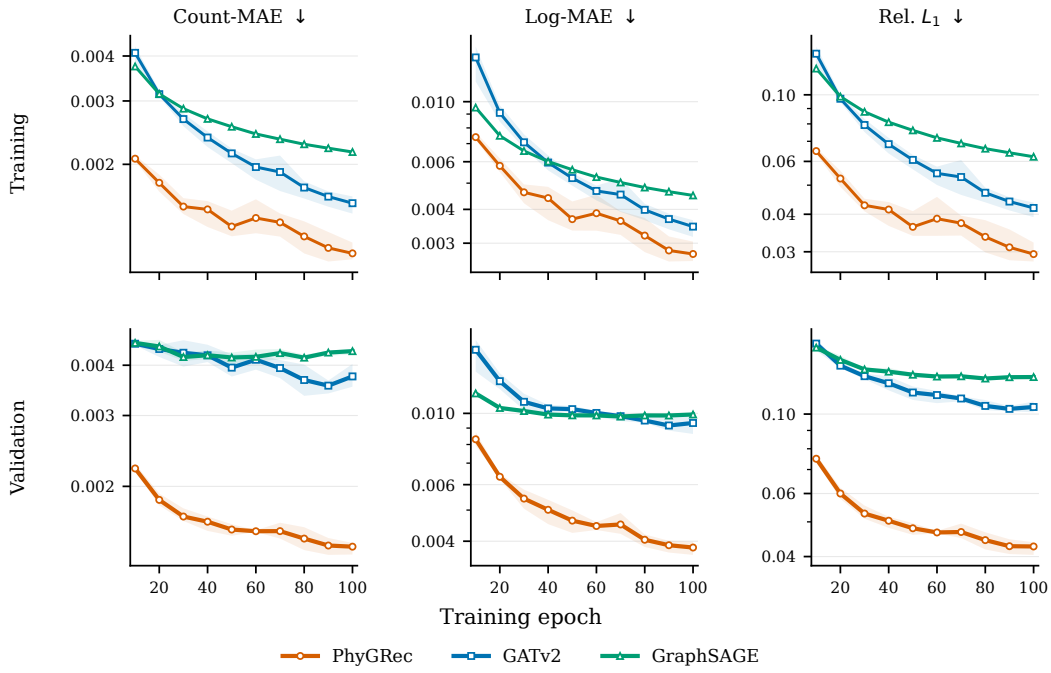

Figure 6: Fixed-split training (top) and validation (bottom) trajectories for Count-MAE, Log-MAE, and Rel. L_1 (left to right). Lines show three-seed means, and shading shows the seed range at ten-epoch intervals. Validation scores use the same 32,904 receiver–scene entries; training aggregation differs by method. Axes are logarithmic, and epochs denote full passes through the training scenes.

E Expression and Structure Results

The fixed-split gain holds across all three chip pairings: PhyGRec has the lowest six expression errors among reconstruction methods (Table 9).

After raw predictions are averaged by physical-cell identity, PhyGRec also has lower Count-MAE (0.00075 versus 0.00212), Log-MAE (0.00749 versus 0.01615), and Rel. L_1 (0.03266 versus 0.08580) than GATv2 (three-seed means; Figure 3). This analysis weights 6,402 cells equally, whereas Table 1 reports receiver–scene means.

Table 9: Test expression errors for each chip pairing, using the six metrics defined in Appendix B. Entries are mean ± sample SD over three fixed training seeds. All errors are lower-is-better; best/second-best within each pairing and metric are bold/underlined.

Method	Count-MAE ↓	Log-MAE ↓	Rel. L_1 ↓	Count-RMSE ↓	Log-RMSE ↓	Rel. L_2 ↓
			A9–L7			
GATv2	<u>0.00330</u> ±0.00009	<u>0.00855</u> ±0.00030	<u>0.09816</u> ±0.00168	<u>0.05992</u> ±0.00181	<u>0.17441</u> ±0.00330	<u>0.12331</u> ±0.00149
GraphSAGE	0.00375 ±0.00002	0.00920 ±0.00018	0.11928 ±0.00114	0.06104 ±0.00062	0.18633 ±0.00204	0.13961 ±0.00043
GAT	0.00529 ±0.00020	0.01703 ±0.00094	0.18962 ±0.00705	0.06993 ±0.00226	0.25591 ±0.00881	0.18383 ±0.00636
MPNN	0.00613 ±0.00017	0.01525 ±0.00106	0.22188 ±0.00410	0.08008 ±0.00114	0.24325 ±0.00959	0.22771 ±0.00579
Physical PGD	0.00380 ±0.00001	0.02996 ±0.00062	0.16406 ±0.00176	0.06140 ±0.00028	0.34366 ±0.00372	0.14949 ±0.00099
GCN	0.00638 ±0.00020	0.02109 ±0.00042	0.22766 ±0.00582	0.07873 ±0.00123	0.28835 ±0.00353	0.21364 ±0.00320
VAE	0.00632 ±0.00004	0.02415 ±0.00043	0.24833 ±0.00084	0.08065 ±0.00016	0.31027 ±0.00285	0.22650 ±0.00067
PhyGRec	**0.00123** ±0.00006	**0.00348** ±0.00022	**0.03777** ±0.00179	**0.02891** ±0.00103	**0.10611** ±0.00358	**0.04571** ±0.00182
			A9–NA			
GATv2	<u>0.00392</u> ±0.00024	<u>0.00976</u> ±0.00049	<u>0.10231</u> ±0.00044	0.06838 ±0.00523	<u>0.18410</u> ±0.00520	<u>0.12641</u> ±0.00199
GraphSAGE	0.00449 ±0.00002	0.01041 ±0.00023	0.12428 ±0.00114	0.06965 ±0.00048	0.19554 ±0.00229	0.14338 ±0.00095
GAT	0.00597 ±0.00018	0.01834 ±0.00092	0.19065 ±0.00647	0.07673 ±0.00185	0.26255 ±0.00814	0.18516 ±0.00575
MPNN	0.00698 ±0.00027	0.01576 ±0.00119	0.22010 ±0.00526	0.08704 ±0.00158	0.24447 ±0.01018	0.22743 ±0.00636
Physical PGD	0.00439 ±0.00001	0.03202 ±0.00063	0.16999 ±0.00180	<u>0.06784</u> ±0.00027	0.35210 ±0.00361	0.15496 ±0.00103
GCN	0.00731 ±0.00029	0.02152 ±0.00058	0.23187 ±0.00696	0.08721 ±0.00003	0.28800 ±0.00488	0.21860 ±0.00404
VAE	0.00706 ±0.00004	0.02564 ±0.00043	0.24976 ±0.00063	0.08781 ±0.00020	0.31647 ±0.00291	0.22869 ±0.00049
PhyGRec	**0.00159** ±0.00005	**0.00412** ±0.00023	**0.04216** ±0.00163	**0.03522** ±0.00046	**0.11453** ±0.00317	**0.05036** ±0.00159
			L7–NA			
GATv2	<u>0.00307</u> ±0.00014	<u>0.00832</u> ±0.00042	<u>0.10042</u> ±0.00191	<u>0.06113</u> ±0.00328	<u>0.17435</u> ±0.00477	<u>0.12529</u> ±0.00090
GraphSAGE	0.00358 ±0.00004	0.00893 ±0.00019	0.12320 ±0.00099	0.06276 ±0.00167	0.18580 ±0.00197	0.14382 ±0.00058
GAT	0.00515 ±0.00021	0.01648 ±0.00099	0.20199 ±0.00613	0.07200 ±0.00114	0.25489 ±0.00900	0.19381 ±0.00495
MPNN	0.00600 ±0.00021	0.01565 ±0.00088	0.23584 ±0.00655	0.08344 ±0.00064	0.24958 ±0.00784	0.23913 ±0.00657
Physical PGD	0.00359 ±0.00002	0.02971 ±0.00066	0.17445 ±0.00204	0.06293 ±0.00035	0.34531 ±0.00390	0.15511 ±0.00118
GCN	0.00623 ±0.00010	0.02171 ±0.00040	0.24378 ±0.00342	0.08242 ±0.00083	0.29645 ±0.00313	0.22638 ±0.00165
VAE	0.00607 ±0.00004	0.02401 ±0.00044	0.26308 ±0.00080	0.08320 ±0.00014	0.31326 ±0.00306	0.23708 ±0.00066
PhyGRec	**0.00115** ±0.00005	**0.00346** ±0.00020	**0.03926** ±0.00159	**0.02803** ±0.00059	**0.10766** ±0.00320	**0.04783** ±0.00153

Table 10: Per-cell agreement with frozen Reference-derived marker annotations on 6,394 matched physical cells. All methods use Reference-fitted gene standardization and five marker programs. Trained rows show mean ± sample SD over three seeds; Mixed is a point estimate. Bal. acc. is macro recall; Macro AP averages one-versus-rest average precision from marker scores. Bold and underlined mark the best and second-best reconstruction methods; Mixed is unranked.

Expression	Macro F1 ↑	Bal. acc. ↑	MCC ↑	Accuracy ↑	Macro prec. ↑	κ ↑	Macro AP ↑
GATv2	<u>0.888</u> ± 0.010	<u>0.889</u> ± 0.009	<u>0.869</u> ± 0.012	<u>0.898</u> ± 0.010	<u>0.888</u> ± 0.011	<u>0.869</u> ± 0.012	<u>0.883</u> ± 0.007
GraphSAGE	0.881 ± 0.004	0.880 ± 0.004	0.858 ± 0.005	0.889 ± 0.004	0.883 ± 0.005	0.858 ± 0.005	0.880 ± 0.005
GAT	0.784 ± 0.009	0.785 ± 0.008	0.741 ± 0.010	0.799 ± 0.008	0.783 ± 0.010	0.741 ± 0.011	0.795 ± 0.008
MPNN	0.806 ± 0.020	0.805 ± 0.021	0.766 ± 0.026	0.818 ± 0.021	0.810 ± 0.017	0.765 ± 0.027	0.818 ± 0.009
GCN	0.745 ± 0.003	0.746 ± 0.003	0.692 ± 0.004	0.761 ± 0.003	0.746 ± 0.004	0.692 ± 0.004	0.761 ± 0.002
VAE	0.710 ± 0.004	0.714 ± 0.003	0.647 ± 0.005	0.725 ± 0.004	0.709 ± 0.004	0.647 ± 0.005	0.725 ± 0.006
Physical PGD	0.680 ± 0.005	0.689 ± 0.005	0.614 ± 0.006	0.696 ± 0.005	0.677 ± 0.005	0.612 ± 0.006	0.690 ± 0.004
Mixed	0.568	0.579	0.470	0.581	0.567	0.468	0.577
PhyGRec	**0.947** ± 0.004	**0.947** ± 0.003	**0.937** ± 0.004	**0.951** ± 0.003	**0.948** ± 0.005	**0.937** ± 0.004	**0.928** ± 0.002

Table 11: Clustering agreement with frozen Reference partitions on 6,394 matched physical cells. For each trained seed, receiver–scene predictions are averaged by cell before independent 20-PC and 30-neighbor cosine Leiden fitting. ARI/NMI average three Leiden initializations; stability is their mean pairwise ARI. Trained rows show mean ± sample SD over three seeds; Reference and Mixed are point estimates. Reference is independently reclustered, so its ARI need not equal one. Bold and underlined mark the best and second-best reconstruction methods; Clusters is unranked.

Expression	ARI ↑	NMI ↑	Silhouette ↑	Stability ARI ↑	Clusters
GATv2	<u>0.686</u>± 0.021	<u>0.635</u>± 0.006	<u>0.078</u>± 0.002	**0.878**± 0.090	5.889± 0.509
GraphSAGE	0.655± 0.005	0.618± 0.011	0.075± 0.002	0.826± 0.025	6.111± 0.192
GAT	0.483± 0.006	0.459± 0.004	0.058± 0.001	0.785± 0.076	5.333± 0.333
MPNN	0.494± 0.049	0.477± 0.035	0.066± 0.013	0.773± 0.057	6.111± 0.962
GCN	0.411± 0.019	0.399± 0.012	0.049± 0.001	0.731± 0.124	5.333± 0.000
VAE	0.306± 0.012	0.312± 0.001	0.037± 0.001	0.651± 0.022	6.222± 0.192
Physical PGD	0.238± 0.024	0.261± 0.011	0.034± 0.001	0.673± 0.069	5.778± 0.192
Reference	0.807	0.769	0.090	0.853	6.667
Mixed	0.122	0.148	0.014	0.622	5.333
PhyGRec	**0.760**± 0.022	**0.719**± 0.009	**0.088**± 0.001	<u>0.870</u>± 0.067	6.444± 0.509

E.1 Fixed-Split Annotation and Clustering

Reference-derived marker labels and cluster partitions provide a complementary view of recovered expression. On 6,394 matched cells, PhyGRec reaches marker Macro F1 0.947 and clustering ARI 0.760, the highest among reconstruction methods (Tables 10 and 11).

E.2 Held-Out Receiver Results

The fixed-split pattern persists when each receiver chip is held out: PhyGRec has the lowest six expression errors in every fold (Table 12) and the highest fold-mean Macro F1 (0.900) and ARI (0.654) among reconstruction methods (Tables 13 and 14).

Table 12: Strict receiver-LOCO expression errors by held-out chip, using the six metrics defined in Appendix B. Entries are mean ± sample SD over three seeds. All errors are lower-is-better; best/second-best within each fold and metric are bold/underlined.

Method	Count-MAE ↓	Log-MAE ↓	Rel. L_1 ↓	Count-RMSE ↓	Log-RMSE ↓	Rel. L_2 ↓
Held-out receiver: A9						
GraphSAGE	0.00501 ±0.00004	<u>0.01251</u> ±0.00008	<u>0.15223</u> ±0.00107	0.07082 ±0.00033	<u>0.21350</u> ±0.00146	0.16432 ±0.00126
GATv2	0.00550 ±0.00051	0.01573 ±0.00223	0.16283 ±0.02498	0.07479 ±0.00370	0.23648 ±0.02106	0.16245 ±0.02114
Physical PGD	<u>0.00390</u> ±0.00002	0.02882 ±0.00064	0.15668 ±0.00002	<u>0.06218</u> ±0.00133	0.33391 ±0.00353	<u>0.14275</u> ±0.00052
GAT	0.00624 ±0.00023	0.01900 ±0.00073	0.20926 ±0.00541	0.07601 ±0.00178	0.26549 ±0.00564	0.19704 ±0.00468
MPNN	0.00716 ±0.00029	0.01686 ±0.00102	0.23366 ±0.00855	0.08608 ±0.00433	0.24979 ±0.00962	0.23167 ±0.00767
GCN	0.00745 ±0.00025	0.02264 ±0.00054	0.24727 ±0.00280	0.08502 ±0.00293	0.29514 ±0.00383	0.22874 ±0.00245
VAE	0.00667 ±0.00024	0.02397 ±0.00136	0.24378 ±0.00531	0.08147 ±0.00068	0.30557 ±0.00981	0.22096 ±0.00464
PhyGRec	**0.00224** ±0.00019	**0.00778** ±0.00072	**0.06842** ±0.00491	**0.04259** ±0.00260	**0.16258** ±0.00829	**0.07777** ±0.00602
Held-out receiver: L7						
GraphSAGE	0.00456 ±0.00003	<u>0.01362</u> ±0.00034	<u>0.17860</u> ±0.00082	0.07380 ±0.00067	<u>0.23340</u> ±0.00297	0.19196 ±0.00033
GATv2	0.00493 ±0.00035	0.01776 ±0.00212	0.18727 ±0.01483	0.07572 ±0.00586	0.26339 ±0.01770	0.18578 ±0.01222
Physical PGD	<u>0.00405</u> ±0.00012	0.03645 ±0.00226	0.21168 ±0.00841	<u>0.07008</u> ±0.00206	0.38374 ±0.01187	<u>0.18298</u> ±0.00614
GAT	0.00610 ±0.00032	0.02396 ±0.00354	0.26437 ±0.00867	0.08103 ±0.00044	0.31097 ±0.02536	0.24021 ±0.00356
MPNN	0.00729 ±0.00060	0.02493 ±0.00490	0.32908 ±0.03442	0.09596 ±0.00396	0.31937 ±0.03229	0.30269 ±0.02109
GCN	0.00638 ±0.00001	0.02774 ±0.00081	0.28246 ±0.00104	0.08305 ±0.00040	0.33897 ±0.00500	0.25280 ±0.00065
VAE	0.00632 ±0.00003	0.02846 ±0.00045	0.29661 ±0.00046	0.08989 ±0.00007	0.34407 ±0.00276	0.26841 ±0.00066
PhyGRec	**0.00180** ±0.00028	**0.00842** ±0.00124	**0.07219** ±0.01166	**0.03855** ±0.00419	**0.17474** ±0.01386	**0.08261** ±0.01417
Held-out receiver: NA						
GraphSAGE	0.00552 ±0.00007	<u>0.01302</u> ±0.00023	0.14927 ±0.00028	0.07288 ±0.00158	<u>0.21397</u> ±0.00222	0.16076 ±0.00150
GATv2	0.00604 ±0.00016	0.01783 ±0.00142	0.15460 ±0.00625	0.07562 ±0.00178	0.24632 ±0.01252	0.15591 ±0.00357
Physical PGD	<u>0.00392</u> ±0.00010	0.02569 ±0.00007	<u>0.13708</u> ±0.00072	<u>0.06415</u> ±0.00231	0.31260 ±0.00100	<u>0.13333</u> ±0.00217
GAT	0.00704 ±0.00062	0.02124 ±0.00212	0.21931 ±0.00819	0.08004 ±0.00299	0.28144 ±0.01551	0.20303 ±0.00841
MPNN	0.00794 ±0.00053	0.01626 ±0.00102	0.22945 ±0.02087	0.08901 ±0.00341	0.24253 ±0.00889	0.23637 ±0.01505
GCN	0.00762 ±0.00039	0.02173 ±0.00136	0.23141 ±0.00665	0.08309 ±0.00334	0.28645 ±0.00934	0.21433 ±0.00716
VAE	0.00680 ±0.00019	0.02209 ±0.00073	0.22358 ±0.00386	0.08325 ±0.00252	0.29057 ±0.00572	0.20586 ±0.00347
PhyGRec	**0.00256** ±0.00043	**0.00784** ±0.00116	**0.07104** ±0.01246	**0.04696** ±0.00723	**0.16055** ±0.01401	**0.08402** ±0.01548

Table 13: Strict receiver-LOCO marker-annotation agreement for the eight primary methods against fold-specific frozen Reference labels. Each fold is scored within each seed before equal-weight fold averaging; entries are mean ± sample SD over three seeds. Bal. acc. is macro recall and AP uses fixed marker scores; higher is better. Best and second-best means are bold and underlined, respectively.

Method	Macro F1 ↑	Bal. acc. ↑	MCC ↑	Accuracy ↑	Macro prec. ↑	κ ↑	Macro AP ↑
GraphSAGE	<u>0.845</u> ±0.003	<u>0.846</u> ±0.003	<u>0.816</u> ±0.003	<u>0.857</u> ±0.002	<u>0.847</u> ±0.002	<u>0.816</u> ±0.003	<u>0.852</u> ±0.001
GATv2	0.811 ±0.005	0.811 ±0.004	0.775 ±0.005	0.825 ±0.004	0.811 ±0.005	0.775 ±0.005	0.823 ±0.005
Physical PGD	0.685 ±0.006	0.694 ±0.005	0.620 ±0.007	0.701 ±0.006	0.684 ±0.006	0.618 ±0.007	0.695 ±0.005
GAT	0.753 ±0.014	0.755 ±0.015	0.702 ±0.019	0.768 ±0.015	0.753 ±0.013	0.701 ±0.019	0.764 ±0.012
MPNN	0.792 ±0.002	0.797 ±0.002	0.752 ±0.003	0.806 ±0.002	0.791 ±0.002	0.751 ±0.003	0.807 ±0.002
GCN	0.721 ±0.005	0.724 ±0.005	0.663 ±0.006	0.737 ±0.005	0.721 ±0.006	0.663 ±0.006	0.739 ±0.005
VAE	0.704 ±0.003	0.707 ±0.003	0.640 ±0.003	0.719 ±0.002	0.705 ±0.002	0.639 ±0.003	0.723 ±0.003
PhyGRec	**0.900** ±0.007	**0.899** ±0.008	**0.882** ±0.008	**0.909** ±0.006	**0.903** ±0.007	**0.882** ±0.008	**0.896** ±0.006

Table 14: Strict receiver-LOCO structure agreement for the eight primary methods against fold-specific frozen Reference partitions. Fold metrics are averaged equally within seed; entries are mean ± sample SD over three seeds. Higher is better for ARI, NMI, silhouette, and stability ARI; best and second-best means are bold and underlined, respectively. Cluster count is descriptive and unranked.

Method	ARI ↑	NMI ↑	Silhouette ↑	Stability ARI ↑	Clusters
GraphSAGE	<u>0.561</u> ± 0.016	<u>0.542</u> ± 0.009	<u>0.0582</u> ± 0.0005	<u>0.834</u> ± 0.026	4.407 ± 0.128
GATv2	0.518 ± 0.016	0.494 ± 0.008	0.0578 ± 0.0015	0.798 ± 0.022	4.259 ± 0.064
Physical PGD	0.200 ± 0.011	0.252 ± 0.011	0.0316 ± 0.0001	0.697 ± 0.006	4.778 ± 0.111
GAT	0.372 ± 0.052	0.379 ± 0.032	0.0438 ± 0.0055	0.687 ± 0.108	4.481 ± 0.170
MPNN	0.403 ± 0.028	0.408 ± 0.021	0.0149 ± 0.0213	0.702 ± 0.057	5.407 ± 0.339
GCN	0.346 ± 0.022	0.350 ± 0.013	0.0410 ± 0.0013	0.667 ± 0.049	3.926 ± 0.170
VAE	0.250 ± 0.043	0.277 ± 0.023	0.0318 ± 0.0008	0.680 ± 0.051	4.481 ± 0.128
PhyGRec	**0.654** ± 0.014	**0.635** ± 0.010	**0.0672** ± 0.0006	**0.857** ± 0.028	4.815 ± 0.064

Table 15: Fixed-split component expression errors on 33,896 receiver–scene entries (three seeds; mean ± sample SD). Count errors use raw counts; log errors use normalized log1p expression. Lower is better; bold and underlined mark the best and second-best means.

Variant	Count-MAE ↓	Log-MAE ↓	Rel. L_1 ↓	Count-RMSE ↓	Log-RMSE ↓	Rel. L_2 ↓
w/o RB	0.00226 ±0.00034	0.00657 ±0.00148	0.06582 ±0.01092	0.04058 ±0.00311	0.14275 ±0.01709	0.07557 ±0.01137
w/o AIM	<u>0.00163</u> ±0.00006	<u>0.00453</u> ±0.00007	<u>0.04637</u> ±0.00136	<u>0.03456</u> ±0.00032	<u>0.11912</u> ±0.00142	<u>0.05522</u> ±0.00056
w/o GCC	0.00261 ±0.00013	0.01075 ±0.00013	0.09121 ±0.00421	0.04847 ±0.00175	0.19659 ±0.00231	0.10595 ±0.00466
PhyGRec (full)	**0.00133** ±0.00005	**0.00369** ±0.00021	**0.03968** ±0.00167	**0.03089** ±0.00061	**0.10944** ±0.00328	**0.04791** ±0.00165

Table 16: Fixed-split component marker-annotation agreement on 6,394 matched cells (three seeds; mean ± sample SD). Labels and standardization are frozen from Reference. Higher is better; bold and underlined mark the best and second-best means.

Variant	Macro F1 ↑	Bal. acc. ↑	MCC ↑	Accuracy ↑	Macro prec. ↑	κ ↑	Macro AP ↑
w/o RB	0.915 ±0.019	0.917 ±0.019	0.901 ±0.023	0.923 ±0.018	0.914 ±0.019	0.901 ±0.023	0.905 ±0.011
w/o AIM	<u>0.945</u> ±0.004	<u>0.945</u> ±0.004	<u>0.934</u> ±0.005	<u>0.949</u> ±0.004	<u>0.945</u> ±0.004	<u>0.934</u> ±0.005	<u>0.926</u> ±0.002
w/o GCC	0.875 ±0.005	0.873 ±0.005	0.853 ±0.005	0.886 ±0.004	0.879 ±0.005	0.853 ±0.005	0.878 ±0.004
PhyGRec (full)	**0.947** ±0.004	**0.947** ±0.003	**0.937** ±0.004	**0.951** ±0.003	**0.948** ±0.005	**0.937** ±0.004	**0.928** ±0.002

Table 17: Fixed-split component clustering agreement on 6,394 matched cells (three seeds; mean ± sample SD). Independently fitted partitions are compared with frozen Reference labels. Higher is better; bold and underlined mark the best and second-best means. Clusters is descriptive and unranked.

Variant	ARI ↑	NMI ↑	Silhouette ↑	Stability ARI ↑	Clusters
w/o RB	0.712 ± 0.020	0.671 ± 0.019	<u>0.0868</u> ± 0.0011	0.852 ± 0.032	6.222 ± 0.694
w/o AIM	<u>0.748</u> ± 0.028	<u>0.709</u> ± 0.014	0.0863 ± 0.0013	**0.876** ± 0.064	6.111 ± 0.192
w/o GCC	0.613 ± 0.049	0.595 ± 0.024	0.0776 ± 0.0004	0.800 ± 0.075	5.556 ± 0.694
PhyGRec (full)	**0.760** ± 0.022	**0.719** ± 0.009	**0.0875** ± 0.0008	<u>0.870</u> ± 0.067	6.444 ± 0.509

E.3 Component Contributions

Each component contributes to fixed-split recovery: removing RB, AIM, or GCC increases Log-MAE in all three matched seeds. The increases are 0.00149–0.00404 without RB, 0.00068–0.00102 without AIM, and 0.00684–0.00734 without GCC; the full model also has the lowest six expression errors among these variants (Table 15). Tables 16 and 17 report the corresponding marker-annotation and clustering results.

F Cell-Level Examples and PBMC Visualizations

F.1 Qualitative Recovery Examples

The ten cases illustrate favorable and high model-to-baseline error-ratio recoveries; Figure 3 reports the cohort-level error distributions.

Illustrative-case selection. The six genes were fixed from test Reference contrasts and generator metadata before examining model errors: donor-enriched BACH2, IL7R, and FOXP1, and receiver-enriched VCAN, LYZ, and S100A9. Selection used first-scene single-donor pairs with at least 10

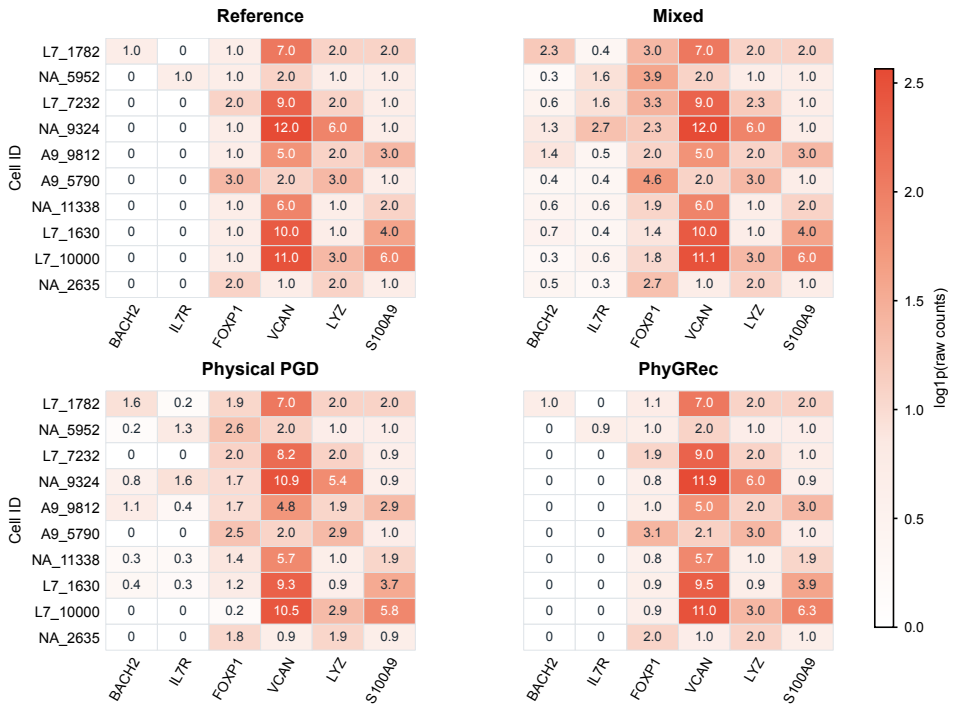

Figure 7: Reference and predicted expression for the ten Figure 5 cases, with five low and five high model-to-baseline error-ratio examples. Columns show Reference, Mixed, Physical PGD, and PhyGRec; color encodes $\log(1 + \text{count})$ on a shared scale.

panel counts per source and injected-count fraction 0.10–0.75; each contrast required at least two counts and $3\max(S_{\text{other}}, 0.5)$ counts, with at least 0.5 injected count for donor markers.

The ten examples come from 65 eligible single-donor cell–scene instances (56 receivers) under the same source-count and injected-fraction limits. Eligibility also required marker enrichment of at least one count and $2\max(S_{\text{other}}, 0.5)$ counts, with at least 0.25 injected count per donor marker. Of these, 32 instances (30 receivers) met the favorable-case rule: PhyGRec had at most half of each baseline's error on both six-marker MAE and full 1,000-gene raw L_1, and reduced the corresponding Mixed errors by at least 80% and 70%. With $R = E_{PhyGRec}/\min(E_{\text{Physical PGD}}, E_{\text{GATv2}})$ for six-marker raw-count MAE, panel (a) shows the five smallest R values among favorable cases and panel (b) the five largest among all eligible cases, excluding identities in (a). Each panel uses at most one scene per receiver; cell–scene order breaks ties. The high-ratio cases have PhyGRec marker MAE of 0.014–0.131 counts. Both panels use seed 20260816 and validation-selected checkpoints.

F.2 PBMC Illustrations

Figure 4(a) shows PBMC expression structure for 6,394 matched test cells with nonzero Reference profiles. For display, each seed takes the gene-wise scene median per cell from validation-selected checkpoints, then the three seed profiles are averaged. Frozen Reference Leiden labels follow cell identities across states. Each state independently fits normalization, scaling, PCA, and UMAP; the Reference layout uses 20 PCs and cosine UMAP with 50 neighbors, minimum distance 0.05, repulsion 1, and seed 20260816. Matched-cell orthogonal Procrustes rotation and translation align Mixed and PhyGRec to Reference without scaling or reflection; these six panels share axis limits. Main-table annotation and clustering scores use expression profiles, not UMAP coordinates.

Figure 4(b) uses the complete, unseen Y40176KE chip as measured input to the frozen three-seed PhyGRec model, without synthetic geometric overlap or independent clean truth. The 8,821 cells with nonzero panel expression are shown. Reference and PhyGRec are each normalized, scaled, projected to 20 PCs, and embedded with cosine UMAP (30 neighbors, minimum distance 0.02, repulsion 1, seed 20260816); the two UMAPs are rigidly aligned and share axis limits. Leiden

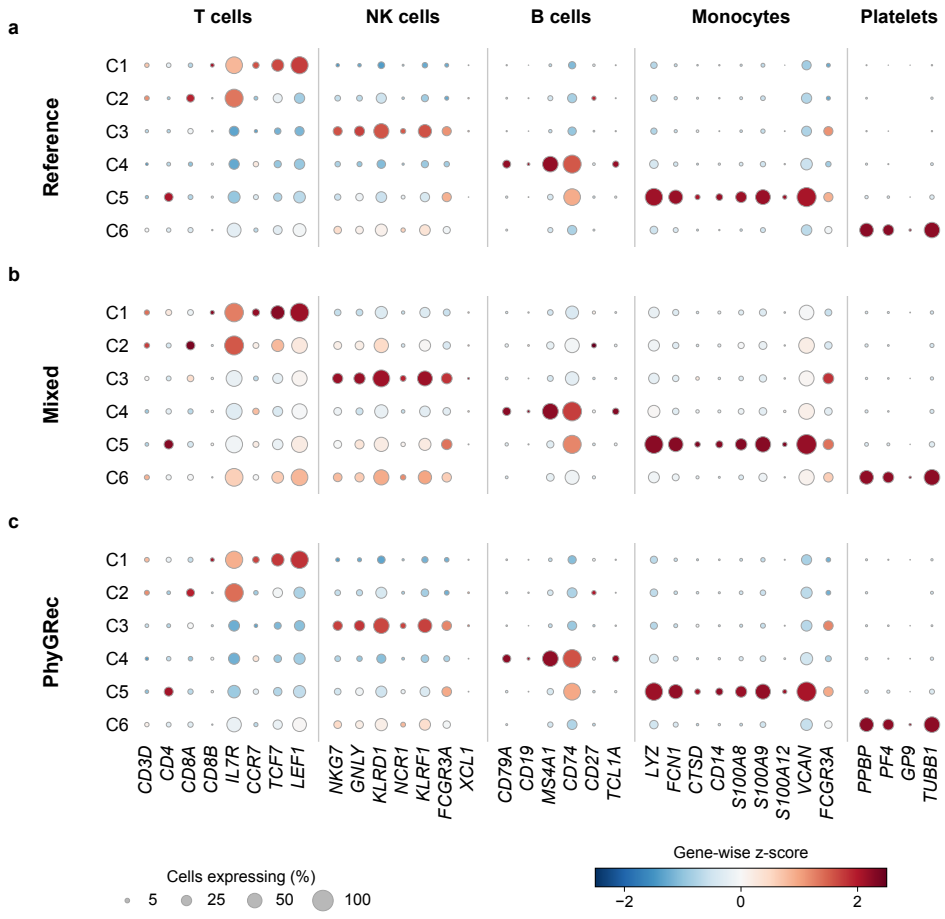

Figure 8: Expression of five fixed annotation programs in frozen Reference clusters C1–C6 for Reference, Mixed, and three-seed mean PhyGRec profiles. The 34 columns represent 33 genes because FCGR3A appears in both NK and monocyte programs. Dot area shows the fraction of cells with positive count-like expression; color shows the gene-wise standardized mean of total-normalized log-expression over the 18 state–cluster combinations.

clustering uses each state's own 20-PC representation, 30 cosine neighbors, resolution 0.35, and the same seed, yielding seven Reference and eight PhyGRec clusters. The bars report Davies–Bouldin and Calinski–Harabasz indices on these two-dimensional coordinates and state-specific clusters; they describe the displayed geometry, not cell-type accuracy.

Figure 8 visualizes the five fixed annotation programs (Appendix B.2.1) underlying the Reference-derived labels across frozen Reference clusters.

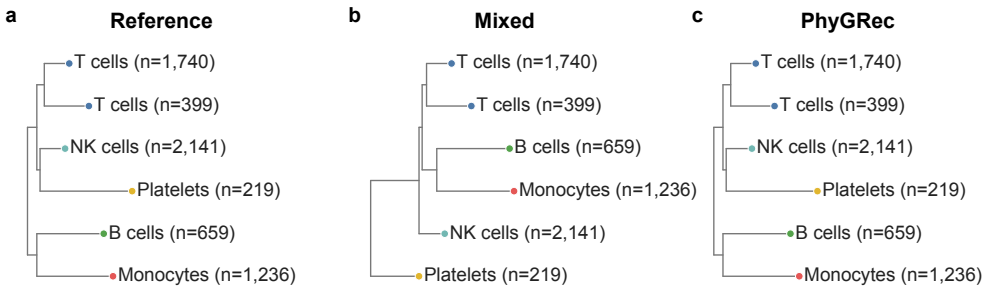

Figure 9: Exploratory Bonsai (de Groot et al., 2026) representations of the six frozen Reference cluster centroids in state-specific 20-PC spaces for Reference, Mixed, and three-seed mean PhyGRec expression. Leaves are labeled by the broad Reference marker annotations; the two Reference T-cell clusters remain distinct leaves. The tree layout summarizes similarity between centroids in expression space.