# DUAL-ORDER RÉNYI–DIRICHLET GRAPH AUTOENCODERS FOR ANOMALY DETECTION ON THE SIMPLEX

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Low- and high-dimensional simplex-valued data arise in microbial communities, spatial transcriptomic profiles, particle-size distributions, and normalized image intensities, where observations may be connected by a graph and corrupted by heterogeneous anomalies. We introduce Dual-Order Rényi–Dirichlet Graph Autoencoders (RD-GAEs), a simplex-geometry preserving graph autoencoder that uses Dirichlet latent variables and Rényi divergences at two distinct orders. The framework utilizes subunit training order emphasizes the dominant, regular structure and reduces the influence of contaminated observations during representation learning. After training, a larger detection order amplifies reconstruction discrepancies in distributional tails, yielding anomaly-sensitive scores without retraining the model. The encoder and decoder preserve simplex geometry through graph aggregation and gated probability transformations, while the graph can be row-stochastic or doubly stochastic. Across controlled-contamination and held-out detection protocols for image, microbiome, and spatial transcriptomics data, RD-GAE achieves superior anomaly-detection performance while retaining interpretable simplex-valued representations, compared with graph, variational, Dirichlet, and Euclidean-coordinate autoencoder baselines.

## 1 INTRODUCTION

Many scientific measurements are compositional rather than unconstrained: microbial relative abundances, cell-type or gene-module proportions at spatial transcriptomic locations, particle-size distributions along spatiotemporal trajectories, and normalized image intensities are all nonnegative vectors whose entries sum to one. Such profiles lie on a simplex/composition and are frequently linked by spatial, temporal, physical, or similarity relations. The learning task must therefore respect simplex-geometry and dependence between observations. This is particularly important for anomaly detection where a corrupted profile may be unusual in its own composition and may also propagate its influence to neighboring nodes through graph message passing.

Autoencoders and variational autoencoders learn nonlinear reconstructions (Hinton & Salakhutdinov, 2006; Kingma & Welling, 2014), while GCNs, GraphSAGE, GAT, and diffusion convolutions incorporate graph structure in transductive, inductive, attention-based, and directed settings (Kipf & Welling, 2017; Hamilton et al., 2017; Velickovic et al., 2018; Atwood & Towsley, 2016). These methods are not designed for a simplex-valued reconstruction problem: an output can be normalized after decoding while the intermediate feature maps, category transformations, and graph messages remain unconstrained. Consequently, negative mass, changes in total mass, and uninterpretable category mixing can enter the representation before the final output constraint is imposed.

Graph autoencoders such as VGAE and GraphMAE commonly reconstruct links or attributed node features (Kipf & Welling, 2016; Hou et al., 2022), and graph anomaly methods such as DOMINANT reconstruct an attributed network (Ding et al., 2019). Robust autoencoders and robust-divergence variational methods reduce the influence of unlikely observations (Zhou & Paffenroth, 2017; Futami et al., 2018; Akrami et al., 2022), but do not address graph-mediated exposure of compositional profiles. Conversely, Dirichlet VAEs provide simplex latent variables (Joo et al., 2020), whereas Dirichlet graph VAEs have mainly focused on adjacency modeling or graph partitioning (Li et al., 2020). Thus existing approach lacks in jointly providing a profile-to-profile graph autoencoder



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.

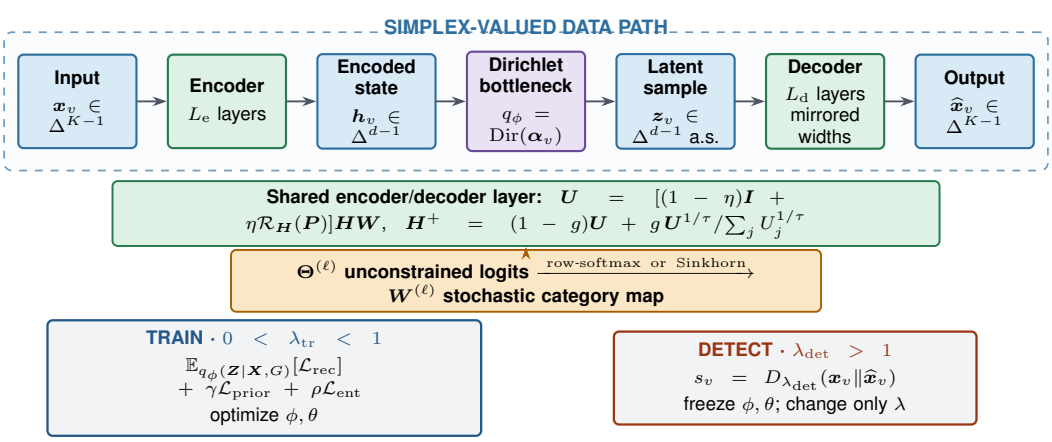



Figure 1: Overview of RD-GAE. Every data-carrying state is simplex-valued; each layer combines a stochastic graph operator $\boldsymbol{P}^{(\ell)}$, a row-stochastic category-mapping matrix $\boldsymbol{W}^{(\ell)}$, and a gated probability-temperature map. The Dirichlet bottleneck maps $\boldsymbol{h}_v$ to $\boldsymbol{\alpha}_v = \alpha_{\min}\mathbf{1} + \kappa\boldsymbol{h}_v$ before a Dirichlet draw. Encoder and decoder widths are structurally mirrored but their parameters are independent. Training uses a subunit reconstruction order and minimizes the posterior-expected objective while detection freezes the fitted reconstruction and changes only the scoring order.

with simplex-valued data-carrying states, an explicit category-mapping interpretation, and distinct mechanisms for robust fitting and tail-sensitive detection.

To address these challenges, we propose RD-GAE, illustrated in Figure 5, a simplex geometry preserving autoencoder with independently parameterized encoder-decoder. The architecture specifically utilizes: a) row-stochastic, or optionally doubly stochastic graph operators aggregate neighboring profiles as convex combinations; b) row-stochastic category-mapping matrices as weight matrix to preserve nonnegativity and unit mass across dimensions; and c) the gated probability-temperature map controls within-profile contrast while residual graph mixing retains local information. The framework prposes a Dirichlet bottleneck to represent latent uncertainty directly on the simplex. Thus, RD-GAE reconstructs the nodewise composition rather than treating the adjacency matrix as its sole target, while retaining an interpretable graph and category operation at every layer.

Additionally, RD-GAE combines this architecture with an exact Rényi reconstruction-plus-prior objective evaluated using stable Log-Beta, `lgamma`, and log-sum-exp operations in PyTorch. A subunit $\lambda_{\mathrm{tr}} < 1$ is used to fit the predominant structure without deleting observed profiles, whereas a validation-calibrated superunit $\lambda_{\mathrm{det}} > 1$ is used only for frozen anomaly scoring. We further support our choice of dual-order approach with a small-order attenuation result, a conditional dual-order separation theorem, and graph-local and global contamination influence for row-stochastic and doubly stochastic propagation. Finally, the our experimental study showcase the efficasy of proposed approach compared to existing approaches adaped for simplex observations on controlled contamination across four simplex data applications including genomics data and image data.

## 2 Rényi–Dirichlet Graph Autoencoder

We consider graph $G = (V, E)$ relating $|V| = N$ observation $\boldsymbol{X} \in \mathbb{R}_{\geq 0}^{N\times K}$ as row-stacked simplex profiles $\boldsymbol{X}\mathbf{1}_K = \mathbf{1}_N$. Let $\mathbf{A} \in \mathbb{R}_{\geq 0}^{N\times N}$ be the graph induced weighted adjacency matrix. One may encode the underlying dependencies through a row-stochastic graph operator $\mathbf{P} = \widetilde{\mathbf{D}}^{-1}\widetilde{\mathbf{A}}$, where $\widetilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}_N$ and $\widetilde{\mathbf{D}} = \mathrm{diag}(\widetilde{\mathbf{A}}\mathbf{1}_N)$, satisfying $\mathbf{P}\mathbf{1}_N = \mathbf{1}_N$. Alternatively, a doubly stochastic operator that additionally satisfies $\mathbf{1}_N^\top\mathbf{P} = \mathbf{1}_N^\top$; it can be constructed using Sinkhorn scaling when the graph support admits the prescribed marginals (Sinkhorn, 1964; Cuturi, 2013). Row stochasticity is the minimal condition for local simplex validity. Double stochasticity is an additional global balance condition and is not forced on directed temporal graphs whose asymmetric flow is scientifically meaningful.

### 2.1 SIMPLEX ENCODER, DIRICHLET BOTTLENECK, AND DECODER

To define the autoencoder, let $k_0 = K$ denote the input dimension, $k_L = d$ denote the latent dimension, and $\mathbf{H}^{(0)} = \mathbf{X}$ be the input simplex profile. The encoder and decoder share a symmetric layer architecture (decoder parameters are entirely un-tied from those of the encoder), where the $l$-th layer is defined as:

$$\mathbf{U}^{(l+1)} = \overline{\mathbf{P}}\eta_l \mathbf{H}^{(l)}\mathbf{W}^{(l)}, \qquad \mathbf{H}^{(l+1)} = Tg_l, \tau_l(\mathbf{U}^{(l+1)}), \tag{1}$$

where each category mapping matrix $\mathbf{W}^{(l)} \geq 0$ is row-stochastic and parameterized via a row-wise softmax: $\mathbf{W}^{(l)} = \text{softmax}_{\text{row}}(\mathbf{\Theta}^{(l)}/\tau_W)$, $T_{g,\tau}(\boldsymbol{u})$ is the node-local escort gate (see section 2.2 for details) and $\overline{\boldsymbol{P}}_\eta = (1-\eta)\boldsymbol{I} + \eta\boldsymbol{P}$ is the graph operator. The robust variant $\overline{\boldsymbol{P}}_\eta$ exponentially downweights graph edges whose sender disagrees in total variation with the receiver's local consensus and then renormalizes rows. While our core model relies on an encoder-only architecture, our ablation study assesses all-layer propagation as well as a parameter-free simplex skip, $(1-\omega)\mathbf{H}_{\text{d}} + \omega\mathbf{H}_{\text{e}}$, for non-final mirrored decoder states, see Appendix Table 6. The final encoder row $\boldsymbol{h}_v \in \Delta^{d-1}$ parameterizes the concentration vector

$$\boldsymbol{\alpha}_v = \alpha_{\min}\mathbf{1}_d + \kappa\boldsymbol{h}_v, \qquad q_\phi(\boldsymbol{Z} \mid \boldsymbol{X}, G) = \prod_{v=1}^{N} \text{Dir}(\boldsymbol{z}_v; \boldsymbol{\alpha}_v), \tag{2}$$

where $\alpha_{\min}, \kappa > 0$ are hyperparameters. Thus every $\boldsymbol{z}_v \in \Delta^{d-1}$ almost surely. The decoder applies Equation (1) in reversed dimension order and returns $\widehat{\boldsymbol{X}}\mathbf{1}_K = \mathbf{1}_N$. Raw logits and concentration vectors are unconstrained parameterizations of stochastic maps and Dirichlet distributions, respectively; they are not data-carrying simplex states. The data-carrying states are $\mathbf{H}^{(\ell)}$, $\mathbf{Z}$, and $\widehat{\mathbf{X}}$, each of which lies on its corresponding simplex.

### 2.2 GATED PROBABILITY TEMPERATURE: SHARPENING AND SATURATION

Every primary layer in (1) applies the node-local escort gate

$$T_{g,\tau}(\boldsymbol{u}) = (1-g)\boldsymbol{u} + gE_\tau(\boldsymbol{u}), \qquad 0 < g < 1, \quad \tau > 0, \tag{3}$$

where $E_\tau(\boldsymbol{u}) = \text{softmax}(\log \boldsymbol{u}/\tau) = \frac{\boldsymbol{u}^{1/\tau}}{\sum_j u_j^{1/\tau}}$. The transformation is a convex combination of two probability vectors with aim to retain simplex geometry. For a positive probability vector $\boldsymbol{u}$, the escort transformation satisfies $\frac{E_\tau(\boldsymbol{u})_i}{E_\tau(\boldsymbol{u})_j} = \left(\frac{u_i}{u_j}\right)^{1/\tau}$. Thus, $\tau$ directly controls the contrast among the components of a node profile. In particular, when $\tau < 1$, every nonzero within-row log-odds magnitude is amplified. The corresponding Jacobian satisfies

$$\|J_{E_\tau}(\boldsymbol{u})\|_1 = \max_j \frac{2E_j(1-E_j)}{\tau u_j}, \tag{4}$$

which makes explicit the dependence of the local sensitivity on both the temperature and the underlying probability profile (see proof in Appendix A.3). We further demonstrate the relevance of the escort gate by comparing its performance with that of the conventional gate $(1-g)\boldsymbol{u} + g\,\text{softmax}(\boldsymbol{u}/\tau)$; see Appendix Table 6.

*Remark* 2.1. A smaller $\tau$ does not always improve node separation. As $\tau \downarrow 0$, the escort distribution concentrates on the dominant components, so profiles sharing the same dominant category may collapse toward the same simplex vertex. This saturation prevents a universal separation guarantee from simply decreasing the temperature. Consistent with this behavior, in our emparical studies, we observed that MNIST diagnostics favored $g \in 0.95, 0.99$ and $\tau \in 0.01, 0.02, 0.03, 0.05$ with limited residual graph mixing, with strong performance near $(g, \tau) = (0.99, 0.03)$. Performance deteriorated at $\tau = 0.003$, so we select $(g, \tau)$ using validation anomaly ranking together with retention and saturation diagnostics.

### 2.3 RÉNYI RECONSTRUCTION LOSS AND DIRICHLET PRIOR MATCHING

For two positive probability vectors $\boldsymbol{p}$ and $\boldsymbol{q}$, the Rényi divergence of order $\lambda > 0$, $\lambda \neq 1$, is defined as $D_\lambda(\boldsymbol{p}\|\boldsymbol{q}) = \frac{1}{\lambda-1}\log\left(\sum_k p_k^\lambda q_k^{1-\lambda}\right)$, with the Kullback–Leibler divergence obtained in

the limiting case $\lambda \to 1$ (van Erven & Harremo"es, 2014). We construct the RDGAE objective in three steps. First, instead of using a Euclidean reconstruction loss, to attain robustness (Akrami et al., 2022), we measure the discrepancy between the observed node profile $\boldsymbol{x}_v$ and its reconstruction $\widehat{\boldsymbol{x}}_v$ using the Rényi divergence with order $\lambda_{tr}$. Since the latent representation is random under the variational posterior, the reconstruction term is defined through its posterior expectation; see Appendix A for details. Second, we regularize the node-specific Dirichlet posterior $q(\boldsymbol{z}_v \mid \boldsymbol{x}_v) = \mathrm{Dir}(\boldsymbol{\alpha}_v)$ toward the prior $p(\boldsymbol{z}_v) = \mathrm{Dir}(\boldsymbol{\beta})$ using a separate Rényi order $\lambda_z$. Collecting the powers of the simplex coordinates yields a Dirichlet kernel with parameter $\boldsymbol{c}_v = \lambda_z \boldsymbol{\alpha}_v + (1 - \lambda_z)\boldsymbol{\beta}$. With $\boldsymbol{c}_v > 0$ componentwise, integration over the simplex gives

$$D_{\lambda_z}\left(\mathrm{Dir}(\boldsymbol{\alpha}_v) \| \, \mathrm{Dir}(\boldsymbol{\beta})\right) = \frac{\log B(\boldsymbol{c}_v) - \lambda_z \log B(\boldsymbol{\alpha}_v) - (1 - \lambda_z) \log B(\boldsymbol{\beta})}{\lambda_z - 1}, \tag{5}$$

where $\log B(\boldsymbol{a}) = \sum_j \log \Gamma(a_j) - \log \Gamma\left(\sum_j a_j\right)$ denotes the log multivariate beta function; see Appendix A for details. The expression follows directly by combining the standard definition of Rényi divergence with the normalizing constant of the Dirichlet distribution (van Erven & Harremo"es, 2014; Ng et al., 2011).

Third, we approximate the posterior reconstruction expectation using $S$ reparameterized samples. The resulting empirical RDGAE objective is

$$\boxed{\widehat{\mathcal{L}}_{\mathrm{RDGAE}} = \frac{1}{SN} \sum_{s=1}^{S} \sum_{v=1}^{N} D_{\lambda_{\mathrm{tr}}}\left(\boldsymbol{x}_v \| \widehat{\boldsymbol{x}}_v^{(s)}\right) + \frac{\gamma}{N} \sum_{v=1}^{N} D_{\lambda_z}\left(\mathrm{Dir}(\boldsymbol{\alpha}_v) \| \, \mathrm{Dir}(\boldsymbol{\beta})\right) + \rho \, \overline{\mathsf{H}}(\boldsymbol{W})} \tag{6}$$

where $\gamma > 0$ controls the contribution of the latent regularization term and $\rho \geq 0$ penalizes high-entropy, nearly uniform transition rows $\mathbf{W}$ and therefore encourages more concentrated category transitions.

### 2.4 Dual-order training and frozen detection

We set $0 < \lambda_{\mathrm{tr}} < 1$ during training. At evaluation, all network parameters, preprocessing, graph operators, Monte Carlo conventions, and $\lambda_z$ are frozen; only the node score $\widehat{\boldsymbol{q}}_v$ at superunit order

$$s_v(\lambda_{\mathrm{det}}) = D_{\lambda_{\mathrm{det}}}(\boldsymbol{x}_v \| \widehat{\boldsymbol{q}}_v), \qquad \lambda_{\mathrm{det}} > 1, \tag{7}$$

is computed. Thus, changing $\lambda_{\mathrm{det}}$ can alter scores and rankings but cannot alter $\widehat{\boldsymbol{q}}_v$. The strategy with $\lambda_{\mathrm{tr}} < 1$ enables us using the data with contamination for training, without deleting or imputing outlying compositions. Our approach moderates categorywise density-ratio sensitivity while the model learns the predominant reconstructible structure. This distinction matters because an arbitrarily small order would also remove the useful optimization signal; see Lemma 2.2.

**Lemma 2.2** (Small-order attenuation)**.** *For positive $\boldsymbol{p}, \boldsymbol{q}$, a softmax-parameterized reconstruction, and $\lambda \downarrow 0$, both $D_\lambda(\boldsymbol{p}\|\boldsymbol{q})$ and its reconstruction-logit gradient are $O(\lambda)$. Hence, a subunit order can moderate sensitivity to distributional outliers, but $\lambda_{\mathrm{tr}}$ cannot approach zero without suppressing learning. The interior first-order expansion is derived in Appendix B.1.*

*Remark* 2.3 (Endpoint order behavior). For a possibly sparse composition $\boldsymbol{p}$, the standard order-zero limit is $D_0(\boldsymbol{p}\|\boldsymbol{q}) = -\log \sum_{k:p_k>0} q_k$; it retains only the support of $\boldsymbol{p}$, not its positive magnitudes. At the other extreme, $\lim_{\lambda\to\infty} D_\lambda(\boldsymbol{p}\|\boldsymbol{q}) = \log \max_k(p_k/q_k)$, so an excessively high detection order behaves as a single-category maximum-ratio score. Both endpoint identities are standard properties of Rényi divergence (van Erven & Harremo"es, 2014).

After training, the reconstruction is fixed and a superunit order is used only to expose distributional discrepancies.

**Theorem 2.4** (Conditional dual-order separation)**.** *Let $(\boldsymbol{p}^A, \boldsymbol{q}^A)$ and $(\boldsymbol{p}^N, \boldsymbol{q}^N)$ be positive anomalous and normal observation–reconstruction pairs, respectively. For $r \in \{A, N\}$, define $Y_k^r = \log(p_k^r/q_k^r)$, $\psi_r(t) = \log \sum_k q_k^r e^{tY_k^r}$, and $\pi_t^r(k) = q_k^r \exp\{tY_k^r - \psi_r(t)\}$. Here, $\pi_t^r$ is the exponentially tilted categorical distribution induced by the log-density-ratio statistic $Y_k^r$. For the anomaly–normal score gap $\Delta(\lambda) = D_\lambda(\boldsymbol{p}^A\|\boldsymbol{q}^A) - D_\lambda(\boldsymbol{p}^N\|\boldsymbol{q}^N)$, suppose that, for some $\lambda > 1$, $\mathrm{Var}_{\pi_t^A}(Y^A) \geq \mathrm{Var}_{\pi_t^N}(Y^N)$ for every $t \in [1, \lambda]$. Then $\Delta(\lambda)$ is nondecreasing on $(1, \lambda]$.*

*Remark* 2.5. Once training is complete, changing $\lambda_{\mathrm{det}}$ changes only the anomaly score; it cannot change the frozen reconstruction $\widehat{\boldsymbol{q}}_v$. Under the exponentially tilted log-density-ratio variance condition in Theorem 2.4, increasing $\lambda_{\mathrm{det}} > 1$ weakly increases anomalous-versus-normal score separation. In particular, if $\boldsymbol{p}^N = \boldsymbol{q}^N$ and $\boldsymbol{p}^A \neq \boldsymbol{q}^A$, the normal score is zero while the anomalous score, and hence the score gap, strictly increases for $\lambda_{\mathrm{det}} > 1$.

In finite validation samples, AUROC can still rise and subsequently fall as the detection order increases because rankings, ties, and sensitivity to extreme coordinates change with order. We therefore evaluate a predeclared superunit grid using known synthetic validation outliers, select the order with the best validation AUROC (breaking ties toward the smaller order), and evaluate the test set once at this frozen choice. Proofs and the exact small-order and tilted-variance formulas are given in Appendix B.1.

### 2.5 Implementation of RD-GAE

We have implemented RD-GAE in PyTorch and optimized with Adam. Dirichlet samples are drawn through PyTorch's reparameterized distribution interface, and automatic differentiation is applied to the complete objective. Stable `lgamma` and `logsumexp` primitives evaluate the Dirichlet and Rényi terms. For each application, we specify the architectures, and than model and optimization hyperparameters are selected by randomized search with k-fold cross-validation strategy.

## 3 Theory properties

Here we study how contamination propagates through the graph and affects reconstructions at nodes whose original profiles remain unchanged. For this, we fix the fitted parameters and graph operators, and consider the sequential encoder–decoder path without skip connections. In this analysis, we are replacing the Dirichlet draw by its posterior mean, i.e., $m(\boldsymbol{h}) = \frac{\alpha_{\min}\mathbf{1}_d + \kappa \boldsymbol{h}}{d\alpha_{\min} + \kappa}$, gives the deterministic map $F_G = D_G \circ m \circ E_G$. This simplex-preserving map contracts $\ell_1$ differences by $a = \frac{\kappa}{d\alpha_{\min} + \kappa} < 1$. We index the encoder and decoder layers in execution order by $r = 0, \ldots, M-1$, with nonnegative row-stochastic graph and category operators $\boldsymbol{P}_r$ and $\boldsymbol{W}_r$, respectively, taking $\boldsymbol{P}_r = \boldsymbol{I}$ for graph-free layers. Let $L_r < \infty$ denote an $\ell_1$ Lipschitz bound for the nodel-local escort gate on a convex domain containing all preactivations induced by the allowed simplex replacements. Define

$$\mathcal{A} = \boldsymbol{P}_{M-1} \cdots \boldsymbol{P}_0, \qquad B = a \prod_r L_r, \qquad c(\boldsymbol{P}) = \max_u \sum_v P_{vu}, \qquad \Gamma = B \prod_r c(\boldsymbol{P}_r). \tag{8}$$

**Theorem 3.1** (Graph exposure and contamination amplification across nodes)**.** *Under this framework, if $\widetilde{\boldsymbol{X}}$ differs from $\boldsymbol{X}$ only on $C \subseteq V$, with all rows remaining in $\Delta^{K-1}$, then*

$$\|F_G(\widetilde{\boldsymbol{X}})_v - F_G(\boldsymbol{X})_v\|_1 \leq 2B[\mathcal{A}\mathbf{1}_C]_v, \qquad \frac{1}{N}\sum_v \|F_G(\widetilde{\boldsymbol{X}})_v - F_G(\boldsymbol{X})_v\|_1 \leq 2\Gamma\frac{|C|}{N}. \tag{9}$$

*Remark* 3.2. The bounds separate graph exposure from sensitivity: at an individual node, $[\mathcal{A}\mathbf{1}_C]_v$ measures its exposure to contamination, while $B = a\prod_r L_r$ bounds the resulting amplification through the gates and attenuation by the posterior mean. Lower exposure or a smaller $B$ gives a smaller certified reconstruction displacement. Exposure vanishes outside the $M$-hop contaminated neighborhood when operators use only graph edges and self-loops; the gates and posterior mean add no cross-node paths. Globally, $\Gamma = B\prod_r c(\boldsymbol{P}_r)$ additionally accounts for broadcasting one source's error to multiple receiving nodes. Doubly stochastic operators give $c(\boldsymbol{P}_r) = 1$ and hence $\Gamma = B$, removing this graph amplification while allowing local diffusion. If also $L_r \leq 1$ uniformly over the allowed replacements, the average displacement is at most $2a|C|/N$. These smaller upper bounds certify stronger resistance to contamination, although reconstruction accuracy remains a separate requirement.

Because simplex outputs are bounded, we define breakdown by a prescribed reconstruction displacement $0 < t \leq 2$. Let $\mathfrak{X}_C(\boldsymbol{X})$ contain all simplex-valued contamination restricted to $C$. We define $\epsilon_t^{\mathrm{pred}}(F_G; \boldsymbol{X})$ as the smallest $|C|/N$, over $C \subsetneq V$, such that $\sup_{\widetilde{\boldsymbol{X}} \in \mathfrak{X}_C(\boldsymbol{X})} \frac{1}{N-|C|}\sum_{v \notin C} \|F_G(\widetilde{\boldsymbol{X}})_v - F_G(\boldsymbol{X})_v\|_1 \geq t$. If no proper contamination set reaches this threshold, we define $\epsilon_t^{\mathrm{pred}}(F_G; \boldsymbol{X}) = 1$, indicating that breakdown is not reached while at least one node remains uncontaminated.

**Theorem 3.3** (Lower bound on clean-node breakdown)**.** *Under the framework of Theorem 3.1, with fitted parameters and graph operators fixed and the constants $L_r$ valid uniformly over all allowed simplex replacements,*

$$\epsilon_t^{\mathrm{pred}}(F_G; \boldsymbol{X}) \geq \frac{\lceil Nt/(2\Gamma + t) \rceil}{N} \geq \frac{t}{2\Gamma + t}. \tag{10}$$

*For doubly stochastic graph operators and $L_r \leq 1$ at every layer, $\epsilon_t^{\mathrm{pred}} \geq t/(2a + t)$.*

*Remark* 3.4. The bound $t/(2\Gamma + t)$ increases as $\Gamma = a \prod_r L_r c(\boldsymbol{P}_r)$ decreases, certifying that a larger contamination fraction is needed to shift clean-node reconstructions by tolerance $t$. Each component controls a different mechanism: $c(\boldsymbol{P}_r)$ bounds the broadcasting of errors across nodes, $L_r$ bounds their amplification by the node-local gate, and $a < 1$ captures attenuation by the posterior mean. Doubly stochastic graph operators eliminate column-sum amplification; if additionally $L_r \leq 1$ uniformly over all allowed replacements, then $\Gamma \leq a$ and $\epsilon_t^{\mathrm{pred}} \geq t/(2a + t)$. Thus graph balance, controlled gate sensitivity, and posterior shrinkage jointly strengthen the breakdown certificate. This certifies stability relative to the original predictions, not reconstruction accuracy: stronger shrinkage may also suppress useful variation. The guarantee applies to fixed fitted parameters and unchanged graph operators. We have relegted proofs to the Appendix C.

## 4 Experimental Result

We investigate the efficacy of RD-GAE using four applications: MNIST, Omniglot, environmental microbiome, and spatial transcriptomics.

### 4.1 Data preparation and Baseline Models

The image applications use MNIST (LeCun et al., 1998) and Omniglot (Lake et al., 2015), obtained through `torchvision`. After preprocessing, MNIST contains 20,000 images from ten digits, whereas Omniglot contains 32,460 images from 1,623 characters. Each image is represented by a composition of 784 pixel intensities. We construct Hellinger nearest-neighbor graphs within digit or character classes, treating class identity as available side information. These applications provide controlled settings for examining sensitivity to localized and distributed image corruption.

The microbiome application comprises NEON surface-water bacterial community profiles and associated sampling metadata (National Ecological Observatory Network (NEON), 2025). Filtering retains 1,494 samples and 3,187 OTUs across 33 sites. The graph combines geographic proximity and collection-time similarity. The human tonsil application comprises 5,778 spatially annotated expression profiles, with 3,234 retained genes and 13 cell types; its graph connects spatial neighbors. In these applications, unusual profiles may arise from technical contamination or genuine ecological and cellular heterogeneity. Their identification can support quality assessment and targeted biological investigation, but does not by itself distinguish artifacts from meaningful variation.

| | Variational Approach | KNN graph | Dirichlet prior | Simplex geometry |
|---|---|---|---|---|
| **RD-GAE** | ✓ | ✓ | ✓ | ✓ |
| G-DirVAE | ✓ | ✓ | ✓ | ✗ |
| G-VAE | ✓ | ✓ | ✗ | ✗ |
| G-AE | ✗ | ✓ | ✗ | ✗ |
| DirVAE | ✓ | ✗ | ✓ | ✗ |
| VAE | ✓ | ✗ | ✗ | ✗ |
| AE | ✗ | ✗ | ✗ | ✗ |

Figure 2: Structural components of RD-GAE and the autoencoder baselines.

Each nonnegative profile is normalized to unit sum after pseudocount addition. We reserve approximately 10% of observations for testing and construct five development folds, stratifying by digit, character, site, or cell type. Contamination is introduced into a selected fraction of profiles through simplex-preserving perturbations, including mass redistribution, spikes in low-abundance components, mixtures, and category rearrangements. These mechanisms represent localized and distributed departures from the original composition; the available mechanisms and their parameters are summarized in Appendix Table 2 . Image experiments introduce known contamination into all partitions, whereas biological experiments introduce contamination only into validation and test profiles.

The latter assess detection of controlled perturbations without assuming that the original biological observations are free of natural anomalies. Detailed preprocessing and graph construction are provided in Appendix D.

We compare RD-GAE with deterministic, Gaussian variational, and Dirichlet variational autoencoders, including their graph-based counterparts. Baselines use either simplex inputs with KL reconstruction loss or ILR-transformed inputs with squared-error loss; Figure 2 summarizes their structural components and modeling capabilities.

### 4.2 Model architecture and training

We use application-specific encoder widths and structurally mirrored decoders, as summarized in Appendix Table 2. Architectures are fixed within each application to facilitate comparison across models. During model training, evaluation and testing, we have assessed the performance in terms of AUROC and area under the precision–recall curve. Training follows three stages. Stage 1 screens 50 randomly sampled hyperparameter configurations using one development fold, with the search space given in Appendix Table 3. Stage 2 evaluates shortlisted configurations, architectural ablations, and baselines through five-fold cross-validation. Stage 3 fits the selected configurations and evaluates their performance on the held-out test observations, reporting variability across fold-specific fits. Early stopping uses validation performance; test outcomes are excluded from model selection. Appendix Table 4 provides the stage-specific settings and computational budget.

### 4.3 Anomaly Detection Performance

Table 1 compares RD-GAE with non-graph and graph-based baseline models across four datasets: MNIST, Omniglot, Microbiome, and SPTR. We evaluate anomaly detection performance using AUROC and AUPRC. RD-GAE achieves the highest mean performance on both metrics across all four datasets, demonstrating consistent improvements over the evaluated baselines across image and biological data.

Among the image dataset, the strongest AUROC improvement is in the MNIST dataset, where RDGAE outperforms G-AE, the best competing model in MNIST, by approximately 0.1, while the strongest AUPRC improvement is in OMNIGLOT, where RDGAE achieves AUPRC of 0.7069, surpassing best baseline method by 0.15. The improvement on image dataset appear in both ROC-based discriminationand precision-recall performance. The biological datasets show similar patterns. RD-GAE achieves an AUROC of 0.9721 and an AUPRC of 0.8689 exceeding VAE, the best baseline on both metrics, by 0.0303 and 0.0684 respectively. On SPTR, RD-GAE achiaves an AUROC of 0.9598, comparing to 0.9331 for G-AE and an AUPRC of 0.7708, compared with 0.6155 for VAE. The strongest baseline models varies across different datasets and metrics, yet RD-GAE maintains the strongest model.

### 4.4 Anomaly Detection Performance

Table 1 compares RD-GAE with the non-graph and graph-based baseline models across MNIST, Omniglot, Microbiome, and SPTR using AUROC and AUPRC. RD-GAE achieves the highest mean performance for both metrics on all four datasets, showing consistent improvement across image and biological applications.

Among the image datasets, RD-GAE attains an AUROC of $0.824$ and an AUPRC of $0.716$ on MNIST, compared with $0.772$ and $0.577$ for G-AE, the strongest baseline on these two metrics. On Omniglot, RD-GAE achieves an AUROC of $0.859$ and an AUPRC of $0.707$, compared with $0.816$ and $0.540$ for G-AE. The improvement is particularly pronounced in AUPRC, indicating stronger precision–recall performance in addition to improved ROC-based discrimination. Similarly, RD-GAE demonstrate superior performance in both genomics applications. On Microbiome, RD-GAE achieves an AUROC of $0.972$ and an AUPRC of $0.869$, compared with $0.942$ and $0.801$ for VAE, the strongest baseline on both metrics. On SPTR, RD-GAE attains an AUROC of $0.960$, compared with $0.933$ for G-AE, and an AUPRC of $0.771$, compared with $0.616$ for VAE. Although the strongest baseline varies across datasets and metrics, RD-GAE consistently provides the best overall anomaly detection performance.

### 4.5 Dual-Order Training and Detection

Figure 3 shows validation and test performance over a range of $\lambda_{\mathrm{det}}$ values for a fixed $\lambda_{\mathrm{tr}}$.

Table 1: Anomaly detection performance of RD-GAE and baseline models across the four datasets. Values are reported as mean with standard deviation in parentheses.

| Model | MNIST | | OMNIGLOT | | Microbiome | | SPTR | |
|---|---|---|---|---|---|---|---|---|
| | AUROC | AUPRC | AUROC | AUPRC | AUROC | AUPRC | AUROC | AUPRC |
| **RD-GAE** ($\lambda_{\text{det}} > \lambda_{\text{tr}}$) | **0.8238** ($\pm$ 0.0268) | **0.7156** ($\pm$ 0.0060) | **0.8594** ($\pm$ 0.0052) | **0.7069** ($\pm$ 0.0116) | **0.9721** ($\pm$ 0.0142) | **0.8689** ($\pm$ 0.0598) | **0.9598** ($\pm$ 0.0007) | **0.7708** ($\pm$ 0.0122) |
| RD-GAE ($\lambda_{\text{det}} = \lambda_{\text{tr}}$) | 0.8066 ($\pm$ 0.0099) | 0.6175 ($\pm$ 0.0161) | 0.7978 ($\pm$ 0.0053) | 0.5059 ($\pm$ 0.0173) | 0.9513 ($\pm$ 0.0223) | 0.8335 ($\pm$ 0.0259) | 0.9257 ($\pm$ 0.0007) | 0.6199 ($\pm$ 0.0019) |
| RD-GAE ($\lambda_{\text{det}} = \lambda_{\text{tr}} = 1$) | 0.7718 ($\pm$ 0.0020) | 0.6026 ($\pm$ 0.0109) | 0.7971 ($\pm$ 0.0047) | 0.4968 ($\pm$ 0.0183) | 0.9488 ($\pm$ 0.0139) | 0.7674 ($\pm$ 0.0752) | 0.9304 ($\pm$ 0.0014) | 0.6456 ($\pm$ 0.0104) |
| AE | 0.7245 ($\pm$ 0.0185) | 0.4619 ($\pm$ 0.0220) | 0.6527 ($\pm$ 0.0073) | 0.3493 ($\pm$ 0.0093) | 0.9200 ($\pm$ 0.0052) | 0.6795 ($\pm$ 0.0221) | 0.9282 ($\pm$ 0.0012) | 0.6124 ($\pm$ 0.0051) |
| VAE | 0.7607 ($\pm$ 0.0006) | 0.5693 ($\pm$ 0.0026) | 0.7697 ($\pm$ 0.0008) | 0.4232 ($\pm$ 0.0021) | 0.9418 ($\pm$ 0.0034) | 0.8005 ($\pm$ 0.0198) | 0.9266 ($\pm$ 0.0009) | 0.6155 ($\pm$ 0.0045) |
| DirVAE | 0.7602 ($\pm$ 0.0001) | 0.5669 ($\pm$ 0.0007) | 0.7302 ($\pm$ 0.0006) | 0.2872 ($\pm$ 0.0008) | 0.8580 ($\pm$ 0.0034) | 0.5683 ($\pm$ 0.0297) | 0.9177 ($\pm$ 0.0052) | 0.5692 ($\pm$ 0.0263) |
| G-AE | 0.7721 ($\pm$ 0.0193) | 0.5774 ($\pm$ 0.0399) | 0.8155 ($\pm$ 0.0025) | 0.5401 ($\pm$ 0.0165) | 0.9307 ($\pm$ 0.0115) | 0.6588 ($\pm$ 0.0159) | 0.9331 ($\pm$ 0.0009) | 0.5639 ($\pm$ 0.0044) |
| G-VAE | 0.7650 ($\pm$ 0.0032) | 0.5752 ($\pm$ 0.0121) | 0.7834 ($\pm$ 0.0009) | 0.4742 ($\pm$ 0.0044) | 0.8798 ($\pm$ 0.0035) | 0.6204 ($\pm$ 0.0277) | 0.9167 ($\pm$ 0.0002) | 0.5623 ($\pm$ 0.0011) |
| G-DirVAE | 0.7608 ($\pm$ 0.0001) | 0.5700 ($\pm$ 0.0002) | 0.7698 ($\pm$ 0.0008) | 0.4230 ($\pm$ 0.0022) | 0.8606 ($\pm$ 0.0010) | 0.6008 ($\pm$ 0.0051) | 0.9163 ($\pm$ 0.0003) | 0.5605 ($\pm$ 0.0008) |

Performance improves initially but deteriorates when $\lambda_{\text{det}}$ becomes too large, motivating selection of $\lambda_{\text{det}}$ based on validation performance. Table 1 further shows that using $\lambda_{\text{det}} > \lambda_{\text{tr}}$ consistently improves performance over using the same order for training and detection. The AUPRC increases from 0.618 to 0.716 on MNIST, 0.506 to 0.707 on Omniglot, 0.834 to 0.869 on Microbiome, and 0.620 to 0.771 on SPTR, with corresponding AUROC improvements across all datasets.

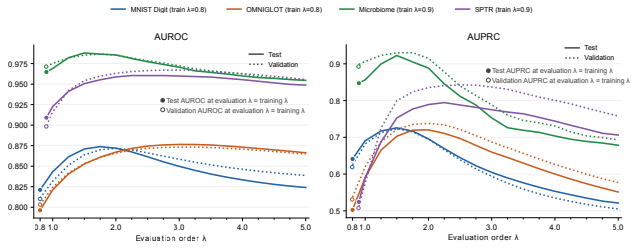



Figure 3: Sensitivity of anomaly detection performance to the detection order $\lambda_{\text{det}}$ for fixed training order $\lambda_{\text{tr}}$.

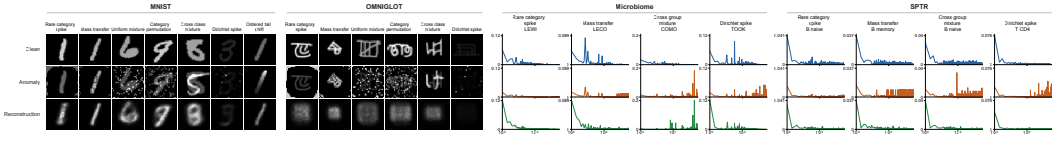



Figure 4: Examples of clean, synthetically contaminated, and RD-GAE reconstructed profiles across the four datasets and different anomaly types.

### 4.6 Reconstruction and Representation

Figure 4 compares the clean profiles, synthetically contaminated profiles, and their RD-GAE reconstructions across the four datasets. RD-GAE suppresses several features introduced by contamination while retaining the main structure of the original observations, although the quality of reconstruction varies across datasets and anomaly types. The discrepancy between the observed and reconstructed profiles is subsequently used for anomaly detection.

Figure 5 compares the observed data with latent representations learned by RD-GAE and G-DirVAE. Since both the observations and latent representations lie on the simplex, we compute pairwise Rényi divergences and use non-metric multidimensional scaling (NMDS) for visualization. While the contaminated observations show clear separation between normal and anomalous samples,

their RD-GAE latent representations show substantial overlap. We further quantify this separation using PERMANOVA. The non-significant PERMANOVA result for RD-GAE indicates no evidence of systematic separation between normal and anomalous latent representations, in contrast to G-DirVAE. This result supports the ability of RD-GAE to learn representations that are less affected by contamination during training.

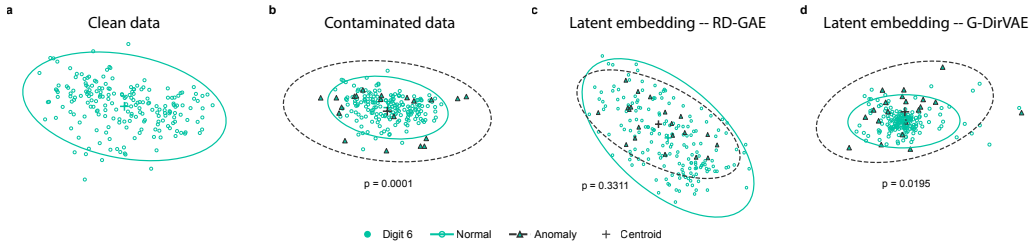



Figure 5: NMDS visualization based on pairwise Rényi divergences for the observed data and latent representations from RD-GAE and G-DirVAE, comparing normal and anomalous samples.

### 4.7 Ablation Study

Table 6 in Appendix shows the contribution of the main architectural components of RD-GAE. Most architectural ablations produce relatively modest and dataset-dependent changes in AUROC and AUPRC, suggesting that no single modification among skip connections, graph propagation, graph weighting, or category-weight constraints dominates performance across all settings.

The strongest deterioration is observed when the proposed escort gate is replaced by the conventional softmax gate, with substantial reductions in both AUROC and AUPRC across all four datasets. A similarly strong effect is observed for the loss-function ablation in Table 1: the KL case, $\lambda_{\mathrm{tr}} = \lambda_{\mathrm{det}} = 1$, performs consistently worse than the proposed dual-order Rényi formulation. Together, these results indicate that both the escort-based gating mechanism and the Rényi-based dual-order objective are important components of RD-GAE.

## 5 Conclusion

We proposed RD-GAE, a graph autoencoding framework that addresses the joint challenges of compositional constraints, neighborhood dependence, and anomaly detection. Unlike unconstrained Euclidean representations, RD-GAE preserves nonnegativity and unit mass throughout its encoder, Dirichlet latent space, and decoder, without requiring a log-ratio reference category. Its gated probability-temperature transformation controls within-profile contrast, while dual-order training and detection separate reconstruction learning from anomaly sensitivity. Across four image and biological applications, RD-GAE achieves the highest mean AUROC and AUPRC among the evaluated baselines; single-order and KL comparisons further support the dual-order formulation. These findings demonstrate the value of integrating simplex-preserving representations, graph information, and order-dependent discrepancies within a unified framework. Generalization beyond the evaluated graphs and synthetic contamination mechanisms requires further validation.

## AI USE STATEMENT

In this work, we used generative AI tools for coding and writing task. LLM-generated code was verified and tested for correctness by 2 authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI. No empirical result was generated or inferred by the system.

## REFERENCES


Haleh Akrami, Anand A. Joshi, Jian Li, Serg"ul Ayd"ore, and Richard M. Leahy. A robust variational autoencoder using beta divergence. *Knowledge-Based Systems*, 238:107886, 2022. doi: 10.1016/j.knosys.2021.107886.

James Atwood and Don Towsley. Diffusion-convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Janez Demsar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

Kaize Ding, Jundong Li, Rohit Bhanushali, and Huan Liu. Deep anomaly detection on attributed networks. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, pp. 594–602, 2019. doi: 10.1137/1.9781611975673.67.

Michael Figurnov, Kirill Struminsky, and Dmitry P. Vetrov. Robust variational inference, 2016.

Futoshi Futami, Issei Sato, and Masashi Sugiyama. Variational inference based on robust divergences. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 813–822, 2018.

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

Geoffrey E. Hinton and Ruslan R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006. doi: 10.1126/science.1127647.

Zhenyu Hou, Xiao Liu, Yukuo Cen, Yuxiao Dong, Hongxia Yang, Chunjie Wang, and Jie Tang. GraphMAE: Self-supervised masked graph autoencoders. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 594–604, 2022. doi: 10.1145/3534678.3539321.

Weonyoung Joo, Wonsung Lee, Sungrae Park, and Il-Chul Moon. Dirichlet variational autoencoder. *Pattern Recognition*, 107:107514, 2020. doi: 10.1016/j.patcog.2020.107514.

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

Thomas N. Kipf and Max Welling. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*, 2016. URL `https://arxiv.org/abs/1611.07308`.

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

Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015. doi: 10.1126/science.aab3050.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Jia Li, Jianwei Yu, Jiajin Li, Honglei Zhang, Kangfei Zhao, Yu Rong, Hong Cheng, and Junzhou Huang. Dirichlet graph variational autoencoder. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

National Ecological Observatory Network (NEON). Surface water microbe community composition (dp1.20141.001), release-2025. Dataset, 2025. URL `https://data.neonscience.org/data-products/DP1.20141.001/RELEASE-2025`.

Kai Wang Ng, Guo-Liang Tian, and Man-Lai Tang. *Dirichlet and Related Distributions: Theory, Methods and Applications*. John Wiley & Sons, 2011. doi: 10.1002/9781119995784.

Min Oh and Liqing Zhang. DeepMicro: Deep representation learning for disease prediction based on microbiome data. *Scientific Reports*, 10:6024, 2020. doi: 10.1038/s41598-020-63159-5.

Oleksandr Shchur, Maximilian Mumme, Aleksandar Bojchevski, and Stephan G"unnemann. Pitfalls of graph neural network evaluation. In *Relational Representation Learning Workshop, Advances in Neural Information Processing Systems*, 2018.

Richard Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *The Annals of Mathematical Statistics*, 35(2):876–879, 1964. doi: 10.1214/aoms/1177703591.

Tim van Erven and Peter Harremo"es. Rényi divergence and Kullback–Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014. doi: 10.1109/TIT.2014.2320500.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

Tianqi Wang, Huitong Zhu, Yunlan Zhou, Weihong Ding, Weichao Ding, Liangxiu Han, and Xueqin Zhang. Graph attention automatic encoder based on contrastive learning for domain recognition of spatial transcriptomics. *Communications Biology*, 7(1):1351, 2024. doi: 10.1038/s42003-024-07037-0.

Chong Zhou and Randy C. Paffenroth. Anomaly detection with robust deep autoencoders. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 665–674. Association for Computing Machinery, 2017. doi: 10.1145/3097983.3098052.

# A Complete Objective Derivation

## A.1 Reconstruction term

Conditional on $\boldsymbol{Z} \sim q_\phi$, the decoder returns positive simplex rows $\widehat{\boldsymbol{x}}_v(\boldsymbol{Z})$. The population reconstruction risk is

$$\mathcal{L}_{\mathrm{rec}}(\phi, \theta) = \frac{1}{N} \mathbb{E}_{q_\phi(\boldsymbol{Z}|\boldsymbol{X},G)} \left[ \sum_{v=1}^{N} D_{\lambda_{\mathrm{tr}}}(\boldsymbol{x}_v \| \widehat{\boldsymbol{x}}_v(\boldsymbol{Z})) \right]. \tag{11}$$

For iid posterior samples $\boldsymbol{Z}^{(1)}, \ldots, \boldsymbol{Z}^{(S)}$, linearity of expectation makes the sample average in Equation (6) unbiased; independence divides the single-sample variance by $S$.

## A.2 Dirichlet prior term

Let the two Dirichlet density functions be $q(\boldsymbol{z}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{j=1}^{d} z_j^{\alpha_j - 1}, p(\boldsymbol{z}) = \frac{1}{B(\boldsymbol{\beta})} \prod_{j=1}^{d} z_j^{\beta_j - 1}$ where $B(\cdot)$ be multivariate beta function. Raising to powers $\lambda_z$ and $1 - \lambda_z$ gives $q(\boldsymbol{z})^{\lambda_z} p(\boldsymbol{z})^{1-\lambda_z} = \frac{1}{B(\boldsymbol{\alpha})^{\lambda_z} B(\boldsymbol{\beta})^{1-\lambda_z}} \prod_{j=1}^{d} z_j^{c_j - 1}$, such that $\boldsymbol{c} = \lambda_z \boldsymbol{\alpha} + (1 - \lambda_z)\boldsymbol{\beta}$. When every $c_j > 0$, integration over $\Delta^{d-1}$ yields

$$\int q(\boldsymbol{z})^{\lambda_z} p(\boldsymbol{z})^{1-\lambda_z} d\boldsymbol{z} = \frac{B(\boldsymbol{c})}{B(\boldsymbol{\alpha})^{\lambda_z} B(\boldsymbol{\beta})^{1-\lambda_z}}.$$

Taking the logarithm and dividing by $\lambda_z - 1$ proves Equation (5). The combined objective is the sum of Equation (11) and the node-mean prior discrepancy weighted by $\gamma$. For computation, Direct Gamma products overflow quickly. Hence, for numerical stability, we compute

$$\log B(\boldsymbol{a}) = \sum_{j=1}^{d} \mathrm{lgamma}(a_j) - \mathrm{lgamma}\left( \sum_{j=1}^{d} a_j \right), \tag{12}$$

and

$$\log I_\lambda(\boldsymbol{p}, \boldsymbol{q}) = \mathrm{logsumexp}_k \left\{ \lambda \log p_k + (1 - \lambda) \log q_k \right\}. \tag{13}$$

The gated decoder ensures $q_k > 0$. If structural zeros occur in $\boldsymbol{p}$ and $\log p_k$ is required, we use masked zero terms or an explicitly defined loss-only smoothing rule, rather than introducing an additional $\varepsilon$ into the network nonlinearity.

The gated output supplies $q_k > 0$. When structural zeros occur in $\boldsymbol{p}$, we replace zero entries by a fixed small constant $\varepsilon > 0$ only for the numerical evaluation of the divergence. This adjustment is restricted to the loss computation and does not modify the network nonlinearity or the decoder output.

## A.3 Escort sharpening and local sensitivity

Let $r = 1/\tau$, $Z = \sum_k u_k^r$, and $E_i = u_i^r / Z$. Cancellation of $Z$ gives $E_i / E_j = (u_i / u_j)^{1/\tau}$. Direct differentiation gives

$$\frac{\partial E_i}{\partial u_j} = r E_i \left( \frac{\mathbf{1}\{i = j\}}{u_i} - \frac{E_j}{u_j} \right).$$

The absolute sum of column $j$ is $2rE_j(1 - E_j)/u_j$, proving Equation (4). The gated Jacobian is $(1 - g)I + gJ_E$, so integration along a segment gives the local upper factor used in Equation (4). Dividing all escort terms by $u_{\max}^{1/\tau}$ proves convergence to the uniform law on the maximizers as $\tau \downarrow 0$. Hence small $\tau$ sharpens odds but cannot provide a uniform positive lower separation bound.

# B Proofs

## B.1 Dual-order separation and its small-order limit

The monotonicity of Rényi divergence in its order is classical (van Erven & Harremo"es, 2014). We give the calculation because the difference of two order derivatives is the certificate used by the proposed detector.

*Proof of Theorem 2.4.* For positive $\boldsymbol{p}, \boldsymbol{q}$, define $Y_k = \log(p_k/q_k)$ and $\psi(t) = \log \sum_k q_k e^{tY_k}$, together $Y = [Y_1, \ldots, Y_k]$. On differentiation of the finite sum, we have

$$\pi_t(k) = \frac{q_k e^{tY_k}}{\sum_j q_j e^{tY_j}}. \qquad \psi'(t) = \mathbb{E}_{\pi_t} Y, \qquad \psi''(t) = \mathrm{Var}_{\pi_t}(Y), \qquad D_t(\boldsymbol{p}\|\boldsymbol{q}) = \frac{\psi(t)}{t-1}.$$

Because $\psi(1) = \log \sum_k p_k = 0$, the numerator of the derivative $\frac{\partial D_t(\boldsymbol{p}\|\boldsymbol{q})}{\partial t}$ obeys

$$\begin{aligned}(\lambda-1)\psi'(\lambda) - \psi(\lambda) &= \int_1^\lambda \frac{d}{dt}\{(t-1)\psi'(t) - \psi(t)\}\, dt \\ &= \int_1^\lambda (t-1)\psi''(t)\, dt.\end{aligned}$$

Division by $(\lambda-1)^2$ proves, for $\lambda > 1$,

$$\frac{\partial}{\partial\lambda} D_\lambda(\boldsymbol{p}\|\boldsymbol{q}) = \frac{1}{(\lambda-1)^2}\int_1^\lambda (t-1)\,\mathrm{Var}_{\pi_t}(Y)\, dt \geq 0. \tag{14}$$

Subtracting the formula for the normal pair from that for the anomalous pair yields

$$\Delta'(\lambda) = \frac{1}{(\lambda-1)^2}\int_1^\lambda (t-1)\{\mathrm{Var}_{\pi_t^A}(Y^A) - \mathrm{Var}_{\pi_t^N}(Y^N)\}\, dt.$$

The stated pointwise variance dominance makes the integrand nonnegative and therefore $\Delta'(\lambda) \geq 0$. If dominance is strict on a subset of positive measure inside a subinterval ending at $\lambda$, the integral is strictly positive at that $\lambda$. □

*Proof of Lemma 2.2.* With the same $\psi$, $\psi(0) = 0$ and $\psi'(0) = \sum_k q_k \log(p_k/q_k) = -\mathrm{KL}(\boldsymbol{q}\|\boldsymbol{p})$. Taylor expansion around zero therefore gives $\psi(\lambda) = -\lambda\,\mathrm{KL}(\boldsymbol{q}\|\boldsymbol{p}) + O(\lambda^2)$. Since $(\lambda - 1)^{-1} = -(1 + \lambda + O(\lambda^2))$, the first expansion in the following display follows. Now let $q_j = \exp(\xi_j)/\sum_i \exp(\xi_i)$. Since $\partial q_j/\partial\xi_k = q_j(1\{j=k\} - q_k)$, direct differentiation gives

$$\frac{\partial}{\partial\xi_k}\mathrm{KL}(\boldsymbol{q}\|\boldsymbol{p}) = q_k\left[\log\frac{q_k}{p_k} - \sum_j q_j \log\frac{q_j}{p_j}\right] = -q_k\left[Y_k - \sum_j q_j Y_j\right].$$

Differentiating the finite-dimensional Taylor expansion establishes

$$D_\lambda(\boldsymbol{p}\|\boldsymbol{q}) = \lambda\,\mathrm{KL}(\boldsymbol{q}\|\boldsymbol{p}) + O(\lambda^2), \qquad \frac{\partial D_\lambda}{\partial\xi_k} = -\lambda q_k\left(Y_k - \sum_j q_j Y_j\right) + O(\lambda^2). \tag{15}$$

Thus the objective is not merely smaller for tiny orders: its parameter-learning signal vanishes to first order. □

The condition above is sufficient, not necessary. It is to be evaluated on validation anomalies, and it does not license order tuning on the test labels. It also explains why an arbitrarily large detection order is unsafe: the tilted distribution can collapse onto a single extreme density ratio, making the score high variance even while its population value is monotone.

## C PROOFS FOR THEORITICAL PROPERTIES

Let us fix $N$ nodes, a reference input $\boldsymbol{X}$ with rows in $\Delta^{K-1}$, the fitted parameters, and the graph operators. For a contaminated set $C \subseteq V$, we define

$$\mathfrak{X}_C(\boldsymbol{X}) = \left\{\widetilde{\boldsymbol{X}} : \widetilde{\boldsymbol{x}}_v = \boldsymbol{x}_v \text{ for } v \notin C, \quad \widetilde{\boldsymbol{x}}_v \in \Delta^{K-1} \text{ for } v \in C\right\}.$$

This set allows arbitrary simplex-valued replacements at nodes in $C$ and preserves the original profiles at all other nodes. We consider $M$ sequential encoder–decoder layers, $\Phi_r(\boldsymbol{H}) = T_{g_r,\tau_r}(\boldsymbol{P}_r\boldsymbol{H}\boldsymbol{W}_r)$, and insert the posterior-mean map $m$ between the encoder and decoder, without bypass skips. Each $\boldsymbol{P}_r$ incorporates any fixed residual graph mixing, and we set $\boldsymbol{P}_r = \boldsymbol{I}_N$ for graph-free layers. The nonnegative, row-stochastic operators $\boldsymbol{P}_r$ and $\boldsymbol{W}_r$ preserve nonnegativity and unit row sums during graph aggregation and category mapping. The gate and posterior-mean map preserve these properties as well. Together, this ensures that every intermediate state and reconstruction lies on its corresponding simplex under every allowed replacement.

### C.1 Uniform sensitivity bound for the node-local escort gate

At layer $r$, let $\mathcal{K}_r$ be a convex subset of the simplex interior containing all preactivation rows generated by the reference input and the allowed replacements. Using the Jacobian bound from Appendix A.3, define

$$L_r = (1 - g_r) + g_r \sup_{\boldsymbol{u} \in \mathcal{K}_r} \max_j \frac{2E_{\tau_r}(\boldsymbol{u})_j \big(1 - E_{\tau_r}(\boldsymbol{u})_j\big)}{\tau_r u_j}.$$

If this supremum is finite, then for any $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{K}_r$,

$$\|T_{g_r,\tau_r}(\boldsymbol{u}) - T_{g_r,\tau_r}(\boldsymbol{v})\|_1 \le L_r \|\boldsymbol{u} - \boldsymbol{v}\|_1,$$

since the line segment joining $\boldsymbol{u}$ and $\boldsymbol{v}$ remains in $\mathcal{K}_r$ by convexity. For a fixed strictly positive category matrix, such a domain can be constructed explicitly. Let $\delta_r = \min_{i,j}(W_r)_{ij} > 0$. Because each row of $\boldsymbol{P}_r\boldsymbol{H}$ is a probability vector, every coordinate of $\boldsymbol{P}_r\boldsymbol{H}\boldsymbol{W}_r$ is a convex combination of entries from the corresponding column of $\boldsymbol{W}_r$ and is therefore bounded below by $\delta_r$. Hence we have

$$\mathcal{K}_r = \left\{ \boldsymbol{u} \in \Delta^{k_{r+1}-1} : u_j \ge \delta_r \text{ for all } j \right\}.$$

Using $E_j(1 - E_j) \le 1/4$ gives

$$L_r \le (1 - g_r) + \frac{g_r}{2\tau_r \delta_r},$$

which provides a uniform, but conservative, Lipschitz bound.

### C.2 Posterior-mean contraction

For $\boldsymbol{h} \in \Delta^{d-1}$, the total Dirichlet concentration is $a_0 = d\alpha_{\min} + \kappa$, so

$$m(\boldsymbol{h}) = \frac{\alpha_{\min}\mathbf{1}_d + \kappa \boldsymbol{h}}{a_0}.$$

For any two simplex vectors $\boldsymbol{h}$ and $\widetilde{\boldsymbol{h}}$,

$$m(\boldsymbol{h}) - m(\widetilde{\boldsymbol{h}}) = a(\boldsymbol{h} - \widetilde{\boldsymbol{h}}), \qquad a = \frac{\kappa}{a_0} \in (0, 1),$$

and therefore the posterior-mean map contracts $\ell_1$ differences exactly by the factor $a$. The map acts independently on each node and thus contributes the identity graph operator.

*Proof of Theorem 3.1.* For a signed row vector $\boldsymbol{b}$ and a nonnegative row-stochastic category matrix $\boldsymbol{W}$, using triangle inequality we have

$$\|\boldsymbol{b}\boldsymbol{W}\|_1 = \sum_j \left| \sum_i b_i W_{ij} \right| \le \sum_i |b_i| \sum_j W_{ij} = \|\boldsymbol{b}\|_1.$$

Let us consider corresponding input states $\boldsymbol{H}, \widetilde{\boldsymbol{H}}$ of one layer and define $d_u = \|\boldsymbol{h}_u - \widetilde{\boldsymbol{h}}_u\|_1$. The uniform gate inequality and the preceding contraction imply

$$\begin{aligned} d_v^+ &\le L_r \left\| \sum_u (P_r)_{vu} (\boldsymbol{h}_u - \widetilde{\boldsymbol{h}}_u) \boldsymbol{W}_r \right\|_1 \\ &\le L_r \sum_u (P_r)_{vu} \|(\boldsymbol{h}_u - \widetilde{\boldsymbol{h}}_u)\boldsymbol{W}_r\|_1 \le L_r \sum_u (P_r)_{vu} d_u. \end{aligned}$$

Putting together componentwise, we can say then $\boldsymbol{d}^+ \le L_r \boldsymbol{P}_r \boldsymbol{d}$. At the bottleneck, the distance vector $\boldsymbol{d}^+$ is multiplied exactly by $a$. Putting together all layers, we have

$$\boldsymbol{d}^{\text{out}} \le a \left( \prod_r L_r \right) \boldsymbol{P}_{M-1} \cdots \boldsymbol{P}_0 \boldsymbol{d}^{\text{in}} = B\mathcal{A}\boldsymbol{d}^{\text{in}}.$$

For each layer, summing its bound over output nodes also gives

$$\sum_v d_v^+ \le L_r \sum_u \left( \sum_v (P_r)_{vu} \right) d_u \le L_r c(\boldsymbol{P}_r) \sum_u d_u.$$

Composition, including the factor $a$, proves $\sum_v d_v^{\text{out}} \le \Gamma \sum_u d_u^{\text{in}}$. Two simplex vectors have distance at most two. If replacements are confined to $C$, then $\boldsymbol{d}^{\text{in}} \le 2\mathbf{1}_C$ and $\sum_u d_u^{\text{in}} \le 2|C|$. Substitution proves both inequalities in Equation 9.

The matrix $\mathcal{A}$ is row-stochastic, so $0 \le [\mathcal{A}\mathbf{1}_C]_v \le 1$. Each entry $\mathcal{A}_{vu}$ is a sum of products of propagation weights along allowed routes from input $u$ to output $v$. If the operators use only edges and self-loops, no such route exists when the graph distance from $C$ to $v$ exceeds $M$; thus exposure is zero there. Finally, double stochasticity gives $c(\boldsymbol{P}_r) = 1$ at every layer, and hence $\Gamma = B$. If every $L_r \le 1$, then $\Gamma \le a$. □

### C.3 Finite-sample reconstruction breakdown

For $0 < t \le 2$, we define

$$R_C = \sup_{\widetilde{\boldsymbol{X}} \in \mathfrak{X}_C(\boldsymbol{X})} \frac{1}{N - |C|} \sum_{v \notin C} \|F_G(\widetilde{\boldsymbol{X}})_v - F_G(\boldsymbol{X})_v\|_1, \qquad C \subsetneq V.$$

If some proper subset has $R_C \ge t$, set $\epsilon_t^{\text{pred}} = N^{-1} \min\{|C| : C \subsetneq V, R_C \ge t\}$; otherwise set $\epsilon_t^{\text{pred}} = 1$. A clean node here means an unreplaced node; its reference reconstruction need not be accurate.

*Proof of Theorem 3.3.* Fix $C \subsetneq V$, put $m = |C|$ and $\epsilon = m/N < 1$. For every admissible replacement, discard the nonnegative contributions of corrupted output indices from the total displacement, and apply Theorem 3.1:

$$\begin{aligned}\frac{1}{N-m} \sum_{v \notin C} \|F_G(\widetilde{\boldsymbol{X}})_v - F_G(\boldsymbol{X})_v\|_1 &\le \frac{1}{N-m} \sum_{v \in V} \|F_G(\widetilde{\boldsymbol{X}})_v - F_G(\boldsymbol{X})_v\|_1 \\ &\le \frac{2\Gamma m}{N-m} = \frac{2\Gamma\epsilon}{1-\epsilon}.\end{aligned}$$

The same $\Gamma$ applies to every allowed replacement by the uniform domain assumption. Taking the supremum therefore leaves this upper bound unchanged. A breaking set must satisfy

$$t \le R_C \le \frac{2\Gamma\epsilon}{1-\epsilon}, \qquad \text{hence} \qquad \epsilon \ge \frac{t}{2\Gamma + t}.$$

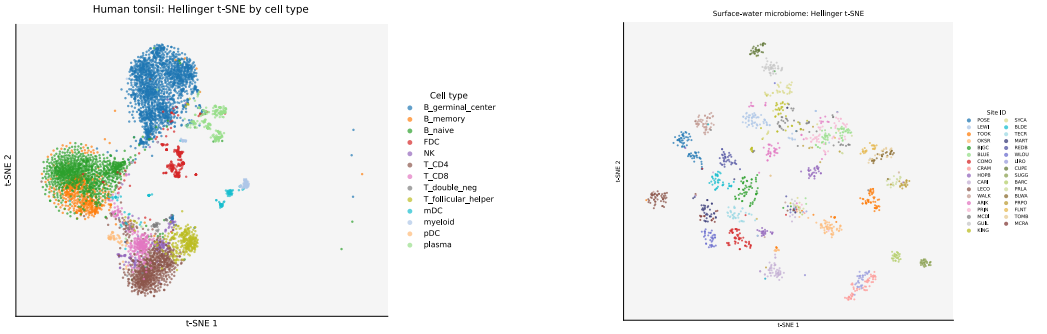
Since $m$ is an integer, it must be at least $\lceil Nt/(2\Gamma + t) \rceil$. Minimizing over breaking sets proves the claimed bound. If there is no breaking set, the convention gives one, which still satisfies the bound since $\lceil Nt/(2\Gamma + t) \rceil \le N$. Under double stochasticity and $L_r \le 1$, we have $\Gamma \le a$. The function $x \mapsto t/(2x + t)$ is decreasing, so substitution gives $\epsilon_t^{\text{pred}} \ge t/(2a + t)$. □

## D Data Preparation and Graph Construction

**Image data.** For MNIST, we randomly select 2,000 observations from each digit in the original training collection. For Omniglot, we combine the background and evaluation collections, which contain 1,623 characters from 50 alphabets, with 20 observations per character. This combined collection is repartitioned for anomaly detection rather than used for the original few-shot recognition task. Omniglot images are resized to $28 \times 28$ and intensity-inverted so that larger values represent foreground strokes. Images are flattened and normalized, yielding $20{,}000 \times 784$ and $32{,}460 \times 784$ data matrices, respectively.

**Microbiome data.** The bacterial count table is aligned with NEON metadata by DNA sample identifier. The metadata provide site identity, collection date, and geographic coordinates. OTUs are retained using fixed thresholds on prevalence, mean abundance, variance, and dispersion, where dispersion is the variance-to-mean ratio. The resulting count matrix contains 1,494 observations and 3,187 OTUs. This filtering avoids specifying an arbitrary number of retained features.

**Spatial expression data.** The human tonsil expression matrix is supplied with genes in rows and observations in columns. Expression profiles are matched to spatial coordinates and cell-type annotations by observation identifier. Supplied log-transformed expression values are returned to the linear scale before feature selection and normalization. Genes are filtered using fixed prevalence, mean, variance, and dispersion thresholds, yielding 5,778 observations and 3,234 genes. The retained observations represent 13 annotated cell types.




(a) Spatial transcriptomics data from human tonsil tissue at various anatomical locations.

(b) Surface water microbiome data from the NEON study at various locations across the United States.

Figure 6: t-SNE visualization of genomics and metagenomics application datasets.

**Simplex representation and splitting.** For a retained nonnegative profile $\boldsymbol{c}_i$, we define $x_{ij} = (c_{ij}+\epsilon)/\sum_{k=1}^{p}(c_{ik}+\epsilon)$, with $\epsilon = 10^{-6}$. Splits are fixed before synthetic contamination is introduced. Validation folds partition the development observations, and the test observations are excluded from every training and validation fold. Stratification preserves the representation of digits, characters, sampling sites, or cell types; these labels are distinct from the synthetic anomaly indicators.

**Image graphs.** Image similarity is measured by the Hellinger distance, $H(\boldsymbol{x}_i, \boldsymbol{x}_j) = \|\sqrt{\boldsymbol{x}_i} - \sqrt{\boldsymbol{x}_j}\|_2/\sqrt{2}$. For each observation, we select up to 20 nearest neighbors within its digit or character class and symmetrize the binary adjacency by taking the union of directed edges. The MNIST clean-graph analysis uses original profiles and therefore represents an oracle graph condition; the observed-graph analysis uses contaminated profiles. Omniglot uses observed profiles. Since each character has only 20 observations, its within-character graph connects all 19 other observations; consequently, the binary topology does not depend on their distance ranking.

**Biological graphs.** For microbiome samples, the affinity between observations $i$ and $j$ is $a_{ij} = \exp\{-d_{ij}^2/(2\sigma_s^2) - |t_i - t_j|/\tau_t\}$, where $d_{ij}$ is the Haversine distance in kilometers and $|t_i - t_j|$ is the collection-time difference in days. The bandwidths $\sigma_s$ and $\tau_t$ are the medians of the respective positive pairwise distances. We retain the 20 strongest affinities per observation and symmetrize using $\max(a_{ij}, a_{ji})$. Neither microbial counts nor anomaly indicators enter this construction. For spatial transcriptomics, we construct a symmetrized binary 20-nearest-neighbor graph from Euclidean distances between the supplied spatial coordinates, without using expression or cell-type labels. Raw graphs exclude self-loops; self-loops and the specified normalization are applied during model fitting.

**Interpretation of anomalies.** Synthetic image perturbations alter the distribution of pixel mass, whereas biological perturbations modify the relative abundance of selected taxa or genes. In environmental data, naturally unusual profiles may reflect community changes or sample-processing artifacts. In tissue data, they may reflect localized cellular programs or technical effects. Accordingly, synthetic detection performance establishes sensitivity to the specified perturbations, not biological validation of naturally occurring outliers. Flagged biological profiles require investigation in their ecological or tissue context.

## E Application Settings, Experimental Search and Ablation study

M, O, B, and S denote MNIST, Omniglot, Microbiome, and spatial transcriptomics. Tables 3–5 summarize the application configurations; fixed architecture choices are separated from randomized optimization parameters.

Table 2: Application settings. Decoder widths reverse the encoder; $p$ is the retained feature count. Mechanism acronyms are defined below.

| Data | Encoder / reference | Contamination | Graph |
|---|---|---|---|
| M | 784–200–200–50; (Figurnov et al., 2016) | 10% train/val/test; $\rho = 0.50$. RS, MT, UM, CP, CM, DS, OT. | Within-digit Hellinger 20-NN; clean and observed profiles give separate binary graphs. |
| O | 784–200–200–50; (Figurnov et al., 2016) | Nominal 10% per character; $\rho = 0.70$. RS, MT, UM, CP, CM, DS. | Observed-profile within-character Hellinger $k$-NN; $k = \min(20, n_c - 1)$. |
| B | $p$–128–64; DeepMicro-inspired adaptation (Oh & Zhang, 2020) | 0% injected in train; 10% val/test; $q_H = 0.50$. RS, LM, GM, DS. | 20 strongest space–time kernel edges per node; symmetrize by maximum. |
| S | $p$–256–64; GAAEST widths (Wang et al., 2024) | 0% injected in train; 10% val/test; $q_H = 0.10$. RS, LM, GM, DS. | Euclidean 20-NN using spatial coordinates; binary undirected union. |

RS: rare-category spike; MT: mass transfer; UM: uniform-support mixture; CP: category permutation; CM: cross-class mixture; DS: Dirichlet spike; OT: ordered tilt; LM: low-mass mixture; GM: cross-group mixture. Widths are fixed; the Microbiome width sequence is an adaptation, not the published DeepMicro SAE. Image graph labels are observed side information. At $n_c = 20$, Omniglot character subgraphs are complete. Omniglot's nominal rate requires confirmation from the saved payload because the local preprocessing call and injector signature differ.

Table 3: Randomized search space. Sets are sampled uniformly; LogU denotes log-uniform sampling. Values are shared unless an application is named.

| Parameter | Randomized values |
|---|---|
| Graph mixing $\eta_\ell$ | $\{0.05, 0.1\}$ |
| Activation temperature $\tau$ | $\{0.01, 0.02, 0.03, 0.05\}$ |
| Gate $g$ | $\{0.95, 0.99\}$ |
| Dropout | $\{0, 0.1\}$ |
| Learning rate (Adam) | $\mathrm{LogU}[0.002, 0.01]$ |
| Weight decay | M/O/S: $\{10^{-5}, 10^{-4}\}$; B: $\{10^{-5}, 10^{-3}\}$ |
| Training order $\lambda_{\mathrm{tr}}$ | $\{0.8, 0.9\}$ |
| Prior order $\lambda_z$ | $\{0.5, 0.8\}$ |
| RD-GAE prior weight $\gamma$ | $\mathrm{LogU}[10^{-5}, 5 \times 10^{-4}]$ |
| Variational baseline prior weight | $\mathrm{LogU}[10^{-4}, 1]$; first candidate fixed at 1 |

Detection order is selected after fitting from $\{1.5, 2, 3, 4, 5, 7, 10, 14, 19, 26, 36, 50\}$ using the same reconstruction draws. Validation average precision (AP, reported as AUPRC) is primary, AUROC breaks ties, and the smaller order breaks remaining ties. This score sweep does not add training fits. The saved validation/test error is $1 - \mathrm{AP}$.

The fixed test set contains approximately 10% of observations; five-fold development splits give approximately 72% training and 18% validation. RD-GAE warm-up and early-stopping patience are each 10% of the epoch cap (3, 5, and 20 epochs), with validation checked each epoch; baselines currently run to the cap. Architectural ablations and baselines use their own Stage-2 winners; final controls inherit the RD-GAE winner. Stage-3 means and standard deviations describe five fitted folds evaluated on the same test set.

Table 4: Three-stage protocol per application and graph condition. A fit is one candidate trained on one fold; counts exclude score-only order selection.

| Stage | Candidates, folds, and training budget |
|---|---|
| 1: Screen | Train 50 randomized RD-GAE candidates on fold 0 for at most 30 epochs (50 fits). Retain the top 5 for M/O and 6 for B/S. |
| 2: Select | Cross-validate each retained candidate and four architectural ablations on five folds, together with 12 baselines $\times$ 3 candidates (M/O) or 6 baselines $\times$ 6 candidates (B/S). At most 50 epochs: $5[5(1+4)+12(3)] = 305$ M/O fits; $5[6(1+4)+6(6)] = 330$ B/S fits. |
| 3: Evaluate | Refit the RD-GAE winner, four architectural winners, baseline winners, and four final controls on five folds, with fresh initialization and at most 200 epochs. Inductive: $5(1+4+12+4) = 105$ M/O fits or $5(1+4+6+4) = 75$ B/S fits; transductive adds 5 fits. |

For B, edge weights before sparsification are $a_{ij} = \exp\{-d_{ij}^2/(2\sigma_s^2) - |t_i - t_j|/\tau_t\}$, where $d_{ij}$ is Haversine distance and both bandwidths are positive pairwise distance medians. Graphs are normalized after adding self-loops; inductive and transductive conditions are evaluated separately. Genomics labels denote synthetic perturbations, not independently verified biological anomalies.

Table 5: Contamination operators and fixed design values. Application assignments appear in Table 2; $M_\rho(x, b) = (1-\rho)x + \rho b$, $u_T = \mathbf{1}_T/|T|$, and $R$ contains the lowest-mean 10% of features.

| Type | Mathematical form | Parameters / support |
|---|---|---|
| RS | $x' = M_\rho(x, u_T)$, $T \subseteq R$. | $\lvert T\rvert = \min(\lvert R\rvert, \lceil 0.10p\rceil)$. |
| MT | $x' = x - \rho(x \odot \mathbf{1}_A) + \rho(\sum_{j\in A} x_j)u_T$. | $A$: largest $\lceil 0.02p\rceil$ entries; $T \subseteq R \setminus A$, $\lvert T\rvert \le \lvert A\rvert$. |
| LM | $x' = M_{\rho_i}(x, u_T)$. | $\lvert T\rvert = \lceil 0.02p\rceil$, drawn from row-wise lowest-mass features. |
| UM | $x' = M_\rho(x, u_T)$. | $\lvert T\rvert = \lceil 0.10p\rceil$; random support across all features. |
| CP | $x' = M_\rho(x, \Pi x)$. | Uniform random category permutation $\Pi$. |
| CM | $x' = M_\rho(x_i, x_j)$, $c_j \neq c_i$. | Random different-class donor in the same partition. |
| GM | $x' = M_{\rho_i}(x_i, x_j)$. | Farthest of up to 64 reference donors, preferably from another group. |
| DS | $x' = M_\rho(x, b)$; $b_T \sim \mathrm{Dir}(0.1\mathbf{1})$, $b_{T^c} = 0$. | Random support: $\lceil 0.30p\rceil$ (M/O), $\lceil 0.10p\rceil$ (B/S). |
| OT | $x'_j = \dfrac{x_j e^{s\rho c a_j}}{\sum_k x_k e^{s\rho c a_k}}$. | $s = \pm 1$, $c = 3$, $\rho = 0.50$; $a_j$ evenly spaced on $[-1, 1]$. |

M/O use fixed $\rho = 0.50/0.70$. B/S select $\rho_i \in [0, 1]$ by bisection to match $\min\{h_q, H(x_i, b_i)\}$, where $H$ is Hellinger distance and $h_q$ is its reference-edge quantile ($q = 0.50/0.10$). Validation references use the training fold; test references use the development set. Support sizes are clipped to available features. MT falls back to non-source features if needed; LM samples from the lowest $\max(\lceil 0.02p\rceil, \lceil 0.10p\rceil)$ entries. These are fixed perturbation controls, not randomized model-search parameters.

Table 6: RD-GAE ablation study: Comparison of default RD-GAE with various ablations in its archicture including SkipLink (partially connecting decoder with encoder via $(1-\omega)\mathbf{H}_{\text{d}} + \omega\mathbf{H}_{\text{e}}$), DualGraph (graph operator in both encoder and decoder), GraphFree ($\boldsymbol{P} = \boldsymbol{I}$), RobustSwitch (robust graph weights changed to normal weights), FreeWeight (unconstrained weights $\mathbf{W}$), and Gated softmax ($T_{g,\tau}(\boldsymbol{u}) = (1-g)\boldsymbol{u} + \text{softmax}(\boldsymbol{u}/\tau)$).

| **Ablation** | **MNIST** | | **Omniglot** | | **Microbiome** | | **SPTR** | |
|---|---|---|---|---|---|---|---|---|
| | AUROC | AUPRC | AUROC | AUPRC | AUROC | AUPRC | AUROC | AUPRC |
| **RD-GAE Default** | 0.8238 (± 0.0268) | 0.7156 (± 0.0060) | 0.8594 (± 0.0052) | 0.7069 (± 0.0116) | 0.9721 (± 0.0142) | 0.8689 (± 0.0598) | 0.9598 (± 0.0007) | 0.7708 (± 0.0122) |
| SkipLink | **0.8287** (± 0.0236) | **0.7158** (± 0.0042) | 0.8598 (± 0.0053) | 0.7057 (± 0.0127) | 0.9659 (± 0.0082) | 0.7483 (± 0.0629) | **0.9602** (± 0.0014) | 0.7783 (± 0.0105) |
| DualGraph | 0.8221 (± 0.0156) | 0.7119 (± 0.0023) | **0.8645** (± 0.0057) | **0.7177** (± 0.0116) | 0.9723 (± 0.0082) | 0.8787 (± 0.0164) | 0.9591 (± 0.0027) | 0.7691 (± 0.0245) |
| GraphFree | 0.8131 (± 0.0066) | 0.7107 (± 0.0026) | 0.8563 (± 0.0038) | 0.6951 (± 0.0120) | 0.9782 (± 0.0074) | 0.8362 (± 0.0638) | 0.9586 (± 0.0033) | 0.7797 (± 0.0103) |
| Robust graph | 0.8138 (± 0.0073) | 0.7032 (± 0.0158) | 0.8604 (± 0.0055) | 0.7044 (± 0.0132) | **0.9787** (± 0.0057) | **0.8872** (± 0.0250) | 0.9600 (± 0.0011) | 0.7687 (± 0.0205) |
| Unconstrained weights | 0.8104 (± 0.0028) | 0.7117 (± 0.0043) | 0.8570 (± 0.0018) | 0.7059 (± 0.0069) | 0.9741 (± 0.0197) | 0.7865 (± 0.0879) | 0.9593 (± 0.0024) | **0.7868** (± 0.0100) |
| Softmax gate | 0.6010 (± 0.0000) | 0.3145 (± 0.0001) | 0.4247 (± 0.0000) | 0.1453 (± 0.0000) | 0.3367 (± 0.0520) | 0.0871 (± 0.0016) | 0.7637 (± 0.0002) | 0.4351 (± 0.0004) |