# RETAIN, THEN CONTRACT: LABEL-FREE GRAPH OOD GENERALIZATION BY ENVIRONMENT CONFUSION

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

## ABSTRACT

Graph-level out-of-distribution (OOD) generalization is critical for deploying graph neural networks in real-world applications where distribution shifts are inevitable. Existing approaches typically require class labels during representation learning to decide which substructures are predictive, limiting their applicability. We propose SCARCE, a label-free representation-learning framework that yields distribution-shift-robust graph representations from an environment partition of the training graphs alone, whether that partition comes from dataset metadata or, when none is available, from clustering a self-supervised embedding. Our method operates in two stages: (1) a variational graph autoencoder is warm-started for graph reconstruction; and (2) an adversarial confusion mechanism forces the encoder to produce representations from which a discriminator cannot distinguish environments, while preserving graph-reconstruction fidelity. We provide theoretical analysis showing that environment confusion contracts environment-discriminative directions in the representation space, which, under regularized downstream classifiers, is equivalent to imposing a stronger directional penalty on environment-associated features. This feature-reweighting mechanism reduces reliance on spurious correlations without erasing information, offering a label-free, continuous relaxation of last-layer reweighting methods. Experiments on a controlled synthetic model, the GOOD-Motif benchmark, and molecular property prediction under assay, scaffold, and size shifts show that SCARCE achieves state-of-the-art OOD generalization among label-free methods, using no class labels during representation learning.

## 1 INTRODUCTION

Machine learning models are routinely deployed on data whose distribution differs from that of their training set, and their tendency to latch onto features that are predictive only in training makes such shifts a primary cause of failure (Geirhos et al., 2020; Gulrajani & Lopez-Paz, 2021). Graph data sharpens the problem: molecules are tested on scaffolds absent from training, and graph size or base-graph structure drifts between collections (Gui et al., 2022; Ji et al., 2023; Li et al., 2025b). A substantial body of work now targets out-of-distribution (OOD) generalization on graphs, typically by extracting an invariant subgraph or enforcing invariance across environments (Wu et al., 2022b; Chen et al., 2022; Miao et al., 2022; Yang et al., 2022a; Wu et al., 2022a; Gui et al., 2023). These methods share a prerequisite that is easy to overlook: they need class labels to decide which substructures are predictive, environment labels to decide which correlations are unstable, or both.

Class labels are the expensive part. In drug discovery a single assay label is the outcome of a wet-lab experiment, and many DrugOOD assays contain only tens of labeled molecules, whereas unlabeled molecular graphs are available by the tens of millions (Ji et al., 2023; Hu et al., 2020; Rong et al., 2020). Environment information, by contrast, is cheap but rough: a benchmark may ship a domain index, a scaffold can be computed although which scaffolds form an environment is a modeling decision, and for proteins or social networks the partition has to be inferred from the data itself (Lin et al., 2022; Yao et al., 2024). A method that learns shift-robust representations from an environment partition alone, with no class labels, and that tolerates a coarse or inferred partition, would therefore cover the common case. This setting remains largely unexplored on graphs: the



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.

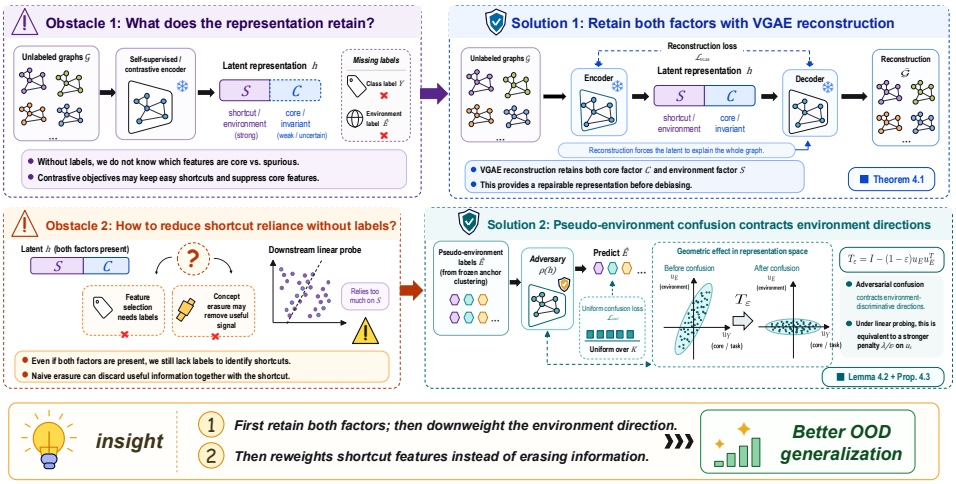


Figure 1: Overview of the SCARCE framework. Given an environment partition of the training graphs (dataset metadata or, when unavailable, clusters of a frozen self-supervised embedding), (1) a VGAE is warm-started for graph reconstruction and (2) an adversarial confusion mechanism forces the encoder to produce representations indistinguishable across environments while preserving reconstruction fidelity.

few existing label-free methods add invariance-inspired augmentations to a contrastive objective (Li et al., 2022b; 2024), and none establishes what the learned representation actually retains.

Removing class supervision raises two obstacles, and each points to one of our contributions. **The first concerns what the representation contains.** Without labels one can neither tell the encoder which features to keep nor check afterwards which it kept, and a contrastive objective offers no guarantee: it is minimized by whichever features most easily discriminate instances and suppresses the rest (Robinson et al., 2021), and a representation that has discarded the core features cannot be repaired downstream. Likelihood objectives behave differently: because the latent must explain the whole input, they keep core features alongside shortcuts (Yang et al., 2022b; Li et al., 2025a; Müller et al., 2024; Qiang et al., 2025). We show that this holds for graph reconstruction without any auxiliary variable: for a variational graph autoencoder (Kipf & Welling, 2016), reconstruction alone forces the latent to retain *both* environment-specific and core factors (Theorem 4.1), and both remain linearly accessible. **The second obstacle concerns what to do with them.** No class label says which factor is spurious, and erasing environment information outright discards useful signal along with the shortcut (Belrose et al., 2023; Lin et al., 2024). We show that neither selection nor erasure is needed: it suffices to *shrink* the environment-associated directions, because any $\ell_2$-regularized downstream classifier then finds them expensive to use and shifts its weight onto the core features. We realize this shrinkage with an adversarial confusion loss against the training environments, which asks nothing of the partition beyond that it separate the spurious factor, and call the framework SCARCE (**S**hortcut **C**ontraction via **A**dversarial **R**epresentation **C**ompression for **E**nvironments).

Our contributions are as follows.

① **Retention by reconstruction.** We prove that VGAE reconstruction alone forces the latent to retain both environment-specific and core features (Theorem 4.1), and verify that both are linearly accessible.

② **Contraction as a ridge penalty.** We propose a confusion loss that scales down environment-discriminative directions rather than erasing them, prove that this contraction (Lemma 4.2) is exactly equivalent to an anisotropic ridge penalty on the downstream probe (Proposition 4.3), and characterize exactly when it helps.

③ **Empirical validation.** Empirically, SCARCE attains state-of-the-art label-free OOD accuracy on GOOD-Motif and is best in 17 of 18 DrugOOD settings, where no environment annotation exists and clustered pseudo-environments suffice; a controlled synthetic study confirms the two preconditions identified by the theory.

Section 2 positions SCARCE against supervised and contrastive alternatives, Section 3 and Section 4 state the setting and the two guarantees, and Section 5 tests the mechanism from a synthetic model up to DrugOOD.

## 2 RELATED WORK

**OOD generalization with labeled supervision.** Most OOD methods prevent a model from relying on features whose correlation with the label is unstable (Geirhos et al., 2020; Gulrajani & Lopez-Paz, 2021): invariant risk minimization (Arjovsky et al., 2019), group-robust optimization (Sagawa et al., 2020), and domain-adversarial alignment (Ganin et al., 2016) use environment or group annotations, and environment inference (Creager et al., 2021) substitutes class labels for them. **In every case some form of supervision decides which correlations are unstable.**

**Graph OOD generalization.** Graph OOD methods either extract an invariant subgraph whose relation to the label is stable (Wu et al., 2022b; Chen et al., 2022; Miao et al., 2022; Li et al., 2022a; Yang et al., 2022a; Chen et al., 2023) or enforce invariance across given or generated environments (Wu et al., 2022a; Gui et al., 2023); in both cases a class label decides which substructures are predictive. **SCARCE uses no class labels during representation learning**; it needs only an environment partition, which may be given or inferred, and it does not select subgraphs.

**Unsupervised graph representation learning.** Unsupervised graph representation learning is mature (Sun et al., 2020; You et al., 2020; Hassani & Khasahmadi, 2020; Hou et al., 2022), and unsupervised domain generalization has been formulated for images (Zhang et al., 2022), but their intersection on graphs is thin. RGCL (Li et al., 2022b) adds rationale-inspired augmentations to a contrastive objective, and OOD-GCL (Li et al., 2024), the closest work to ours, learns a channel-wise disentangled contrastive encoder and penalizes each factor's contrastive loss for varying across inferred environments. Both are contrastive at the core, which suppresses features not needed to tell instances apart (Robinson et al., 2021), and both make environment information invariant; SCARCE instead trains by reconstruction, which provably retains core and environment factors alike (Theorem 4.1), and contracts environment directions rather than erasing them (Section 4.3). When no environment annotation is available we borrow the construction of Yao et al. (2024), who cluster a frozen self-supervised embedding in place of environment labels but then train a supervised graph OOD model; we use the same pseudo-environments inside a label-free objective (Section 5.3).

**Feature reweighting and concept erasure.** Two lines of work act on a representation after it is learned. Last-layer retraining (Kirichenko et al., 2023) re-fits the classifier on group-balanced data, reweighting core over spurious features at the cost of group labels. Linear concept erasure (Ravfogel et al., 2022; Belrose et al., 2023) removes every linearly decodable trace of a protected attribute, discarding task-useful environment signal as well (Lin et al., 2024), and the probes that certify removal are themselves unreliable (Kumar et al., 2022). **SCARCE sits between the two:** it reweights rather than erases, acts in representation space rather than on the classifier, and, since the contracted direction is chosen by an adversary against environments, needs neither group-balanced labeled data nor a protected attribute.

## 3 PROBLEM SETTING

We consider graph-level *unsupervised domain generalization*: learning graph representations that are robust to distribution shifts, without access to class labels during representation learning.

Let the training set contain $N$ graphs $\{G_1, G_2, \ldots, G_N\}$, where each graph $G_i = (\mathbf{X}_i, \mathbf{A}_i)$ consists of a node feature matrix $\mathbf{X}_i \in \mathbb{R}^{n_i \times d_{\text{in}}}$ and an adjacency structure $\mathbf{A}_i$. The training graphs are drawn from a mixture of environments (domains), and we seek representations that generalize to unseen test environments. Each training graph carries an environment index $e_i \in \{1, \ldots, K\}$. The partition may be supplied by the dataset (a domain index) or, absent such metadata, inferred without labels by clustering a self-supervised embedding (Section 5.3); nothing below depends on which.

**Latent generative factors.** We posit that each graph is generated from three latent factors through an unknown map $g$,

$$G = g(C, S, \xi), \tag{1}$$

where $C$ is the *core* (invariant) factor that determines the label $Y$ in every environment, $S$ is the *environment* (spurious) factor whose correlation with $Y$ varies across environments $e$, and $\xi$ collects residual variation independent of $(C, S)$. A predictor that relies on $S$ can succeed in training environments where the $S$–$Y$ correlation is strong yet fail on test environments where it weakens or reverses, so $C$ is the only reliable predictor of $Y$. We write $Z = f_\theta(G)$ for the latent representation produced by a graph encoder $f_\theta$, viewed as a random variable, and $\mathbf{h}_G \in \mathbb{R}^d$ for its value on a given graph $G$. For a reconstruction-based encoder–decoder pair, $L_{\text{rec}} = \mathbb{E}[-\log p_\psi(G \mid Z)]$ denotes the expected reconstruction loss (negative log-likelihood, in nats), i.e., the reconstruction term of the evidence lower bound.

**Encoder.** We encode graphs with message-passing GNNs (Xu et al., 2019; Kipf & Welling, 2017): $L$ layers of neighborhood aggregation produce node embeddings $\mathbf{h}_v^{(L)}$, and the graph-level latent is obtained by sum pooling, $\mathbf{h}_G = \sum_{v \in G} \mathbf{h}_v^{(L)}$. GIN is our default backbone; we also report GCN.

**Objective.** Let $e \in \mathcal{E}$ index the environments. The goal is to learn a graph encoder $f_\theta$ whose representations, when paired with a downstream classifier $c_\phi$, minimize the worst-case risk across all environments including unseen ones:

$$\min_{\theta} \sup_{e \in \mathcal{E}_{\text{all}}} R^e(\theta), \qquad R^e(\theta) = \mathbb{E}_{(G,Y)\sim P^e}[\ell(c_\phi(f_\theta(G)),\, Y)]\,. \tag{2}$$

We call this setting *label-free*: no class label $Y$ is available during representation learning, the environment index $e$ is the only side information, and it may itself be inferred. Class labels are used *only* for fitting a linear probe at evaluation time. This constraint rules out existing invariant learning methods (Arjovsky et al., 2019; Wu et al., 2022a; Chen et al., 2022), which need class labels to decide which correlations are unstable.

## 4 Method

We introduce SCARCE, a framework for graph-level unsupervised domain generalization. The key insight is that a standard VGAE learns representations encoding *all* predictive variation, including environment-specific features that may serve as spurious shortcuts. SCARCE addresses this by augmenting the VGAE objective with an adversarial confusion mechanism that selectively suppresses environment-discriminative directions in the representation, while preserving graph-reconstruction fidelity. An overview is presented in Figure 1 and Algorithm 1.

### 4.1 What Features Does a VGAE Retain?

Our method relies on a VGAE whose latent encodes *both* generative factors: without $S$ there is nothing for the adversary to contract, and without $C$ contraction cannot help. Identifiability results for VAE-style models (Khemakhem et al., 2020; Zheng et al., 2022; Kong et al., 2022; Sun et al., 2021; Wang et al., 2023) and the shortcut-robustness of generative classifiers (Li et al., 2025a; Qiang et al., 2025; Müller et al., 2024) both trace factor retention to the reconstruction objective. We ask whether VGAE reconstruction *forces* retention of both factors, and answer affirmatively.

**Theorem 4.1** (Reconstruction forces retention of both factors)**.** *Let $G = g(C, S, \xi)$ as in Equation (1) and let $Z = f_\theta(G)$ be the encoder latent. Assume* (A1) *the generative map $g$ is injective in each argument with the others held fixed, and* (A2) *$G$ is discrete with $H(S \mid C, \xi)$ and $H(C \mid S, \xi)$ finite and positive. Then any encoder–decoder pair attaining reconstruction loss $L_{\text{rec}}$ satisfies* simultaneously

$$I(Z; S \mid C, \xi) \;\geq\; H(S \mid C, \xi) - L_{\text{rec}}, \qquad I(Z; C \mid S, \xi) \;\geq\; H(C \mid S, \xi) - L_{\text{rec}}, \tag{3}$$

*both bounds being strictly positive whenever $L_{\text{rec}} < \min\{H(S \mid C, \xi),\, H(C \mid S, \xi)\}$.*

The key distinction from discriminative training is that the reconstruction objective imposes *per-factor* constraints whose strength scales with each factor's generative footprint on the observed graph, whereas a classification loss yields only a single aggregate constraint bounded by $H(Y)$, which is satisfiable by an encoder that discards $C$ entirely when $S$ alone predicts the training labels. Importantly, the VGAE does not itself learn a "shortcut": it retains $S$ as a generative factor, and the shortcut arises only downstream when a predictor exploits the spurious $S$–$Y$ correlation. Theorem 4.1 establishes information-theoretic retention; it does not guarantee linear accessibility or preservation after graph pooling. The full proof is deferred to Appendix A.1.

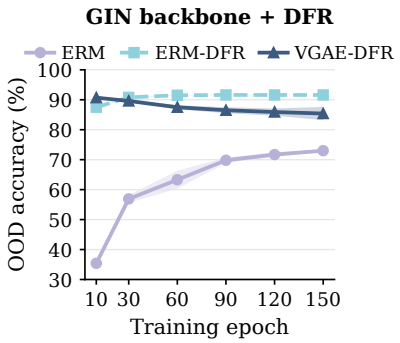


Figure 2: DFR on frozen GIN representations across training, GOOD-Motif concept shift (OOD accuracy, mean±std over 3 seeds; $y$-axis starts at 30%). ERM improves slowly, yet DFR on the same encoder is already above 87% at epoch 10; DFR on the unsupervised VGAE starts at 90.7% and degrades mildly with further training.

Algorithm 1: SCARCE Training Pipeline

**Require:** graphs $\{G_i\}_{i=1}^{N}$, environment indices $\{e_i\}$, weights $\{w_i\}$
1: Pre-train VGAE on $\mathcal{L}_{\text{VGAE}}$; save warm-start checkpoint
2: Initialize VGAE from warm-start checkpoint; initialize adversary $\rho$
3: **for** $t = 1, \ldots, T$ **do**
4: **for** each mini-batch $\mathcal{B}$ **do**
5: Compute $\{\mathbf{h}_G^{(i)}\}_{i\in\mathcal{B}}$ with the VGAE encoder
6: **Adversary step:** freeze encoder; $k_{\text{adv}}$ updates of $\rho$ on $\text{sg}(\mathbf{h}_G)$ with Equation (4)
7: **Encoder step:** freeze $\rho$; update encoder with Equation (6)
8: **end for**
9: **end for**
10: **Evaluation:** freeze encoder; fit a linear probe on training labels

**Empirical verification via feature reweighting.** Theorem 4.1 is information-theoretic; we test whether both factors are also *linearly accessible* with Deep Feature Reweighting (DFR; Kirichenko et al. 2023), which freezes the encoder and refits a sparse linear head on a group-balanced subset, so its OOD accuracy measures the invariant signal the latent exposes once the spurious $S$–$Y$ correlation is removed. We apply DFR to GIN checkpoints of a label-trained ERM encoder and of an unsupervised VGAE on GOOD-Motif concept shift (3 seeds; protocol in Appendix B). Figure 2 shows two things. First, ERM climbs slowly from 35.4% to 73.0%, yet DFR on the *same* frozen encoder is already at 87.5% at epoch 10: $C$ is linearly accessible early, and the ERM head merely underweights it relative to $S$. Second, DFR on the VGAE, which never sees a label, reaches 90.7% at epoch 10, on par with the label-trained encoder: reconstruction exposes $C$ in linear form from the outset, and, since the same latent is what a naive probe exploits via $S$, it exposes $S$ as well. With longer VGAE training, DFR accuracy drifts down to 85.4% with growing variance, suggesting that prolonged reconstruction slowly reallocates latent capacity away from the directions a linear head needs. Since DFR probes the pooled $\mathbf{h}_G$ rather than the latent the decoder sees, retention evidently survives pooling as well.

### 4.2 ADVERSARIAL ENVIRONMENT CONFUSION

Since the VGAE latent retains environment-specific features alongside invariant ones (Section 4.1), SCARCE adds an encoder–adversary alternating optimization that discourages reliance on the environment-discriminative directions.

**Adversary.** The adversarial classifier $\rho : \mathbb{R}^d \to \mathbb{R}^K$ is a two-layer MLP that maps graph representations to environment logits. In each mini-batch, the encoder is frozen and the adversary is updated for $k_{\text{adv}}$ steps on the *detached* representations $\mathbf{h}_{\text{adv}} = \text{sg}(\mathbf{h}_G)$, minimizing the rebalanced cross-entropy:

$$\mathcal{L}_{\text{adv}} = \frac{\sum_i w_i \cdot \text{CE}(\rho(\mathbf{h}_{\text{adv}}^{(i)}),\ e_i)}{\sum_i w_i}, \tag{4}$$

where $e_i$ are the environment indices of Section 3 and $w_i \propto n_{e_i}^{-\alpha}$ are rebalancing weights that compensate for uneven environment sizes, with $n_e$ the number of training graphs in environment $e$ and $\alpha \geq 0$ controlling the strength of rebalancing ($\alpha = 0$ gives uniform weights). Rebalancing matters most when the partition is inferred, since clusters are rarely of equal size.

**Encoder update via uniform confusion.** The adversary is then frozen, and the encoder is updated. The confusion loss drives the adversary's output toward the uniform distribution $\mathbf{u} =$

$(1/K, \ldots, 1/K)$:

$$\mathcal{L}_{\text{conf}} = \frac{\sum_i w_i \cdot \left( -\sum_{k=1}^{K} \frac{1}{K} \log \frac{\exp(\rho_k(\mathbf{h}_G^{(i)}))}{\sum_j \exp(\rho_j(\mathbf{h}_G^{(i)}))} \right)}{\sum_i w_i}. \tag{5}$$

For each sample, this equals $H(U_K, p_\rho) = \log K + \mathrm{KL}(U_K \| p_\rho)$, minimized when the adversary outputs $1/K$ for every environment. The alternating game works as follows: the adversary identifies representation directions that separate environments; the encoder receives gradients that weaken those distinctions; and the reconstruction objective constrains the encoder from discarding all variation.

**Encoder total loss.** The encoder loss combines reconstruction and confusion:

$$\mathcal{L}_{\text{enc}} = \mathcal{L}_{\text{VGAE}} + \beta \cdot \mathcal{L}_{\text{conf}}, \tag{6}$$

where $\beta$ controls the strength of the confusion penalty: $\mathcal{L}_{\text{VGAE}}$ keeps the latent informative (Section 4.1) while $\mathcal{L}_{\text{conf}}$ contracts the directions the adversary identifies.

### 4.3 Theoretical Analysis: Why Confusion Helps OOD Generalization

Erasing environment information discards useful signal together with the shortcut (Belrose et al., 2023; Lin et al., 2024). We show instead that confusion training *contracts* environment-discriminative directions (Lemma 4.2), and that under a regularized downstream classifier this geometric change is exactly a directional penalty that makes spurious features expensive to use (Proposition 4.3). Proofs are in Appendix A.

**Lemma 4.2** (Local contraction)**.** *Fix a linear $K$-class softmax adversary with logits $z(h) = Wh+b$, probabilities $p(h) = \mathrm{softmax}(z(h))$, and uniform target $q = K^{-1}\mathbf{1}$. Define the centered adversary weights $W_c = CW$ where $C = I - K^{-1}\mathbf{1}\mathbf{1}^\top$. Then a gradient step $h^+ = h - \alpha\beta\,\nabla_h L_{\text{conf}}(h)$ contracts each adversary-discriminative direction $v_j$ (right singular vector of $W_c$ with singular value $s_j$) by a factor $1 - \alpha\beta s_j^2/K$ to first order, while directions orthogonal to $\mathrm{range}(W_c^\top)$ are unchanged. In particular, when one singular value dominates ($s_1 \gg s_2$), the cumulative effect of confusion training is well approximated by a rank-one contraction along $u_E = v_1$:*

$$\mathbf{h}_G^{(\beta)} \approx T_\varepsilon\, \mathbf{h}_G^{(0)}, \qquad T_\varepsilon = I - (1-\varepsilon)\, u_E u_E^\top, \quad 0 < \varepsilon \le 1, \tag{7}$$

*where $\varepsilon = 1$ corresponds to no contraction and smaller $\varepsilon$ means the environment direction $u_E$ receives less relative scale.*

**Uniform confusion is thus a *targeted* contraction:** only the directions the adversary uses to separate environments are compressed, more strongly the larger their singular value $s_j$, while directions the adversary ignores, which may carry stable task-relevant information, are left intact. This selectivity is what distinguishes confusion from naive information removal: it narrows the environment-discriminative subspace without collateral damage to invariant features. The rank-one model equation 7 is a first-order consequence of a dominant singular value, and it ties $\varepsilon$ to the training hyperparameters: after $T$ encoder steps, $\varepsilon \approx \prod_{t=1}^{T}(1 - \alpha\beta s_1^2/K)$, so a larger confusion weight $\beta$ or longer training yields a stronger contraction. A neural encoder may additionally rotate or nonlinearly deform the representation, so we verify the model empirically rather than assume it globally.

**Proposition 4.3** (Ridge identity)**.** *For any loss $\ell$, any $\lambda > 0$, and any $\varepsilon \in (0, 1]$, the $\ell_2$-regularized linear probe on the contracted representation is equivalent to an anisotropic ridge penalty on the original representation:*

$$\min_{w,b}\ \mathbb{E}\big[\ell(Y, w^\top T_\varepsilon h + b)\big] + \lambda\|w\|_2^2 \;=\; \min_{v,b}\ \mathbb{E}\big[\ell(Y, v^\top h + b)\big] + \lambda\big[\|v\|_2^2 + (\varepsilon^{-2}-1)(v^\top u_E)^2\big], \tag{8}$$

*where optima correspond under $v = T_\varepsilon w$. The effective penalty on $\mathrm{span}(u_E)$ is $\lambda/\varepsilon^2$, while orthogonal directions retain penalty $\lambda$.*

**Contracting $u_E$ in the representation is *exactly* equivalent to increasing the downstream ridge penalty in that direction from $\lambda$ to $\lambda/\varepsilon^2$, while leaving other directions unaffected.**

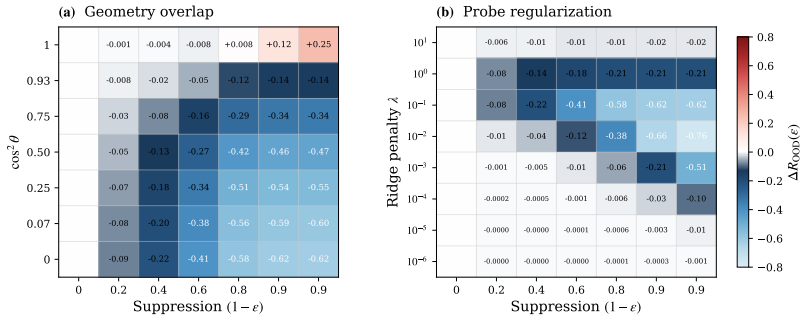


Figure 3: **Controlled contraction of an environment-associated direction.** Cells report $\Delta R_{\mathrm{OOD}}$ (10 seeds; negative = improvement). **(a)** Contraction helps when environment and task directions are separated but hurts at full alignment. **(b)** The effect vanishes as $\lambda \to 0$ and peaks at intermediate $\lambda$, consistent with the induced penalty $\lambda/\varepsilon^2$.

The added term $(\varepsilon^{-2} - 1)(v^\top u_E)^2$ charges only the component of $v$ along $u_E$ and leaves any $v \perp u_E$ untouched. This is **feature reweighting, not feature erasure**: the probe can still use the environment-associated direction, but doing so becomes progressively more expensive as $\varepsilon$ decreases, and the information along $u_E$ is never removed. Proposition 4.3 thus performs the reweighting of DFR (Kirichenko et al., 2023) in representation space, as an anisotropic ridge penalty, and therefore needs neither class labels nor group annotations. **Regularization is essential:** $T_\varepsilon$ is invertible for $\varepsilon > 0$, so an unregularized classifier simply rescales its weight along $u_E$ by $1/\varepsilon$ and contraction has no downstream effect.

**When does contraction help?** A two-dimensional Gaussian model (Proposition A.1, Appendix A.4) makes the answer exact: when the environment feature is spuriously correlated with $Y$ in training but not at test time, the OOD risk is *strictly decreasing* in the contraction amount; as $\lambda \to 0^+$ the probe undoes the contraction and the effect vanishes; and when $u_E = u_Y$ contraction hurts because the compressed direction carries the stable signal. **The mechanism therefore requires exactly two things:** ① a spurious training correlation along a direction separated from the core signal, and ② a finitely regularized probe.

## 5 Experiments

### 5.1 Synthetic Evaluation

We isolate the mechanism of Section 4.3 by directly contracting an environment-associated direction in a synthetic representation and checking downstream behavior against the theory.

**Setup.** We sample $z_c, z_e \sim \mathcal{N}(0,1)$, $\eta \sim \mathcal{N}(0, \sigma_\eta^2)$, and set $Y = z_c$. The representation is $\mathbf{h} = (z_c + \eta)\mathbf{u}_Y + \sigma_s(r z_c + \sqrt{1-r^2}\, z_e)\mathbf{u}_E$, where $\mathbf{u}_E$ and $\mathbf{u}_Y$ have squared overlap $\cos^2\theta$. Training uses $r_{\mathrm{tr}} = 0.9$; testing uses $r_{\mathrm{te}} = 0$. We apply $T_\varepsilon = I - (1-\varepsilon)\mathbf{u}_E\mathbf{u}_E^\top$ directly and fit a ridge probe on 10k training examples, evaluating OOD MSE on 50k test examples. We report $\Delta R_{\mathrm{OOD}} = R_{\mathrm{OOD}}(\varepsilon) - R_{\mathrm{OOD}}(1)$.

**Finding 1: Geometry governs the cost of contraction.** We fix $\lambda = 0.1$ and sweep $\cos^2\theta$. When the two directions are orthogonal, contraction reduces OOD error; the benefit shrinks with increasing overlap; and at full alignment contraction *increases* error because it suppresses stable signal alongside the shortcut (Figure 3a), matching the trade-off in Proposition A.1(iii).

**Finding 2: Finite regularization is necessary.** At near-zero regularization ($\lambda = 10^{-6}$), contraction has negligible effect: the probe rescales its coefficient to undo the contraction, as Proposition A.1(ii) predicts. The benefit peaks at intermediate $\lambda$ and diminishes under excessive regularization, which suppresses all directions uniformly (Figure 3b). **Contraction acts through the regularization geometry, not by erasing information.**

Table 1: **GOOD-Motif concept shift**, OOD test accuracy (%). **Bold**: best; underline: second best.

| Method | Acc. |
|---|---|
| ERM (sup.) | 53.0 ±8.5 |
| GraphCL | 60.9 ±3.4 |
| MVGRL | 54.5 ±4.7 |
| OOD-GCL | 64.1 ±3.6 |
| RGCL | **70.1** ±6.5 |
| VGAE ($\beta$=0) | 65.9 ±2.0 |
| SCARCE | 69.6 ±2.2 |

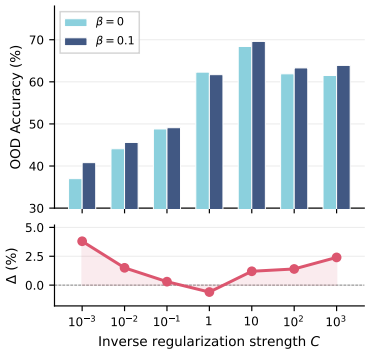


Figure 4: **Probe regularization sweep on GOOD-Motif.** *Top*: OOD accuracy for $\beta$=0 vs. $\beta$=0.1. *Bottom*: gain $\Delta$ from confusion.

### 5.2 CONTROLLED GRAPH BENCHMARK

The synthetic study applies $T_\varepsilon$ to a fixed representation; we now test whether the mechanism survives end-to-end training on realistic graphs.

**Setup.** We use GOOD-Motif concept shift (Gui et al., 2022), where the base-graph distribution shifts across domains while the motif–label relationship is stable. The benchmark provides a domain index for every training graph, which we use directly as the environment $e$ ($K$ equals the number of training domains); class labels are never seen by the encoder. The encoder is a 3-layer GCN with a 64-dimensional latent space, trained with the VGAE reconstruction loss and fine-tuned with adversarial confusion. We freeze the encoder and fit an $\ell_2$-regularized linear classifier on the training labels, sweeping the inverse regularization strength over $C \in \{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$. To isolate the effect of representation geometry from trivial scale differences, we fit a per-dimension standardizer on the $\beta$=0 training representations and apply it, frozen, to all encoder variants before probing. We report OOD test accuracy averaged over 3 seeds.

**Results.** Table 1 reports OOD test accuracy. **SCARCE outperforms GraphCL, MVGRL, and OOD-GCL, and improves over the plain VGAE warm-start.** RGCL performs on par with SCARCE on this setting.

**Ablation: probe regularization.** Figure 4 sweeps $C$ over six orders of magnitude. Accuracy for both $\beta$=0 and $\beta$=0.1 is low when the probe is too constrained to use any feature, peaks near $C$=10, and drops once weak regularization lets the probe exploit shortcut features. The gain $\Delta$ from confusion is non-monotonic: it is largest (+3.8%) at $C$=$10^{-3}$, where a large base penalty $\lambda$ is inflated to $\lambda/\varepsilon^2$ on the environment direction (Proposition 4.3); it vanishes at $C$=1, where the probe has enough capacity to partially undo the contraction (Proposition A.1(ii)); and it returns at $C \geq 10^2$ (+1.4, +2.4%), where the baseline probe freely exploits shortcuts that the confusion-trained encoder has already contracted.

### 5.3 REAL-WORLD MOLECULAR EVALUATION

**Setup.** We use EC50 and IC50 from DrugOOD (Ji et al., 2023) under assay, scaffold, and size splits. Each encoder is frozen and a logistic-regression probe is fitted on standardized representations using 5%, 10%, or 20% of the labeled training split; $(\alpha, \beta, K)$ and checkpoints are selected on validation and the test split is used once. Baselines are GraphCL, RGCL, OOD-GCL, MVGRL, and VGAE ($\beta$=0), i.e. SCARCE without confusion. Details are in Appendix C.

**Environment partition.** Unlike GOOD-Motif, DrugOOD offers no usable environment index inside the training split: the assay, scaffold, and size attributes define the train/validation/test split, but within training they are respectively thousands of sparsely populated assay IDs, near-unique scaffolds, and a continuous graph size, none of which yields a partition of the kind Section 3 assumes. We therefore construct *pseudo-environments* without labels, following Yao et al. (2024): a frozen

Table 2: **DrugOOD under limited labels.** Test ROC-AUC (%), mean ± std over three seeds (IC50-size VGAE at 20%: two), for linear probes on frozen representations. **Bold**: best per column; Avg.: mean over the six splits; shaded row: SCARCE.

| Method | EC50 | | | IC50 | | | |
|---|---|---|---|---|---|---|---|
| | Assay | Scaffold | Size | Assay | Scaffold | Size | Avg. |
| *5% labels* | | | | | | | |
| MVGRL | 68.9±3.7 | 52.1±3.7 | 57.8±2.9 | 67.1±0.3 | 57.4±1.7 | **62.1±0.8** | 60.9 |
| VGAE ($\beta$=0) | 58.2±2.8 | 50.0±6.2 | 52.9±4.1 | 61.5±2.2 | 55.7±2.2 | 56.8±0.7 | 55.9 |
| GraphCL | 65.1±2.0 | 54.6±9.0 | 56.3±1.8 | 67.0±1.0 | 61.8±1.9 | 61.5±1.1 | 61.1 |
| RGCL | 69.5±3.1 | 54.8±7.5 | 56.5±2.9 | 65.8±0.1 | 60.0±1.6 | 60.5±1.1 | 61.2 |
| OOD-GCL | 59.5±6.5 | 57.3±6.2 | 53.4±1.1 | 58.1±2.2 | 57.3±4.0 | 57.1±2.9 | 57.1 |
| SCARCE | **70.6±1.0** | **65.0±1.0** | **61.0±1.8** | **69.5±0.5** | **64.3±1.2** | 61.6±1.0 | **65.3** |
| *10% labels* | | | | | | | |
| MVGRL | 67.5±2.5 | 54.4±3.5 | 56.0±1.8 | 67.8±0.2 | 59.8±1.0 | 61.7±1.1 | 61.2 |
| VGAE ($\beta$=0) | 59.6±2.0 | 49.9±1.9 | 53.2±1.9 | 62.1±1.0 | 58.4±1.6 | 58.6±0.3 | 57.0 |
| GraphCL | 66.9±1.7 | 58.8±2.5 | 58.4±0.7 | 67.3±0.7 | 62.8±0.8 | 61.5±0.4 | 62.6 |
| RGCL | 69.8±1.5 | 58.8±2.3 | 57.7±2.6 | 66.9±1.3 | 60.8±1.7 | 61.3±0.6 | 62.6 |
| OOD-GCL | 63.9±4.7 | 57.2±3.1 | 53.2±1.0 | 59.0±1.6 | 58.8±4.0 | 58.3±2.9 | 58.4 |
| SCARCE | **71.8±1.7** | **64.9±1.2** | **62.9±1.5** | **69.5±0.1** | **66.0±0.7** | **62.6±0.9** | **66.3** |
| *20% labels* | | | | | | | |
| MVGRL | 67.8±1.6 | 56.3±1.2 | 57.1±2.4 | 67.8±0.5 | 61.4±0.8 | 61.6±0.3 | 62.0 |
| VGAE ($\beta$=0) | 62.5±4.6 | 53.1±0.6 | 55.6±1.4 | 63.9±0.4 | 60.4±0.7 | 60.2±0.0 | 59.3 |
| GraphCL | 70.7±3.7 | 59.1±3.7 | 60.2±0.4 | 67.5±0.3 | 63.5±0.7 | 62.3±0.7 | 63.9 |
| RGCL | 73.0±6.2 | 55.0±4.3 | 61.0±0.6 | 67.2±0.4 | 61.3±1.1 | 62.3±1.0 | 63.3 |
| OOD-GCL | 64.3±4.0 | 58.7±4.3 | 53.9±1.0 | 60.1±1.9 | 58.7±3.6 | 59.0±3.1 | 59.1 |
| SCARCE | **73.6±1.3** | **64.4±1.7** | **62.8±0.5** | **69.7±0.1** | **65.9±0.4** | **63.2±0.5** | **66.6** |

MVGRL encoder embeds the training graphs, $k$-means with $K$ clusters is run on the embeddings, and the cluster index serves as $e_i$. Cluster imbalance is absorbed by the rebalancing exponent $\alpha$ of Section 4.2. No annotation of any kind is used during representation learning, and $K$ is treated as a hyperparameter selected on validation.

**Results.** Table 2 reports test ROC-AUC. **SCARCE is best in 17 of the 18 dataset–ratio cells,** the sole exception being IC50-size at 5% labels, and the margin is largest on the scaffold splits, where test molecules differ structurally from training ones. The VGAE ($\beta$=0) row is the ablation without confusion: it is the weakest representation on most splits, trailing SCARCE by 11–15 points on EC50-scaffold and by 3–10 points elsewhere, **so the gain is attributable to confusion training rather than to the reconstruction backbone.** MVGRL, whose embedding only supplies the pseudo-environment partition, is likewise well below SCARCE, so a rough, label-free partition is enough for confusion to act on. Averaged over the six splits, SCARCE leads the strongest baseline by 4.1, 3.7, and 2.7 points at 5%, 10%, and 20% labels, so the gain holds at every label budget.

**Sensitivity.** Across the top-ranked configurations (Appendix C.1), the strongest confusion weight $\beta = 10$ is preferred in nearly every setting, whereas the best $K$ varies across splits and is selected on validation. The preference for large $\beta$ matches Proposition 4.3: a probe fitted on 5–20% of the labels sits in the strongly regularized regime where Figure 3(b) shows contraction to help most. Component-loss curves (Appendix C.2) confirm that the alternating optimization is stable.

## 6 CONCLUSION

We introduced SCARCE, a label-free framework for graph OOD generalization that pairs a VGAE with adversarial confusion against the training environments, given or, as on DrugOOD, inferred by clustering. Both components come with a guarantee: reconstruction forces the latent to retain core and environment factors alike, and confusion contracts, rather than erases, the environment-discriminative directions, which for any $\ell_2$-regularized probe is exactly an anisotropic ridge penalty. Experiments on a controlled synthetic model, GOOD-Motif, and DrugOOD confirm both the mechanism and its preconditions. The recipe, i.e., **retain by reconstruction, then reweight by contraction**, is not specific to graphs and extends naturally to other likelihood-based encoders.

## ETHICS STATEMENT

This work studies unsupervised representation learning for graph out-of-distribution generalization. All experiments use publicly available benchmarks that contain no personal data or human subjects. By removing the need for class labels during representation learning and tolerating coarse or inferred environment partitions, SCARCE lowers the cost of building predictors that remain reliable under distribution shift, which is beneficial in settings such as molecular property prediction where labeled and annotated data are scarce. The method does not create new data-collection or surveillance capabilities, and we do not foresee any direct negative societal impact beyond those generic to representation learning.

## REFERENCES

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *NeurIPS*, 2023.

Yongqiang Chen, Yonggang Zhang, Yatao Bian, Han Yang, Kaili Ma, Binghui Xie, Tongliang Liu, Bo Han, and James Cheng. Learning causally invariant representations for out-of-distribution generalization on graphs. In *NeurIPS*, 2022.

Yongqiang Chen, Yatao Bian, Kaiwen Zhou, Binghui Xie, Bo Han, and James Cheng. Does invariant graph learning via environment augmentation learn invariance? In *NeurIPS*, 2023. URL https://openreview.net/forum?id=EqpR9Vtt13.

Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *ICML*, 2021.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Shurui Gui, Xiner Li, Limei Wang, and Shuiwang Ji. GOOD: A graph out-of-distribution benchmark. In *NeurIPS Datasets and Benchmarks Track*, 2022.

Shurui Gui, Meng Liu, Xiner Li, Youzhi Luo, and Shuiwang Ji. Joint learning of label and environment causal independence for graph out-of-distribution generalization. In *NeurIPS*, 2023.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *ICLR*, 2021.

Kaveh Hassani and Amir Hosein Khasahmadi. Contrastive multi-view representation learning on graphs. In *ICML*, 2020.

Zhenyu Hou, Xiao Liu, Yukuo Cen, Yuxiao Dong, Hongxia Yang, Chunjie Wang, and Jie Tang. GraphMAE: Self-supervised masked graph autoencoders. In *KDD*, 2022.

Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. In *ICLR*, 2020.

Yuanfeng Ji, Lu Zhang, Jiaxiang Wu, Bingzhe Wu, Long-Kai Huang, Tingyang Xu, Yu Rong, Lanqing Li, Jie Ren, Ding Xue, et al. DrugOOD: Out-of-distribution dataset curator and benchmark for AI-aided drug discovery. In *AAAI*, 2023.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *AISTATS*, pp. 2207–2217. PMLR, 2020.

Thomas N Kipf and Max Welling. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*, 2016.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *ICML*, 2023.

Lingjing Kong, Shaoan Xie, Weiran Yao, Yujia Zheng, Guangyi Chen, Petar Stojanov, Victor Akinwande, and Kun Zhang. Partial disentanglement for domain adaptation. In *ICML*, pp. 11455–11472, 2022.

Abhinav Kumar, Chenhao Tan, and Amit Sharma. Probing classifiers are unreliable for concept removal and detection. In *NeurIPS*, 2022.

Alexander C. Li, Ananya Kumar, and Deepak Pathak. Generative classifiers avoid shortcut solutions. In *ICLR*, 2025a.

Haoyang Li, Ziwei Zhang, Xin Wang, and Wenwu Zhu. Learning invariant graph representations for out-of-distribution generalization. In *NeurIPS*, 2022a.

Haoyang Li, Xin Wang, Zeyang Zhang, Haibo Chen, Ziwei Zhang, and Wenwu Zhu. Disentangled graph self-supervised learning for out-of-distribution generalization. In *ICML*, pp. 28890–28904, 2024.

Haoyang Li, Xin Wang, Ziwei Zhang, and Wenwu Zhu. Out-of-distribution generalization on graphs: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(11): 10490–10512, 2025b.

Sihang Li, Xiang Wang, An Zhang, Yingxin Wu, Xiangnan He, and Tat-Seng Chua. Let invariant rationale discovery inspire graph contrastive learning. In *ICML*, pp. 13052–13065. PMLR, 2022b.

Yong Lin, Shengyu Zhu, Lu Tan, and Peng Cui. Zin: When and how to learn invariance without environment partition? In *NeurIPS*, 2022.

Yong Lin, Lu Tan, Yifan Hao, Honam Wong, Hanze Dong, Weizhong Zhang, Yujiu Yang, and Tong Zhang. Spurious feature diversification improves out-of-distribution generalization. In *ICLR*, 2024.

Siqi Miao, Mia Liu, and Pan Li. Interpretable and generalizable graph learning via stochastic attention mechanism. In *ICML*, pp. 15524–15543. PMLR, 2022.

Nicolas M Müller, Simon Roschmann, Shahbaz Khan, Philip Sperl, and Konstantin Böttinger. Shortcut detection with variational autoencoders. In *IJCNN*, pp. 1–7. IEEE, 2024.

Wenwen Qiang, Jingyao Wang, Zeen Song, Jiangmeng Li, and Changwen Zheng. On the out-of-distribution generalization of self-supervised learning. In *ICML*, pp. 50086–50112, 2025.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan Cotterell. Linear adversarial concept erasure. In *ICML*, 2022.

Joshua Robinson, Li Sun, Ke Yu, Kayhan Batmanghelich, Stefanie Jegelka, and Suvrit Sra. Can contrastive learning avoid shortcut solutions? *Advances in neural information processing systems*, 34:4974–4986, 2021.

Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. Self-supervised graph transformer on large-scale molecular data. *Advances in neural information processing systems*, 33:12559–12571, 2020.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *ICLR*, 2020.

Fan-Yun Sun, Jordan Hoffmann, Vikas Verma, and Jian Tang. Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization. In *ICLR*, 2020.

Xinwei Sun, Botong Wu, Xiangyu Zheng, Chang Liu, Wei Chen, Tao Qin, and Tie-Yan Liu. Recovering latent causal factor for generalization to distributional shifts. *Advances in Neural Information Processing Systems*, 34:16846–16859, 2021.

Xinyi Wang, Michael Saxon, Jiachen Li, Hongyang Zhang, Kun Zhang, and William Yang Wang. Causal balancing for domain generalization. In *ICLR*, 2023.

Qitian Wu, Hengrui Zhang, Junchi Yan, and David Wipf. Handling distribution shifts on graphs: An invariance perspective. In *ICLR*, 2022a.

Ying-Xin Wu, Xiang Wang, An Zhang, Xiangnan He, and Tat-Seng Chua. Discovering invariant rationales for graph neural networks. In *ICLR*, 2022b.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *ICLR*, 2019.

Nianzu Yang, Kaipeng Zeng, Qitian Wu, Xiaosong Jia, and Junchi Yan. Learning substructure invariance for out-of-distribution molecular representations. *Advances in Neural Information Processing Systems*, 35:12964–12978, 2022a.

Wanqian Yang, Polina Kirichenko, Micah Goldblum, and Andrew G Wilson. Chroma-vae: Mitigating shortcut learning with generative classifiers. *Advances in Neural Information Processing Systems*, 35:20351–20365, 2022b.

Tianjun Yao, Yongqiang Chen, Zhenhao Chen, Kai Hu, Zhiqiang Shen, and Kun Zhang. Empowering graph invariance learning with deep spurious infomax. In *ICML*, pp. 56860–56884, 2024.

Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. In *NeurIPS*, 2020.

Xingxuan Zhang, Linjun Zhou, Renzhe Xu, Peng Cui, Zheyan Shen, and Haoxin Liu. Towards unsupervised domain generalization. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4900–4910. IEEE, 2022.

Yujia Zheng, Ignavier Ng, and Kun Zhang. On the identifiability of nonlinear ICA: Sparsity and beyond. *Advances in Neural Information Processing Systems*, 35:16411–16422, 2022.

# APPENDIX

## A PROOFS

### A.1 PROOF OF THEOREM 4.1 (RECONSTRUCTION FORCES RETENTION)

Throughout, $G = g(C, S, \xi)$ is a discrete random variable, $Z = f_\theta(G)$ is the (possibly stochastic) encoder latent, so that $(C, S, \xi) \to G \to Z$ forms a Markov chain, and $p_\psi(\cdot \mid Z)$ is the decoder distribution over graphs. All entropies are in nats.

**Step 1: Reconstruction loss upper-bounds the conditional entropy of $G$.** For every value $z$ of $Z$, the Gibbs inequality gives $\mathbb{E}[-\log p_\psi(G \mid z) \mid Z = z] \geq H(G \mid Z = z)$, with equality iff $p_\psi(\cdot \mid z)$ equals the true posterior $p(\cdot \mid z)$. Taking expectation over $Z$,

$$L_{\text{rec}} = \mathbb{E}[-\log p_\psi(G \mid Z)] \;\geq\; H(G \mid Z). \tag{9}$$

This is the only place the decoder enters: any encoder–decoder pair achieving reconstruction loss $L_{\text{rec}}$ leaves at most $L_{\text{rec}}$ nats of $G$ unexplained by $Z$.

**Step 2: Injectivity turns $S$ into a function of $(G, C, \xi)$.** By (A1), for each fixed $(c, \xi')$ the map $s \mapsto g(c, s, \xi')$ is injective, so there exists a measurable $\phi_S$ with $S = \phi_S(G, C, \xi)$ almost surely. Symmetrically, injectivity in the first argument yields $\phi_C$ with $C = \phi_C(G, S, \xi)$.

**Step 3: Chain of conditioning inequalities.** Since $S$ is a deterministic function of $(G, C, \xi)$, conditioning on $(Z, C, \xi)$ and using that the entropy of a function is at most the entropy of its argument,

$$H(S \mid Z, C, \xi) \;=\; H(\phi_S(G, C, \xi) \mid Z, C, \xi) \;\leq\; H(G \mid Z, C, \xi) \;\leq\; H(G \mid Z), \tag{10}$$

where the last step is "conditioning reduces entropy." Combining with equation 9, $H(S \mid Z, C, \xi) \leq L_{\text{rec}}$.

**Step 4: Conclusion.** By the chain rule for conditional mutual information,

$$I(Z; S \mid C, \xi) \;=\; H(S \mid C, \xi) - H(S \mid Z, C, \xi) \;\geq\; H(S \mid C, \xi) - L_{\text{rec}}, \tag{11}$$

which is the first bound in equation 3. The second bound follows by exchanging the roles of $C$ and $S$ in Steps 2–3 using $\phi_C$. Both bounds are strictly positive whenever $L_{\text{rec}} < \min\{H(S \mid C, \xi),\, H(C \mid S, \xi)\}$, and (A2) guarantees the right-hand sides are finite so the statement is non-vacuous. □

**Why the bounds are per-factor.** The two inequalities hold *simultaneously* for the same $L_{\text{rec}}$ because each factor is recoverable from $G$ once the others are fixed; the reconstruction loss therefore has to pay for every factor separately. A discriminative objective offers no such guarantee: minimizing $\mathbb{E}[-\log q(Y \mid Z)]$ only constrains $I(Z; Y) \leq H(Y)$, and an encoder that discards $C$ can still attain this whenever $S$ predicts $Y$ on the training environments.

### A.2 PROOF OF LEMMA 4.2 (LOCAL CONTRACTION)

We first establish the local contraction result in four steps (Steps 1–4): deriving the confusion gradient, introducing centering, linearizing near uniform predictions, and showing per-direction contraction. We then derive the rank-one cumulative approximation in Steps 5–8.

**Step 1: Confusion gradient in representation space.** The per-example confusion loss is cross-entropy with the uniform target $q = K^{-1}\mathbf{1}$:

$$L_{\text{conf}}(h) = -q^\top \log p(h), \qquad p(h) = \text{softmax}(Wh + b). \tag{12}$$

The standard logit-space gradient of cross-entropy with target $q$ is $\nabla_z L_{\text{conf}} = p - q$. Using $z = Wh + b$, the chain rule gives:

$$\begin{aligned} \nabla_h L_{\text{conf}} &= \frac{\partial z}{\partial h}^\top \nabla_z L_{\text{conf}} \\ &= W^\top (p - q). \end{aligned} \tag{13}$$

That is, the representation gradient is determined entirely by the directions the adversary uses to distinguish environments.

**Step 2: Centering removes the common logit shift.** Since both $p$ and $q$ are probability vectors summing to one, their difference satisfies $\mathbf{1}^\top(p-q)=0$, so $p-q\in\mathbf{1}^\perp$ and $C(p-q)=p-q$ where $C=I-K^{-1}\mathbf{1}\mathbf{1}^\top$. Therefore:

$$\begin{aligned}\nabla_h L_{\text{conf}} &= W^\top(p-q)\\ &= W^\top C(p-q)\\ &= (CW)^\top(p-q)\\ &= W_c^\top(p-q),\end{aligned} \tag{14}$$

which shows the gradient lies in $\mathrm{range}(W_c^\top)$. Since $C$ has rank $K-1$, this subspace has dimension at most $K-1$. Centering removes the common logit shift and keeps only the environment-discriminative directions.

**Step 3: Linearization near equal-logit points.** At $h_0$ where $W_c h_0 + b_c = 0$, all centered logits vanish, hence all logits are equal and $p(h_0) = q$. The Jacobian of softmax evaluated at equal probabilities is:

$$\begin{aligned}J_{\text{sm}}(z(h_0)) &= \mathrm{diag}(q) - qq^\top\\ &= \frac{1}{K}\left(I - \frac{1}{K}\mathbf{1}\mathbf{1}^\top\right)\\ &= \frac{1}{K}C.\end{aligned} \tag{15}$$

A first-order Taylor expansion of $p$ around $h_0$, using $\delta h = h - h_0$ and noting that $CW\delta h = W_c\delta h$, gives:

$$p(h) - q = \frac{1}{K}W_c\delta h + O(\|W_c\delta h\|_2^2). \tag{16}$$

Substituting into the gradient expression from Step 2:

$$\begin{aligned}\nabla_h L_{\text{conf}} &= W_c^\top(p-q)\\ &= W_c^\top\left(\frac{1}{K}W_c\delta h + O(\|W_c\delta h\|_2^2)\right)\\ &= \frac{1}{K}W_c^\top W_c\,\delta h + O(\|W_c\delta h\|_2^2).\end{aligned} \tag{17}$$

**Step 4: Contraction per singular direction.** A gradient step $h^+ = h - \alpha\beta\,\nabla_h L_{\text{conf}}$ gives, in terms of $\delta h$:

$$\delta h^+ = \left(I - \frac{\alpha\beta}{K}W_c^\top W_c\right)\delta h + O(\alpha\beta\|W_c\delta h\|_2^2). \tag{18}$$

Let $W_c = U\,\mathrm{diag}(s_1,\dots,s_r)V^\top$ be a thin SVD. Then:

$$W_c^\top W_c = V\,\mathrm{diag}(s_1^2,\dots,s_r^2)V^\top, \tag{19}$$

and projecting onto each right singular direction $v_j$:

$$v_j^\top\delta h^+ = \left(1 - \frac{\alpha\beta}{K}s_j^2\right)v_j^\top\delta h + O(\alpha\beta\|W_c\delta h\|_2^2). \tag{20}$$

When $0 < \alpha\beta s_j^2/K < 1$, the leading-order factor is a contraction. Directions with larger adversarial singular values are contracted more strongly, meaning the representation selectively suppresses environment-discriminative components. For any direction $v\perp\mathrm{range}(V)=\mathrm{range}(W_c^\top)$, we have $W_c^\top W_c\,v = 0$, so $v^\top\delta h^+ = v^\top\delta h$ exactly; non-discriminative directions are unchanged.

**Step 5: Rank-one approximation.** When the adversary's centered weight matrix has a dominant singular value $s_1 \gg s_2$, the leading-order update from Step 4 is well approximated by retaining only the $j=1$ term. Writing $u_E := v_1$ for the leading right singular direction and defining $\gamma := \alpha\beta s_1^2/K$, the single-step update becomes:

$$h^{(1)} \approx (I - \gamma\,u_E u_E^\top)\,h^{(0)}. \tag{21}$$

**Step 6: Decomposition into environment and orthogonal components.** To see how the update acts on different components, decompose any representation as $h = h_E + h_\perp$, where $h_E = (u_E^\top h)\, u_E$ is the environment-direction component and $u_E^\top h_\perp = 0$. Denoting the projection $P = u_E u_E^\top$ (so $P^2 = P$), one confusion step gives:

$$\begin{aligned} h^{(1)} &= (I - \gamma P)(h_E + h_\perp) \\ &= h_E + h_\perp - \gamma P\, h_E \\ &= (1-\gamma)\, h_E + h_\perp. \end{aligned} \tag{22}$$

That is, the environment component shrinks by a factor $(1-\gamma)$, while the orthogonal component is unchanged.

**Step 7: Cumulative effect of $m$ confusion steps.** Repeating the same rank-one update $m$ times and using $P^2 = P$:

$$\begin{aligned} h^{(m)} &= (I - \gamma P)^m\, h^{(0)} \\ &= (1-\gamma)^m\, h_E^{(0)} + h_\perp^{(0)}. \end{aligned} \tag{23}$$

Each step multiplicatively shrinks the environment component while leaving the orthogonal complement intact. Defining $\varepsilon := (1-\gamma)^m \in (0, 1]$, we can write:

$$\begin{aligned} h^{(m)} &= \varepsilon\, h_E^{(0)} + h_\perp^{(0)} \\ &= h^{(0)} - (1-\varepsilon)(u_E^\top h^{(0)})\, u_E. \end{aligned} \tag{24}$$

**Step 8: The contraction model $T_\varepsilon$.** The cumulative multi-step effect can therefore be expressed as a single linear map:

$$\mathbf{h}_G^{(\beta)} \approx T_\varepsilon\, \mathbf{h}_G^{(0)}, \qquad T_\varepsilon = I - (1-\varepsilon)\, u_E u_E^\top, \quad 0 < \varepsilon \le 1. \tag{25}$$

This map has a transparent geometric action:

$$\begin{aligned} &u_E^\top T_\varepsilon h = \varepsilon\, (u_E^\top h), \\ &v^\top T_\varepsilon h = v^\top h \quad \text{for } v \perp u_E. \end{aligned} \tag{26}$$

When $\varepsilon = 1$, $T_\varepsilon = I$ (no contraction); as $\varepsilon \to 0$, the environment-discriminative component is increasingly suppressed, while all other directions are preserved. ■

**Scope.** Steps 1–4 are exact for the gradient direction and first-order for the contraction magnitude, holding for a single step with a fixed linear adversary and a directly updated representation. Steps 5–8 model the cumulative multi-step effect as a rank-one contraction under three simplifying assumptions: (a) the adversary weights (hence $u_E$) remain approximately stable across steps; (b) the encoder update acts directly on $h$ rather than through neural network parameters; and (c) the dominant singular value $s_1$ governs the contraction. In the implemented method, the encoder is updated through its parameter Jacobian (introducing a preconditioner), the adversary is nonlinear and alternately retrained, and the reconstruction loss acts simultaneously. The rank-one model $T_\varepsilon$ is therefore a first-order analytical approximation that captures the leading effect of confusion training, suppressing the dominant environment-discriminative direction while preserving orthogonal components, and is verified empirically rather than derived as a theorem.

### A.3 Proof of Proposition 4.3 (Ridge Identity)

The key insight is that contraction $T_\varepsilon$ does not remove information; rather, it changes the downstream regularization geometry. We establish this through a change of variables.

**Step 1: Invertibility of $T_\varepsilon$.** For $\varepsilon > 0$, $T_\varepsilon$ is symmetric and invertible with:

$$T_\varepsilon^{-1} = I + (\varepsilon^{-1} - 1)\, u_E u_E^\top. \tag{27}$$

To verify, using $u_E^\top u_E = 1$:

$$\begin{aligned} T_\varepsilon \cdot T_\varepsilon^{-1} &= (I - (1-\varepsilon)\, u_E u_E^\top)(I + (\varepsilon^{-1} - 1)\, u_E u_E^\top) \\ &= I + \big[(\varepsilon^{-1} - 1) - (1-\varepsilon) - (1-\varepsilon)(\varepsilon^{-1} - 1)\big] u_E u_E^\top \\ &= I, \end{aligned} \tag{28}$$

since the bracketed coefficient simplifies to $\varepsilon^{-1} - 1 - 1 + \varepsilon - \varepsilon^{-1} + 1 + 1 - \varepsilon = 0$.

**Step 2: Squared inverse.** By squaring $T_\varepsilon^{-1}$ using the identity $(I + \gamma\, u_E u_E^\top)^2 = I + ((1+\gamma)^2 - 1)\, u_E u_E^\top$ with $\gamma = \varepsilon^{-1} - 1$:

$$T_\varepsilon^{-2} = I + (\varepsilon^{-2} - 1)\, u_E u_E^\top. \tag{29}$$

**Step 3: Change of variables.** Substitute $v := T_\varepsilon w$, which is a bijection by Step 1, leaving $b$ unchanged. The prediction function is unchanged after reparameterization:

$$\begin{aligned} w^\top T_\varepsilon h &= (T_\varepsilon w)^\top h \\ &= v^\top h, \end{aligned} \tag{30}$$

so the loss terms agree pointwise.

**Step 4: Transform the regularizer.** For the penalty term, $w = T_\varepsilon^{-1} v$, and only the environment direction receives an additional penalty:

$$\begin{aligned} \|w\|_2^2 &= \|T_\varepsilon^{-1} v\|_2^2 \\ &= v^\top T_\varepsilon^{-2} v \\ &= v^\top \left[ I + (\varepsilon^{-2} - 1)\, u_E u_E^\top \right] v \\ &= \|v\|_2^2 + (\varepsilon^{-2} - 1)(v^\top u_E)^2. \end{aligned} \tag{31}$$

The two objectives are therefore equal as functions under the bijection $v = T_\varepsilon w$, so their infima agree and optima correspond whenever they exist. The penalty matrix $\lambda T_\varepsilon^{-2}$ has eigenvalue $\lambda/\varepsilon^2$ on $\mathrm{span}(u_E)$ and $\lambda$ on $u_E^\perp$. Therefore, contraction is equivalent to directional feature reweighting: using the environment direction becomes disproportionately expensive under $\ell_2$ regularization. ■

### A.4 Exact Worked Example: Statement and Proof

The mechanism improves OOD generalization when the contracted direction captures unstable environment-associated variation while remaining separated from invariant task-relevant information. We formalize this intuition using a two-dimensional Gaussian model; the statement is the one summarized in Section 4.3. The analysis highlights two essential requirements: (1) the environment-associated feature induces a spurious training correlation, and (2) the downstream classifier applies finite regularization.

**Proposition A.1** (Exact worked example)**.** *Let $h = (z_c + \eta)\, u_Y + \sigma_s (r\, z_c + \sqrt{1 - r^2}\, z_e)\, u_E$ with $Y = z_c$, orthonormal $(u_Y, u_E)$, $z_c, z_e \sim \mathcal{N}(0,1)$ and $\eta \sim \mathcal{N}(0, \sigma_\eta^2)$ mutually independent, $\sigma_\eta^2 > 0$, $\sigma_s > 0$, and training correlation $r = r_{\mathrm{tr}}$. Under population ridge regression with regularization $\lambda > 0$:*

- *(i) If $r_{\mathrm{tr}} \neq 0$ and $r_{\mathrm{te}} = 0$ (correlation shift), the OOD risk $R(\varepsilon; r_{\mathrm{te}})$ is* strictly decreasing *in the contraction amount: every reduction of $\varepsilon$ helps.*
- *(ii) As $\lambda \to 0^+$ (no regularization), both fitted coefficients become independent of $\varepsilon$: contraction has no downstream effect.*
- *(iii) If $u_E = u_Y$ (full alignment), contraction hurts because the compressed direction also contains task-relevant information.*

This proposition makes the two requirements concrete. Part (i) shows the positive case: when a spurious correlation exists in training ($r_{\mathrm{tr}} \neq 0$) but vanishes at test time ($r_{\mathrm{te}} = 0$), contraction *strictly* reduces OOD risk. Two effects contribute simultaneously: the probe's loading on the spurious feature decreases, and its core coefficient, which was deflated at $\varepsilon = 1$ because the environment feature could proxy for the label, is restored toward its test-optimal value. Parts (ii) and (iii) characterize the failure modes: without regularization the probe can undo the contraction for free, and when the environment direction overlaps the stable signal, contraction damages invariant information. Together, these results precisely delineate when SCARCE is expected to help and when it is not.

**Proof.** The proof follows the chain: data model → covariance $\Sigma$ and cross-covariance $b$ → ridge solution $\hat{w}_\varepsilon$ → effective coefficients $a_c, a_s$ → OOD risk.

**Step 1: Setup and covariance derivation.** We work in the orthonormal basis $(u_Y, u_E)$, so each representation is $h = (h_1, h_2)$ with $h_1 = z_c + \eta$ (core feature with noise) and $h_2 = \sigma_s(r\, z_c + \sqrt{1-r^2}\, z_e)$ (environment feature). The label is $Y = z_c$. In this basis, the contraction $T_\varepsilon$ acts as $S_\varepsilon = \mathrm{diag}(1, \varepsilon)$, leaving the core feature unchanged and compressing the environment feature.

Since all variables are zero-mean and $z_c, z_e, \eta$ are mutually independent with $\mathbb{E}[z_c^2] = \mathbb{E}[z_e^2] = 1$ and $\mathbb{E}[\eta^2] = \sigma_\eta^2$, the core variance is:

$$\begin{aligned} \mathbb{E}[h_1^2] &= \mathbb{E}[(z_c + \eta)^2] \\ &= \mathbb{E}[z_c^2] + \mathbb{E}[\eta^2] \\ &= 1 + \sigma_\eta^2. \end{aligned} \tag{32}$$

The cross-covariance between core and environment features is:

$$\begin{aligned} \mathbb{E}[h_1 h_2] &= \sigma_s \mathbb{E}\Big[(z_c + \eta)(r\, z_c + \sqrt{1-r^2}\, z_e)\Big] \\ &= r\sigma_s. \end{aligned} \tag{33}$$

This off-diagonal term directly captures the spurious correlation. Similarly, $\mathbb{E}[h_2^2] = \sigma_s^2$, giving:

$$\Sigma = \mathbb{E}[hh^\top] = \begin{pmatrix} 1 + \sigma_\eta^2 & r\sigma_s \\ r\sigma_s & \sigma_s^2 \end{pmatrix}, \qquad b = \mathbb{E}[hY] = \begin{pmatrix} 1 \\ r\sigma_s \end{pmatrix}. \tag{34}$$

**Step 2: Ridge solution on contracted features.** The contracted features are $x = S_\varepsilon h$, so $\mathbb{E}[xx^\top] = S_\varepsilon \Sigma S_\varepsilon$ and $\mathbb{E}[xY] = S_\varepsilon b$. The population ridge regression solution is:

$$\begin{aligned} \hat{w}_\varepsilon &= (\mathbb{E}[xx^\top] + \lambda I)^{-1} \mathbb{E}[xY] \\ &= (S_\varepsilon \Sigma S_\varepsilon + \lambda I)^{-1} S_\varepsilon b. \end{aligned} \tag{35}$$

The determinant of the $2 \times 2$ system is:

$$\Delta_\varepsilon = (1 + \sigma_\eta^2 + \lambda)(\varepsilon^2 \sigma_s^2 + \lambda) - \varepsilon^2 r^2 \sigma_s^2 > 0 \quad (\lambda > 0). \tag{36}$$

The effective coefficients on the original (unshrunk) features, defined by writing the prediction as $\hat{Y} = a_c h_1 + a_s h_2$ with $a_c = [\hat{w}_\varepsilon]_1$ and $a_s = \varepsilon[\hat{w}_\varepsilon]_2$, are:

$$a_c(\varepsilon) = \frac{\lambda + \varepsilon^2 \sigma_s^2 (1 - r^2)}{\Delta_\varepsilon}, \qquad a_s(\varepsilon) = \frac{\varepsilon^2 r \sigma_s (\sigma_\eta^2 + \lambda)}{\Delta_\varepsilon}. \tag{37}$$

**Step 3: OOD risk derivation.** At test time, $h_2 = \sigma_s(r_{\mathrm{te}}\, z_c + \sqrt{1 - r_{\mathrm{te}}^2}\, z_e)$. The prediction error is:

$$\begin{aligned} Y - \hat{Y} &= z_c - a_c(z_c + \eta) - a_s \sigma_s (r_{\mathrm{te}}\, z_c + \sqrt{1 - r_{\mathrm{te}}^2}\, z_e) \\ &= (1 - a_c - a_s \sigma_s r_{\mathrm{te}})\, z_c - a_c\, \eta - a_s \sigma_s \sqrt{1 - r_{\mathrm{te}}^2}\, z_e. \end{aligned} \tag{38}$$

Since $z_c, \eta, z_e$ are mutually independent and zero-mean with $\mathbb{E}[z_c^2] = \mathbb{E}[z_e^2] = 1$ and $\mathbb{E}[\eta^2] = \sigma_\eta^2$, squaring and taking expectations gives:

$$\begin{aligned} R(\varepsilon; r_{\mathrm{te}}) &= \mathbb{E}[(Y - \hat{Y})^2] \\ &= (1 - a_c - a_s \sigma_s r_{\mathrm{te}})^2 \\ &\quad + a_c^2 \sigma_\eta^2 \\ &\quad + a_s^2 \sigma_s^2 (1 - r_{\mathrm{te}}^2). \end{aligned} \tag{39}$$

This decomposition is transparent: the first term is the squared bias from under- or over-relying on the core feature, the second is the noise penalty from using the noisy core channel, and the third is the variance penalty from loading on the environment feature whose correlation shifts.

We now prove each claim.

**Claim (i): Strict benefit under correlation shift.** Let $r = r_{\mathrm{tr}} \neq 0$, $r_{\mathrm{te}} = 0$, $\lambda > 0$. Define $t := \varepsilon^2$; note that $t$ increases as contraction weakens (larger $\varepsilon$), so it suffices to show $R$ is strictly increasing

in $t$. Setting $r_{\text{te}} = 0$, the risk simplifies to $R(\varepsilon; 0) = (1 - a_c)^2 + a_c^2\sigma_\eta^2 + a_s^2\sigma_s^2$. We show each summand is strictly increasing in $t$.

For the spurious term: $a_s(t) = c_1 t/(c_2 + c_3 t)$ with $c_1 = r\sigma_s(\sigma_\eta^2 + \lambda)$, $c_2 = \lambda(1 + \sigma_\eta^2 + \lambda) > 0$, and $c_3 = \sigma_s^2(1 + \sigma_\eta^2 + \lambda - r^2) > 0$. Its derivative $c_1 c_2/(c_2 + c_3 t)^2$ has sign $\text{sign}(r)$ and never vanishes, so $a_s^2$ is strictly increasing in $t$.

For the core term, write $a_c(t) = N(t)/\Delta(t)$ where both $N(t)$ and $\Delta(t)$ are affine in $t$. Direct differentiation gives:

$$a_c'(t) = \frac{N'(t)\,\Delta(t) - N(t)\,\Delta'(t)}{\Delta(t)^2}. \tag{40}$$

Since both functions are affine, the numerator $N'\Delta - N\Delta' = n_1 d_0 - n_0 d_1$ is constant in $t$. Substituting $n_0 = \lambda$, $n_1 = \sigma_s^2(1 - r^2)$, $d_0 = \lambda(1 + \sigma_\eta^2 + \lambda)$, and $d_1 = \sigma_s^2(1 + \sigma_\eta^2 + \lambda - r^2)$:

$$\begin{aligned} n_1 d_0 - n_0 d_1 &= \sigma_s^2(1 - r^2)\,\lambda(1 + \sigma_\eta^2 + \lambda) - \lambda\sigma_s^2(1 + \sigma_\eta^2 + \lambda - r^2) \\ &= -\lambda\sigma_s^2 r^2(\sigma_\eta^2 + \lambda) < 0. \end{aligned} \tag{41}$$

Since $\Delta(t)^2 > 0$, we conclude $a_c'(t) < 0$, so $a_c$ is strictly decreasing in $t$. Since $a_c < 1/(1 + \sigma_\eta^2) =: a^*$ (the test-optimal core coefficient), and $f(a) := (1 - a)^2 + \sigma_\eta^2 a^2$ has $f'(a) < 0$ for $a < a^*$, the composition $f(a_c(t))$ is strictly increasing in $t$. Summing the two strictly increasing terms shows $R(\varepsilon; 0)$ is strictly increasing in $t = \varepsilon^2$, hence strictly decreasing as $\varepsilon$ decreases (stronger contraction).

**Claim (ii): Regularization is necessary.** As $\lambda \to 0^+$, $\Delta_\varepsilon \to \varepsilon^2\sigma_s^2(1 + \sigma_\eta^2 - r^2)$ and the factors of $\varepsilon^2$ cancel in both expressions of equation 37:

$$\begin{aligned} a_c &\to \frac{1 - r^2}{1 + \sigma_\eta^2 - r^2}, \\ a_s &\to \frac{r\sigma_\eta^2}{\sigma_s(1 + \sigma_\eta^2 - r^2)}. \end{aligned} \tag{42}$$

Both coefficients, hence the entire fitted predictor, become independent of $\varepsilon$. This confirms that without regularization, the downstream probe can freely rescale its weight along the contracted direction by $1/\varepsilon$, undoing the contraction entirely.

**Claim (iii): Alignment counterexample.** When $u_E = u_Y$, the representation collapses to one dimension $h = z_c + \eta + \sigma_s z_e$ (taking $r = 0$ for clarity, $\sigma_\eta^2 = \sigma_s^2 = 1$, $\lambda = 0.1$). The contraction rescales this single coordinate by $\varepsilon$. The ridge coefficient on the unshrunk feature is:

$$v(\varepsilon) = \frac{\varepsilon^2}{3\varepsilon^2 + 0.1}, \tag{43}$$

giving:

$$R(\varepsilon) = 1 - 2v(\varepsilon) + 3v(\varepsilon)^2. \tag{44}$$

Numerically, $R(1) \approx 0.667$ while $R(0.1) \approx 0.864$: contraction strictly hurts because the contracted direction carries the core signal. This confirms that the separation between $u_E$ and $u_Y$ is essential for the mechanism to succeed. ■

## B DFR PROTOCOL

We follow Deep Feature Reweighting (DFR; Kirichenko et al. 2023), which retrains only the last layer of a fixed encoder and thereby reweights the features the encoder already exposes. Given an encoder checkpoint $f_\theta$, the protocol is:

1. **Freeze and extract.** Freeze $f_\theta$, discard its original classification head, and extract pooled representations $\mathbf{h}_G = f_\theta(G)$ for the training, validation, and test graphs.
2. **Group-balanced reweighting set.** Define groups by the pair (label, environment). Keep all graphs of the smallest group and randomly subsample every other group to the same size, yielding a balanced set $\hat{\mathcal{D}}$ drawn from the group-labeled validation split; group labels are used only here and never during encoder training.
3. **Standardize.** Standardize each latent dimension with the mean and standard deviation computed on $\hat{\mathcal{D}}$, and apply the same transform to validation and test representations.
4. **Sparse linear head.** Fit a logistic-regression head on $\hat{\mathcal{D}}$ with $\ell_1$ regularization, $\min_{W,b} \sum_{(\mathbf{h},y)\in\hat{\mathcal{D}}} \ell(W\mathbf{h} + b,\, y) + \lambda\|W\|_1$. The sparsity penalty suits the regime $|\hat{\mathcal{D}}| \ll d$ and pushes the coefficients of irrelevant or spurious directions toward zero. The inverse strength $C \propto 1/\lambda$ is selected from $\{1, 0.7, 0.3, 0.1, 0.07, 0.03, 0.01\}$ by worst-group accuracy on a held-out half of the validation split, after which the head is refit on the full balanced set.
5. **Average over resamples.** Steps 2–4 are repeated 10 times with independently resampled balanced subsets, and the resulting weights and biases are averaged, $W_{\mathrm{DFR}} = \frac{1}{10}\sum_{r=1}^{10} W^{(r)}$, which stabilizes the head.

The final predictor is $\hat{y} = \arg\max W_{\mathrm{DFR}} f_\theta(G) + b_{\mathrm{DFR}}$, evaluated on the OOD test split. In Figure 2, $f_\theta$ is a GIN encoder taken from checkpoints at epochs $\{10, 30, 60, 90, 120, 150\}$ of either supervised ERM training or unsupervised VGAE training on GOOD-Motif concept shift; the *ERM* curve reports the accuracy of the ERM model with its original head. All numbers are means and standard deviations over 3 seeds. DFR does not modify the representation: it changes only how the frozen features are weighted, so its accuracy measures what the encoder already makes linearly accessible.

## C ADDITIONAL REAL-WORLD RESULTS

The real-world experiments freeze the learned graph encoders and fit linear probes using 5%, 10%, or 20% of the labeled training split. Model and configuration selection uses validation performance, and the test split is reserved for final evaluation. This section expands the representative analyses in the main text across all available molecular shifts.

### C.1 HYPERPARAMETER SENSITIVITY ACROSS DRUGOOD

Tables 3, 4, and 5 list the six highest-scoring evaluated $(\alpha, \beta, K)$ configurations for each of the six DrugOOD settings at 5%, 10%, and 20% labeled-training ratios, respectively. Within each dataset–domain block, rows are ordered by test ROC-AUC at that ratio; untested configurations are not included or interpolated.

### C.2 COMPLETE TRAINING DYNAMICS

Across the complete set of runs, the three objectives remain bounded, although their scales and transient behavior depend on the dataset and split. These plots are optimization diagnostics: stable losses rule out persistent numerical divergence but do not by themselves establish invariance or causal removal of environment information.

Table 3: **Top evaluated pseudo-environment configurations at 5% labels.** Within each dataset–domain setting, the six displayed $(\alpha, \beta, K)$ configurations are ordered from highest to lowest test ROC-AUC (%). Untested configurations are not interpolated.

| Dataset | Domain | $\alpha$ | $\beta$ | $K$ | Test ROC-AUC (%) |
|---|---|---|---|---|---|
| EC50 | Assay | 0 | 10 | 3 | 69.84 |
| | | 0 | 10 | 20 | 65.70 |
| | | 0.5 | 10 | 20 | 65.63 |
| | | 0 | 10 | 5 | 65.61 |
| | | 0.5 | 10 | 3 | 64.28 |
| | | 1 | 10 | 20 | 63.96 |
| | Scaffold | 1 | 10 | 20 | 63.75 |
| | | 0 | 10 | 20 | 61.71 |
| | | 0 | 0.01 | 20 | 61.00 |
| | | 0 | 10 | 5 | 60.30 |
| | | 0 | 1 | 20 | 60.27 |
| | | 0 | 0 | 20 | 50.0 |
| | Size | 1 | 10 | 20 | 58.83 |
| | | 1 | 10 | 10 | 58.65 |
| | | 0.5 | 10 | 3 | 58.59 |
| | | 0 | 10 | 3 | 58.24 |
| | | 0 | 10 | 10 | 57.32 |
| | | 0 | 10 | 5 | 57.14 |
| IC50 | Assay | 1 | 10 | 20 | 65.70 |
| | | 0 | 10 | 5 | 64.87 |
| | | 0.5 | 1 | 3 | 64.74 |
| | | 1 | 10 | 3 | 64.38 |
| | | 0.5 | 10 | 3 | 64.32 |
| | | 0 | 0.1 | 3 | 64.24 |
| | Scaffold | 0 | 10 | 3 | 60.57 |
| | | 0.5 | 10 | 3 | 60.40 |
| | | 0 | 10 | 5 | 59.46 |
| | | 0.5 | 10 | 5 | 59.23 |
| | | 0 | 10 | 10 | 58.53 |
| | | 1 | 10 | 5 | 58.48 |
| | Size | 0 | 10 | 20 | 58.22 |
| | | 1 | 10 | 20 | 58.14 |
| | | 0 | 10 | 5 | 58.11 |
| | | 0.5 | 10 | 3 | 57.83 |
| | | 0 | 10 | 3 | 57.71 |
| | | 0.5 | 1 | 20 | 57.55 |

Table 4: **Top evaluated pseudo-environment configurations at 10% labels.** Within each dataset–domain setting, the six displayed $(\alpha, \beta, K)$ configurations are ordered from highest to lowest test ROC-AUC (%). Untested configurations are not interpolated.

| Dataset | Domain | $\alpha$ | $\beta$ | $K$ | Test ROC-AUC (%) |
|---|---|---|---|---|---|
| EC50 | Assay | 0 | 10 | 3 | 70.38 |
| | | 0 | 10 | 10 | 68.45 |
| | | 0 | 10 | 5 | 68.32 |
| | | 0.5 | 10 | 5 | 67.17 |
| | | 0 | 0.1 | 3 | 66.59 |
| | | 0.5 | 1 | 20 | 66.39 |
| | Scaffold | 1 | 10 | 10 | 61.03 |
| | | 0 | 10 | 10 | 61.00 |
| | | 0.5 | 10 | 10 | 60.99 |
| | | 0 | 0.01 | 20 | 60.53 |
| | | 1 | 10 | 20 | 60.31 |
| | | 0 | 0 | 20 | 49.9 |
| | Size | 1 | 10 | 20 | 60.90 |
| | | 0 | 10 | 3 | 60.06 |
| | | 0.5 | 10 | 3 | 59.62 |
| | | 0.5 | 1 | 20 | 59.16 |
| | | 1 | 10 | 10 | 59.11 |
| | | 0 | 1 | 20 | 59.06 |
| IC50 | Assay | 1 | 10 | 20 | 66.90 |
| | | 0 | 0.1 | 3 | 65.77 |
| | | 0.5 | 1 | 3 | 65.48 |
| | | 0 | 10 | 20 | 65.16 |
| | | 1 | 10 | 3 | 65.00 |
| | | 0 | 10 | 5 | 64.88 |
| | Scaffold | 0 | 10 | 3 | 61.52 |
| | | 1 | 10 | 5 | 61.11 |
| | | 0.5 | 10 | 3 | 60.94 |
| | | 1 | 10 | 20 | 60.84 |
| | | 0.5 | 10 | 10 | 60.83 |
| | | 0 | 10 | 20 | 60.55 |
| | Size | 0 | 10 | 20 | 59.44 |
| | | 0.5 | 1 | 20 | 59.44 |
| | | 1 | 10 | 20 | 59.18 |
| | | 0.5 | 1 | 3 | 58.94 |
| | | 0.5 | 10 | 3 | 58.84 |
| | | 0 | 10 | 3 | 58.74 |

Table 5: **Top evaluated pseudo-environment configurations at 20% labels.** Within each dataset–domain setting, the six displayed $(\alpha, \beta, K)$ configurations are ordered from highest to lowest test ROC-AUC (%). Untested configurations are not interpolated.

| Dataset | Domain | $\alpha$ | $\beta$ | $K$ | Test ROC-AUC (%) |
|---|---|---|---|---|---|
| EC50 | Assay | 0 | 10 | 3 | 73.07 |
| | | 0.5 | 10 | 3 | 71.55 |
| | | 0 | 10 | 5 | 69.67 |
| | | 0.5 | 1 | 20 | 69.25 |
| | | 0 | 1 | 3 | 68.31 |
| | | 0.5 | 1 | 3 | 68.31 |
| | Scaffold | 0 | 10 | 10 | 61.23 |
| | | 0.5 | 10 | 10 | 60.59 |
| | | 0 | 0.01 | 20 | 60.42 |
| | | 0 | 10 | 20 | 60.06 |
| | | 0 | 10 | 5 | 59.93 |
| | | 0 | 0 | 20 | 53.1 |
| | Size | 0.5 | 10 | 3 | 61.61 |
| | | 0 | 10 | 3 | 60.71 |
| | | 1 | 10 | 20 | 60.71 |
| | | 0 | 10 | 20 | 59.84 |
| | | 0.5 | 10 | 5 | 59.76 |
| | | 0 | 1 | 3 | 59.53 |
| IC50 | Assay | 1 | 10 | 20 | 67.39 |
| | | 0 | 0.1 | 3 | 66.60 |
| | | 0 | 10 | 20 | 66.29 |
| | | 0.5 | 1 | 3 | 66.22 |
| | | 1 | 10 | 3 | 65.56 |
| | | 0 | 10 | 5 | 65.25 |
| | Scaffold | 0.5 | 10 | 3 | 61.92 |
| | | 0.5 | 10 | 10 | 61.74 |
| | | 0 | 10 | 3 | 61.54 |
| | | 0 | 10 | 5 | 61.47 |
| | | 0.5 | 10 | 5 | 61.46 |
| | | 0 | 10 | 20 | 61.41 |
| | Size | 0.5 | 1 | 20 | 60.73 |
| | | 0 | 10 | 20 | 60.09 |
| | | 1 | 10 | 20 | 60.04 |
| | | 0.5 | 1 | 3 | 59.91 |
| | | 0 | 1 | 20 | 59.43 |
| | | 0 | 10 | 3 | 59.36 |

Table 6: **GOOD-Motif concept hyperparameter sensitivity.** The evaluated configurations use fixed $K = 8$ and $k_{\text{adv}} = 2$. Test accuracy is the best-epoch linear-probe result.

| $\alpha$ | $\beta$ | Test accuracy (%) |
|---|---|---|
| 0 | 0.01 | 65.0 |
| 0 | 0.1 | 62.8 |
| 0 | 1 | 66.0 |
| 0 | 10 | 61.7 |
| 0.5 | 0.01 | 62.3 |
| 0.5 | 0.1 | 69.6 |
| 0.5 | 1 | 63.6 |
| 0.5 | 10 | 53.8 |
| 1.0 | 0.01 | 65.3 |
| 1.0 | 0.1 | 69.6 |

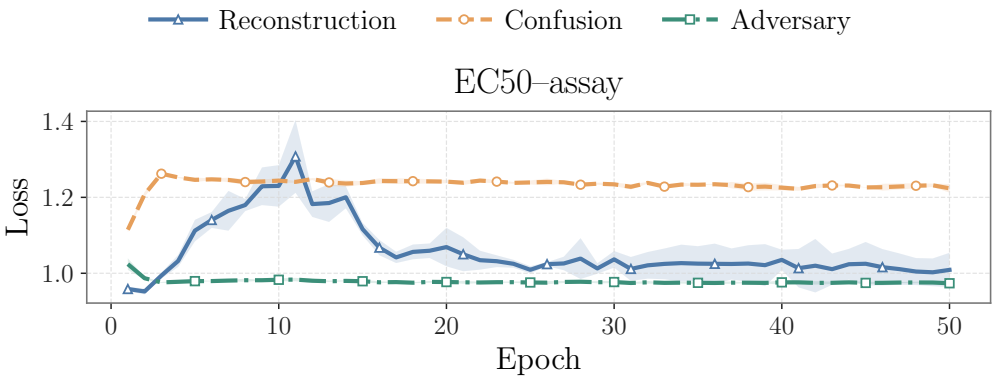


Figure 5: **Training dynamics on EC50-assay.** The plot uses the validation-selected $(\alpha, \beta, K) = (0, 10, 3)$. Reconstruction, confusion, and adversary losses use solid, dashed, and dash-dotted lines with triangle, circle, and square markers, respectively. Lines are means over three seeds and shaded regions denote mean $\pm$ sample standard deviation.

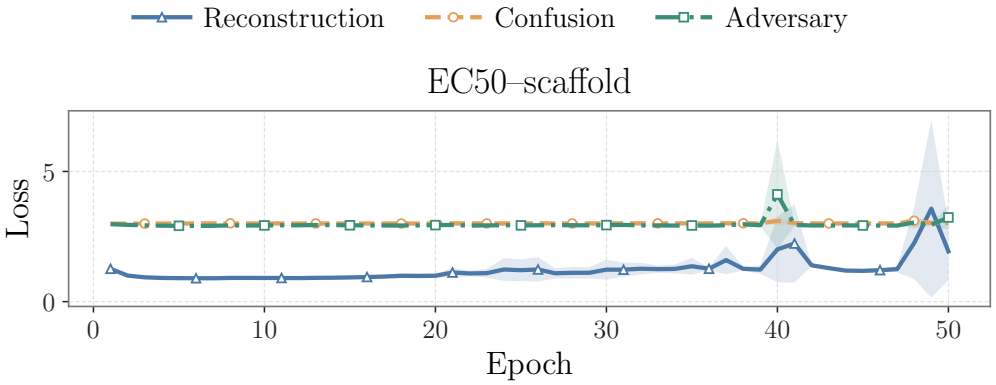


Figure 6: **Training dynamics on EC50-scaffold.** The validation-selected $(\alpha, \beta, K)$ configuration is $(1, 10, 20)$. Lines and shaded regions follow the convention of Figure 5.

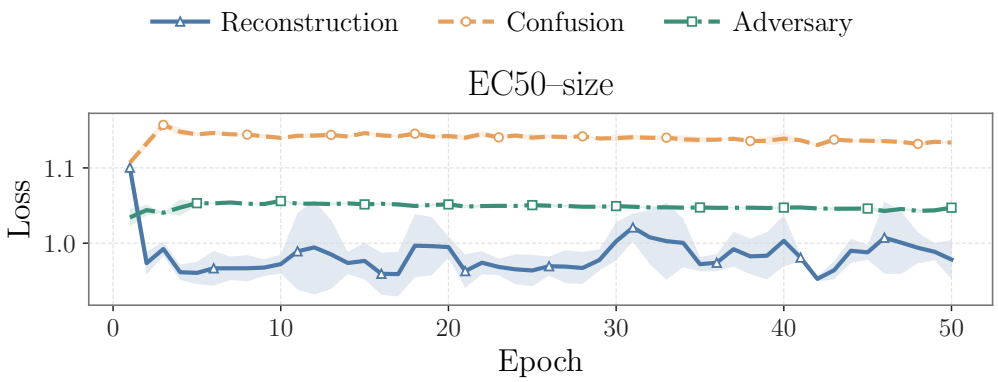


Figure 7: **Training dynamics on EC50-size.** The validation-selected $(\alpha, \beta, K)$ configuration is $(0.5, 10, 3)$. Lines and shaded regions follow the convention of Figure 5.

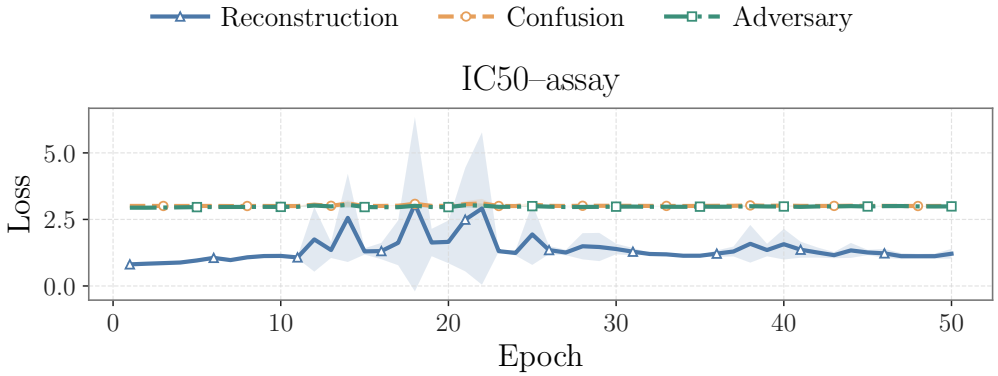


Figure 8: **Training dynamics on IC50-assay.** The validation-selected $(\alpha, \beta, K)$ configuration is $(1, 10, 20)$. Lines and shaded regions follow the convention of Figure 5.

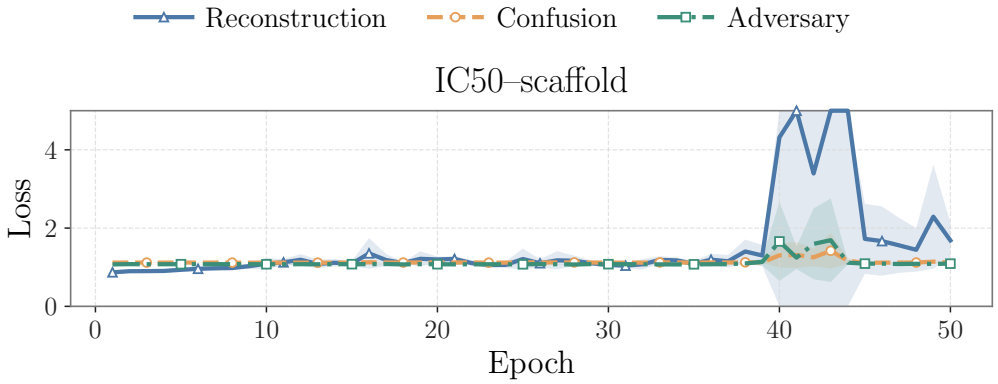


Figure 9: **Training dynamics on IC50-scaffold.** The validation-selected $(\alpha, \beta, K)$ configuration is $(0.5, 10, 3)$. The vertical axis is truncated at 5 to show the stable-loss range; values above this limit are placed at the boundary and connected to the next sampled vertex. Lines and shaded regions follow the convention of Figure 5.

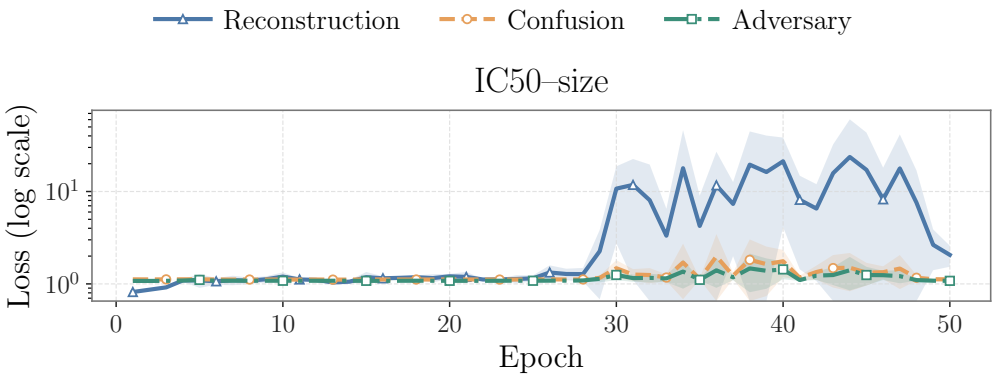


Figure 10: **Training dynamics on IC50-size.** The validation-selected $(\alpha, \beta, K)$ configuration is $(0.5, 10, 3)$. The vertical axis is logarithmic because reconstruction has a substantially larger transient scale. Lines and shaded regions follow the convention of Figure 5.

## D AI USAGE STATEMENT

The authors used generative AI tools to polish the writing, to assist with the literature survey, to assist with the implementation of experiments, and to check mathematical derivations. All research ideas, method design, experimental setups, results, and conclusions were developed and verified by the authors, who take full responsibility for the content of this paper.