MAPPERTURB: DIMENSION AND DIRECTION IN COMPACT GNN ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Low-dimensional weight maps reduce the trainable parameters of a GNN, but their coordinate count does not determine which prediction changes they can realize. We study how update directions affect the benefit of adding coordinates. MAPPERTURB parameterizes weight updates around a pretrained anchor through a fixed affine map. We analyze first-order changes in centered logits, separating the dimension of the reachable response space from its alignment with a specified target response. On three citation graphs, gradient directions yield larger reconstruction gains from additional coordinates than random directions at matched weight-update budgets. This gradient-direction advantage persists under direct radius-constrained fitting. Additionally, weight-space retention and held-out response reconstruction also rank the tested maps differently. Low-rank gradient factors retain a reconstruction advantage over spectrum-matched random factors, use 12.97–$15.46\times$ fewer stored basis values than dense maps, and match or exceed LoRA, VeRA, and full encoder tuning on further graphs. In transductive few-shot transfer, directions constructed from DGI or masked-feature reconstruction without task labels improve pooled accuracy over head-only adaptation by 2.65 and 3.19 percentage points, respectively, with eight trainable encoder coordinates and a full linear head.

1 INTRODUCTION

Adapting a pretrained graph neural network (GNN) from few labels requires choosing which parameters to learn. Adapters, low-rank updates, and graph prompts reduce trainable parameters (Hu et al., 2022; Li et al., 2024; Fang et al., 2023; Fu et al., 2025). Another approach optimizes low-dimensional coordinates and maps them into weight space (Li et al., 2018a; Sen & Mukherjee, 2026). This reduces optimization state, but maps with the same coordinate count can reach different prediction changes.

Random-subspace studies address how many coordinates are needed to recover a target performance level (Li et al., 2018a; Aghajanyan et al., 2021). LoRA-GA, PiSSA, and GiVA instead use gradients or pretrained weights to choose update directions (Wang et al., 2024; Meng et al., 2024; Gangwar et al., 2026). In a GNN, features, message passing, nonlinearities, and the classifier can make independent weight directions induce redundant prediction changes or leave predictions unchanged. We ask how orientation changes the benefit of adding coordinates, and whether retaining a target weight update also retains its prediction response.

We introduce MAPPERTURB as a framework for studying this question and constructing compact updates. Given pretrained weights θ^*, which we call the anchor, a fixed map E converts a trainable coordinate vector z into a weight update:

$$\theta(z) = \theta^* + \beta E z, \qquad z \in \mathbb{R}^d, \qquad d \ll P_{\mathrm{w}}, \tag{1}$$

where P_{w} is the number of mapped weights and $\beta > 0$ sets the update scale. Each coordinate weights one column of E: d counts the trainable coordinates, and $\mathrm{col}(E)$ is their reachable weight subspace. With the classifier fixed, their first-order effect on a graph is

$$T_G = C J_G E. \tag{2}$$

Here J_G is the logit Jacobian at the anchor on a specified node set, and C centers each node's logits across classes, removing shifts that leave softmax probabilities unchanged. The columns of T_G are

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.

directions of first-order change in these centered logits, with $\mathrm{rank}(T_G) \le \mathrm{rank}(E) \le d$. Neither weight-space overlap nor the dimension of this image determines how well the map reconstructs a particular target response. We fit coordinates to the centered-logit changes of a separately fitted reference network, then evaluate reconstruction on held-out nodes. This gives different maps a common prediction target; separate classification experiments measure task accuracy.

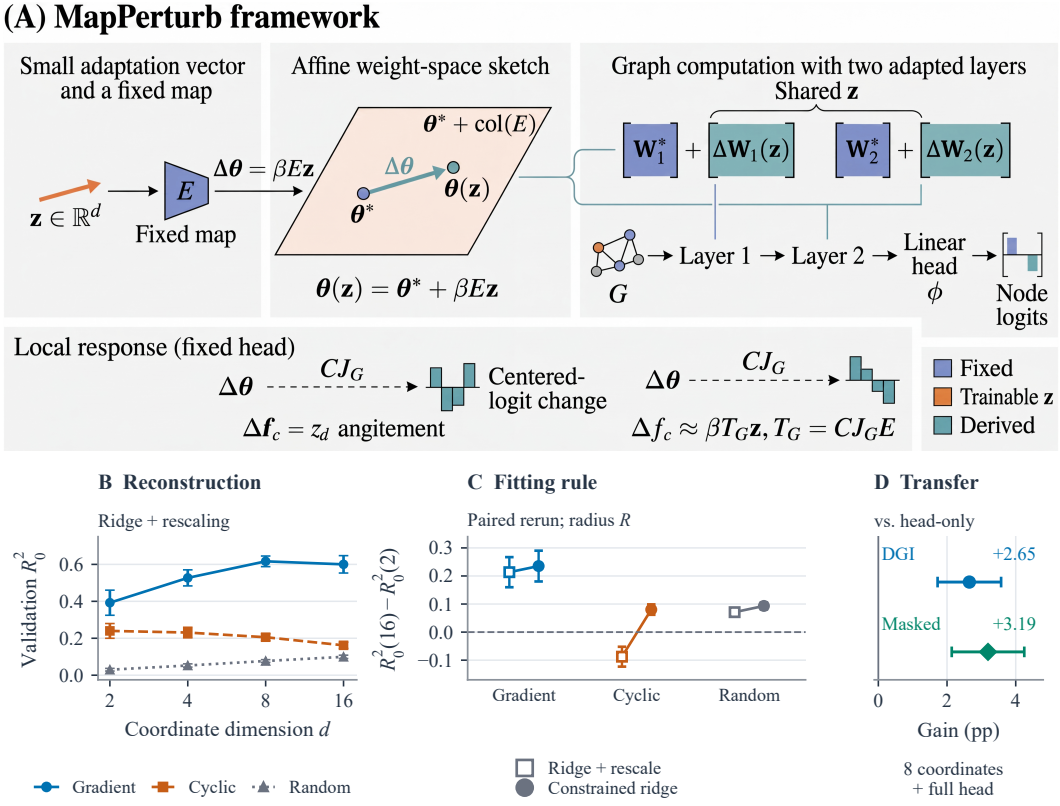

Figure 1: From weight directions to graph-conditioned predictions. (A) An anchor-centered map turns coordinates into weight updates; the graph-conditioned Jacobian maps these to first-order centered-logit changes. The weight-space plane and logits are schematic. (B) Reconstruction R_0^2 versus coordinate dimension for gradient, random, and cyclic bases under ridge-and-rescaling fitting. (C) The corresponding $R_0^2(16) - R_0^2(2)$ gains in a paired rerun with reference and candidate radii both R; constrained fitting reverses the cyclic trend. (D) Few-shot accuracy gains of label-free DGI and masked-feature irections over head-only adaptation. The head is fixed in B/C and jointly trained with z in D. Bars are 95% t intervals.

We factor gradient directions to reduce basis storage and construct label-free directions for few-shot adaptation. Figure 1 connects the map to the response and task measurements.

Our contributions can be summarized as follows:

- We characterize how update directions change the benefit of additional coordinates. Through graph-conditioned response analysis and controlled experiments, we show that weight-space retention and held-out response reconstruction can rank the tested maps differently.
- We construct compact gradient-factor maps that preserve a reconstruction advantage over spectrum-matched random factors, use 12.97–15.46× fewer stored basis values than dense maps, and match or exceed LoRA, VeRA, and full encoder tuning on further graphs.
- We construct directions from DGI and masked-feature objectives without task labels. With eight encoder coordinates and a full linear head, both improve pooled few-shot accuracy over head-only and objective-matched random controls across three citation graphs.

2 Related Work

Low-dimensional optimization and weight generation. Random-subspace training estimates the coordinate budget needed to recover a target performance level (Li et al., 2018a; Aghajanyan et al., 2021). Mode connectivity and trajectory analyses characterize structure in weight space (Garipov et al., 2018; Frankle et al., 2020; Mao et al., 2024), while hypernetworks and mapping networks generate weights from auxiliary networks or latent vectors (Ha et al., 2017; Sen & Mukherjee, 2026). Our fixed affine parameterization lets us vary the weight subspace and measure the prediction changes it supports on a graph.

Parameter-efficient fine-tuning. Adapters introduce trainable modules, and LoRA learns low-rank weight updates (Houlsby et al., 2019; Hu et al., 2022). rsLoRA studies rank-dependent scaling; VeRA learns scaling vectors over shared frozen matrices (Kalajdzievski, 2023; Kopiczko et al., 2024). LoRA-GA initializes low-rank factors to approximate the full fine-tuning gradient at the first step (Wang et al., 2024); PiSSA initializes from principal components of pretrained weights (Meng et al., 2024). GiVA uses gradient-informed frozen bases and learns scaling vectors (Gangwar et al., 2026). Our compact maps combine multiple fixed gradient atoms with a coordinate vector shared across encoder matrices. The response experiments compare these directions at matched coordinate counts and weight-update radii, separating their weight-space and prediction-space behavior. Related work studies gradient intrinsic dimensionality and Fisher subspaces (Ye et al., 2026; Feng et al., 2026).

Adaptation and pretraining on graphs. GraphLoRA, G-Adapter, AdapterGNN, and EdgePrompt adapt GNNs through low-rank updates, adapters, and edge prompts (Yang et al., 2025; Gui et al., 2024; Li et al., 2024; Fu et al., 2025). Other approaches use meta-learning, augmentation, or graph prompts (Zhou et al., 2019; Ding et al., 2022; Wang et al., 2021; Fang et al., 2023; Liu et al., 2023). We use DGI (Veličković et al., 2019) and masked-feature reconstruction to derive directions before introducing few-shot labels. The experiments isolate the contribution of these directions on a common backbone.

Graph-conditioned responses. GNN expressiveness studies which graph structures an architecture can distinguish (Xu et al., 2019). We study local changes in predictions around a pretrained model. Node features, message passing, and depth enter through the logit Jacobian (Kipf & Welling, 2017; Veličković et al., 2018; Li et al., 2018b; Chen et al., 2020). The resulting tangent describes which responses are reachable and how well they align with a target.

3 Low-Dimensional Weight Maps

MapPerturb denotes the framework of shared coordinates over fixed weight directions around a pretrained anchor. Its tested instances include cyclic, random, dense-gradient, and compact gradient-factor maps. We also construct gradient factors without task labels. The classifier is fixed for response analysis and trained with the coordinates for primary few-shot transfer.

3.1 Weight directions and a common radius

Using the output-first weight convention, a GCN layer is $H_l = \sigma(\widehat{A}H_{l-1}W_l^\top)$ and $W_l \in \mathbb{R}^{o_l \times i_l}$. The input-shared cyclic map uses

$$W_l(z) = W_l^* + \beta \mathbf{1}_{o_l} v_l(z)^\top, \qquad [v_l(z)]_i = z_{i \bmod d}. \tag{3}$$

The same z is used in every selected layer. At fixed input activations, its direct pre-activation update is $\beta(\widehat{A}H_{l-1}v_l(z))\mathbf{1}_{o_l}^\top$, a graph-filtered scalar broadcast across channels. Input indices start at zero. If coordinate j is copied m_j times, then $\|Ez\|_2^2 = \sum_j m_j z_j^2$. Equal norms of z can therefore produce different weight-update norms across maps.

To compare directions at a common weight-update radius, we express each span in orthonormal coordinates, absorbing the scale into the parameterization:

$$\Delta\theta = Qz, \qquad Q^\top Q = I_d, \qquad \|z\|_2 \le r_{\mathrm{w}}. \tag{4}$$

Column normalization preserves the cyclic span. Random bases are obtained by thin QR decomposition of Gaussian directions. The gradient basis uses the top d right singular vectors of a matrix containing 16 normalized encoder gradients from class-balanced subsets of discovery nodes. A fold is invalid if the source gradients have rank below d.

The mapped encoder requires d trainable values, d gradient values, and $2d$ persistent Adam moments. Basis storage is additional to the common frozen anchor: a dense basis stores $P_{\mathrm{w}}d$ values, whereas the cyclic map is represented implicitly. These counts exclude activation storage, temporary construction memory, and the cost of message passing.

3.2 Compact gradient factors

To reduce the storage of a dense gradient basis, we compress the discovery gradients before selecting their shared coordinates. Let g_t, $t = 1, \ldots, M$ with $M = 16$, be normalized discovery gradients. For each layer, reshape their blocks into $G_{t,l}$ and retain rank-r SVD factors,

$$G_{t,l} \approx L_{t,l} R_{t,l}^{\top}, \qquad L_{t,l} \in \mathbb{R}^{o_l \times r}, \quad R_{t,l} \in \mathbb{R}^{i_l \times r}. \tag{5}$$

We split the retained singular values symmetrically between the factors. Concatenating the reconstructed layer blocks and normalizing the result gives one gradient *atom*; we absorb this normalization into its factors. Stack the atoms in $A_r \in \mathbb{R}^{P_{\mathrm{w}} \times M}$. If $A_r^{\top} A_r V_d = V_d \Lambda_d$ is its top-d positive eigensystem, the weight-mixed map is

$$K_r = V_d \Lambda_d^{-1/2}, \qquad Q_r = A_r K_r, \qquad Q_r^{\top} Q_r = I_d. \tag{6}$$

The map stores the factors and the mixing matrix K_r. For $\alpha = K_r z$, the layer update and the number of stored frozen values are

$$\Delta W_l(z) = \sum_{t=1}^{M} \alpha_t L_{t,l} R_{t,l}^{\top}, \tag{7}$$

$$N_{\mathrm{compact}}(r) = Mr \sum_{l} (o_l + i_l) + Md. \tag{8}$$

The rank r applies to each atom's layer block. Their sum can have higher rank. Equation (8) counts the retained factors and mixing coefficients; construction still requires full-weight gradients, and the diagnostics temporarily materialize dense bases.

A matched random control randomizes each atom's left and right orthonormal frames while preserving its layerwise singular values and factor shapes from the same discovery objective. Storage, coordinate dimension, optimizer, and scale or radius settings are matched; graph-conditioned tangent spectra may differ.

3.3 Label-free gradient directions

To construct directions without task labels, we collect gradients at a DGI-pretrained anchor using either DGI's discrimination objective (Veličković et al., 2019) or masked-feature reconstruction. DGI uses random feature permutations; masked reconstruction uses a decoder trained on frozen encoder representations. Independent corruptions yield 16 atoms, and discovery auxiliary modules are discarded afterward. Both objectives use the same anchor, features, and edges; masked reconstruction tests directions whose discovery objective differs from pretraining.

In transfer, each atom's layer block is normalized separately after rank-two truncation. We stack these atoms in $\widetilde{A}_r$ and compute K from its top-d Gram eigensystem, with eigenvalues floored at 10^{-6} for numerical stability. We then apply fixed layer gains $s_l = \sqrt{o_l i_l / d}$, so the realized encoder lift is

$$E = D_s \widetilde{A}_r K, \qquad D_s = \mathrm{blockdiag}(s_l I_{o_l i_l}). \tag{9}$$

Few-shot adaptation initializes $z = 0$ and jointly trains the coordinates and a full linear classifier with $d = 8$, $\beta = 0.01$, and no projection onto a coordinate ball. Discovery is label-free; adaptation uses the few-shot labels. This protocol shares scale settings across methods without forcing equal update norms. Supervised reconstruction uses globally normalized atoms. For comparisons of functional responses, we orthonormalize each evaluated encoder span and impose a common weight radius.

Algorithm 1 Label-free map construction and few-shot adaptation

Require: Anchor encoder θ^* and common initial head ϕ_0; graph G; target layers $\mathcal{I}$; few-shot set S_k; discovery loss $\mathcal{J}$
Ensure: Frozen factors (L, R, K), adapted coordinates z_T, and classifier ϕ_T
$M \leftarrow 16$, $r \leftarrow 2$, $d \leftarrow 8$, $\beta \leftarrow 0.01$, $T \leftarrow 100$; $\mathcal{J} \in \{\mathcal{J}_{\mathrm{DGI}}, \mathcal{J}_{\mathrm{mask}}\}$
Fix discovery auxiliary parameters (fit a masked decoder on frozen features if needed)
for $t = 1, \dots, M$ **do**
  $\xi_t \sim \mathcal{P}_{\mathcal{J}}$; $g_t \leftarrow \nabla_{\theta_{\mathcal{I}}} \mathcal{J}(\theta^*; G, \xi_t)$; $g_t \leftarrow g_t / \max(\|g_t\|_2, 10^{-12})$
  for $l \in \mathcal{I}$ **do**
    $(U, \Sigma, V) \leftarrow \mathrm{SVD}_r(\mathrm{reshape}(g_{t,l}))$; $\nu_{t,l} \leftarrow \max(\|\Sigma\|_F, 10^{-12})$
    $L_{t,l} \leftarrow U\Sigma^{1/2}/\sqrt{\nu_{t,l}}$; $R_{t,l} \leftarrow V\Sigma^{1/2}/\sqrt{\nu_{t,l}}$ ▷ Normalize each layer block
$a_t \leftarrow \mathrm{concat}_{l \in \mathcal{I}} \mathrm{vec}(L_{t,l} R_{t,l}^\top)$, $t = 1, \dots, M$; $\widetilde{A}_r \leftarrow [a_1, \dots, a_M]$
$(V_d, \Lambda_d) \leftarrow \mathrm{TopEig}_d(\widetilde{A}_r^\top \widetilde{A}_r)$; $K \leftarrow V_d \, \mathrm{diag}_{j=1}^{d}[\max(\lambda_j, 10^{-6})^{-1/2}]$
$s_l \leftarrow \sqrt{o_l i_l / d}$; freeze (θ^*, L, R, K); discard discovery auxiliary modules
$w_0 \leftarrow (0, \phi_0)$; $(m_0, v_0) \leftarrow (0, 0)$ ▷ $w = (z, \phi)$ and Adam moments
for $q = 0, \dots, T-1$ **do**
  $\Delta W_l(z_q) \leftarrow \beta s_l \sum_{t=1}^{M} (K z_q)_t L_{t,l} R_{t,l}^\top$, $l \in \mathcal{I}$
  $h_q \leftarrow \nabla_w \mathrm{CE}_{S_k}(f_{\theta^* + \Delta\theta(z_q), \phi_q}(G))$
  $(w_{q+1}, m_{q+1}, v_{q+1}) \leftarrow \mathrm{AdamStep}(w_q, h_q, m_q, v_q)$
return (L, R, K, z_T, ϕ_T) ▷ No coordinate-ball projection in transfer

3.4 Functional mixing and factor-rank selection

Weight mixing selects directions by their energy in parameter space. Functional mixing selects directions by local response sensitivity. We test whether this criterion improves reconstruction; it does not use the reference response when selecting directions. For an atom matrix A, write its positive Gram eigensystem as $A^\top A = U\Lambda U^\top$ and form $W = AU\Lambda^{-1/2}$. Let $J_{G,D}$ be the logit Jacobian on discovery nodes D, and let C_D remove the per-node mean across the c classes. Define

$$F_D = (|D|c)^{-1/2} C_D J_{G,D} W, \qquad H_D = F_D^\top F_D. \tag{10}$$

If S_d contains the top d eigenvectors of H_D, then $Q_F = AU\Lambda^{-1/2} S_d$ maximizes discovery centered-logit sensitivity over d orthonormal weight directions. This changes the mixing matrix while preserving the factors and stored size. The objective uses the unweighted centered-logit Jacobian.

Factor-rank selection. A separate fidelity criterion compares compact and dense responses. For the supervised compact study, an inner class-balanced cross-fit on discovery labels scores

$$\mathcal{F}(T_0, T_r) = 1 - \frac{\min_{O^\top O = I} \|T_0 - T_r O\|_F^2}{\|T_0\|_F^2}, \tag{11}$$

where $T_0, T_r \in \mathbb{R}^{m \times d}$ are dense and compact tangents, $O \in \mathbb{R}^{d \times d}$, and $\|T_0\|_F > 0$. Orthogonal alignment removes coordinate rotations while retaining amplitude differences (Appendix E.2). We average the two cross-fit directions and select the smallest rank meeting the fixed fidelity and storage criteria. Selection uses discovery data; downstream fitting labels and official validation outcomes are held out.

The fixed fidelity and storage thresholds select $r = 2$ in development, before confirmation on independent anchors. Weight and functional mixing use the same factors in that comparison. This selection concerns fidelity to the dense reference; Algorithm 1 specifies the separate label-free adaptation protocol.

4 Graph-Conditioned Responses and Target Alignment

Let $f_{G,S}(\theta)$ collect logits on nodes S in node-major order, and let $C_S = I_{|S|} \otimes (I_c - \mathbf{1}_c \mathbf{1}_c^\top / c)$. At a differentiable anchor, the graph-conditioned tangent of a weight map is

$$T_G = C_S J_{G,S}(\theta^*) E, \qquad C_S[f_{G,S}(\theta^* + \beta E z) - f_{G,S}(\theta^*)] = \beta T_G z + o(\|z\|_2). \tag{12}$$

Its rank is the local functional dimension for the specified graph, anchor, node set, and map, and is bounded by $\mathrm{rank}(E) \le d$. The column space describes reachable directions. At a finite weight

radius, response amplitudes also matter: for an orthonormal basis Q, the first-order responses form the ellipsoid $\{C_S J_{G,S} Qz : \|z\|_2 \le r_{\mathrm{w}}\}$. We use this local description to compare directions, then evaluate finite updates through the nonlinear GNN, including any activation changes.

Target alignment. For any specified centered-logit target y and tangent T, least squares gives $\min_z \|y - Tz\|_2^2 = \|(I - P_T)y\|_2^2$, where P_T is the orthogonal projector onto $\mathrm{col}(T)$. Sensitivity alone does not ensure alignment. For orthonormal e_1, e_2, the tangent $T = ae_1$ has arbitrarily large norm as a grows, yet cannot represent $y = e_2$. A radius constraint also limits response amplitudes. For nested feasible sets, the optimal fit to the same target on the same nodes cannot worsen as directions are added. Our measurements also depend on the fitting rule and evaluate nonlinear responses on held-out nodes; they need not be monotone (Appendix B.3).

An exact null direction. For any classifier $Wh + b$, replacing W by $W + \mathbf{1}_c v^\top$ adds $(v^\top h)\mathbf{1}_c$ to its logits. Softmax probabilities are unchanged. The cyclic output-head update therefore has an exactly zero centered-logit image (Proposition 1). This softmax invariance applies to the output head. Cyclic encoder updates can change the hidden representation h and thereby change class probabilities.

Alignment in a quadratic model. Consider $q(\delta) = q(0) + g^\top \delta + \frac{1}{2}\delta^\top H \delta$ with $H \succ 0$, and write $\bar{E} = \beta E$. Completing the square yields

$$\min_z q(\bar{E}z) - \min_\delta q(\delta) = \frac{1}{2}\left\|(I - P_{H^{1/2}\bar{E}})H^{-1/2}g\right\|_2^2, \tag{13}$$

where $P_{H^{1/2}\bar{E}}$ projects onto the indicated column space. The gap is determined by the curvature-scaled descent component outside the transformed subspace. Thus dimension alone does not determine the best attainable loss, even in this quadratic setting. A finite radius can add approximation error. Appendix E.1 gives the derivation and the extension to positive semidefinite curvature.

Dimension and scale. The affine map gives an exact d-coordinate optimization space whenever E has rank d, but its Lipschitz constant is $\beta\|E\|_2$. For a balanced cyclic expansion, $\sqrt{d}\|E\|_2 \approx \sqrt{P_{\mathrm{w}}}$. Thus capacity bounds depend on the lift's scale as well as d. Appendix B gives additional covering, capacity, and convergence results under tube, i.i.d., and coordinate-space Polyak–Łojasiewicz conditions, respectively; these are conditional statements, not guarantees for the transductive Adam experiments. Proofs appear in Appendix E.

5 Experiments

We test dimension–direction interactions (Q1), compact factors (Q2), label-free few-shot adaptation (Q3), and other graphs and backbones (Q4). Table 1 distinguishes the response and classification protocols.

Table 1: Evaluation protocols with transductive DGI anchors. Labels used for direction discovery, head training, and endpoints differ between studies.

Study	Direction discovery	Fitting labels / head	Endpoint / comparators
Q1/Q2; Q4 reconstruction	Training fold D_{basis}	Disjoint D_{fit}; fixed head	Validation R_0^2; random/cyclic/dense
Q3 transfer	DGI or masked features	$1, 5, 10$/class; jointly trained head	Test accuracy; head-only/matched random
Q4 classification	20 training labels/class	Same 20/class; frozen fitted head	Test accuracy; random/PEFT/full encoder

5.1 Experimental setup

The main studies use Cora, CiteSeer, and PubMed with their official Planetoid masks (Yang et al., 2016) and a two-layer width-64 GCN pretrained with DGI (Veličković et al., 2019). Pretraining and direction discovery are transductive: they use all graph features and edges. Adaptation costs exclude pretraining.

Functional reconstruction. For each dataset, we use ten DGI anchors and split the 20 official training nodes per class into two ten-node folds, D_{basis} and D_{fit}. We use D_{basis} to fit a common classifier, construct the maps, and set the radius $r_{\mathrm{w}} = \min(\|\Delta_{\mathrm{probe}}\|_2, 0.05\|\theta^*_{\mathrm{w}}\|_2)$; we then freeze the classifier and fit a reference network from the same anchor on D_{fit}, updating encoder weights within this radius. We fit ridge coefficients to its centered-logit change using a finite-difference tangent, rescale to the weight radius, and evaluate the nonlinear response on official validation nodes. Test endpoints are unused. We swap folds and average folds and random repetitions within each anchor. Zero-baseline reconstruction is

$$R_0^2 = 1 - \frac{\|\Delta f - \widehat{\Delta f}\|_F^2}{\|\Delta f\|_F^2}, \tag{14}$$

where $\Delta f \neq 0$ is the reference network's centered-logit change and $\widehat{\Delta f}$ is the nonlinear response produced by the fitted map. A zero score matches the no-change baseline; negative scores are worse. The full protocol is in Appendix C.4.

Few-shot classification. Q3 shares the anchor, initial linear head, support nodes, and optimizer random stream across methods. The coordinate vector starts at zero and is trained for 100 Adam steps ($\mathrm{lr} = 0.01$, zero weight decay). Settings are fixed before training; test accuracy is measured at the final step without validation checkpoint selection. We form paired 95% t intervals after averaging datasets within ten anchor indices for Q1/Q2 and shots within 30 dataset–anchor pairs for Q3. Holm correction applies within each comparison family. Appendix C.10 gives seed rules and complete primary results.

5.2 Dimension and orientation interact (Q1)

We vary $d \in \{2, 4, 8, 16\}$ for gradient, random, and cyclic bases at a common radius (Figure 1B), keeping the anchor, classifier, gradient bank, reference network, and evaluation nodes fixed within each fold. Gradient and cyclic spans are nested at these dimensions; random bases are sampled independently for each d.

From $d = 2$ to 16, gradient-basis R_0^2 increases by 0.208 ($[0.137, 0.278]$), random by 0.070 ($[0.062, 0.078]$), and cyclic decreases by 0.078 (change interval $[-0.118, -0.038]$). The gradient–cyclic gap therefore widens by 0.286 ($[0.194, 0.377]$). Gradient bases lead at every tested d (eight Holm-adjusted $p \leq 0.0014$).

Changing the fitting rule reveals which trends depend on the estimator. In a paired rerun (Figure 1C), the cyclic gain changes from -0.088 with ridge-and-rescaling to $+0.080$ with constrained ridge. Gradient and random gains become $+0.235$ and $+0.093$. Both gradient–control interactions remain positive at matched reference/candidate radii $\{0.5, 1, 2\}r_{\mathrm{w}}$; their intervals include zero when only the candidate radius is doubled against a fixed target (Appendix C.6). Thus the cyclic decline does not establish a loss of capacity in its nested span. Gradient directions also yield the largest fitting-CE decrease at $d = 8$ (Appendix Figure 3).

Weight-space retention and response reconstruction also differ. At $d = 8$, gradient directions capture less reference-displacement energy than cyclic directions (0.0367 versus 0.0792), yet reconstruct responses better (0.621 versus 0.221). Displacement and trajectory bases retain much more weight energy (0.4308, 0.4333) but give similar reconstruction (0.605, 0.633; Appendix Table 18).

5.3 Compact factorization of the aligned map (Q2)

Dense gradient bases store $P_{\mathrm{w}}d$ values. We test rank-two factors selected by discovery-only tangent fidelity on development anchors and fixed before confirmation on ten new anchors per dataset (Appendix C.9).

At the same storage and dimension, gradient factors outperform spectrum-matched random factors by 0.442 in R_0^2 ($[0.398, 0.485]$; Holm $p = 5.48 \times 10^{-9}$; Figure 2A).

Replacing weight mixing with functional mixing on the same gradient factors adds 0.021 ($[-0.002, 0.045]$; Holm $p = 0.074$; Figure 2B). The comparison supports retaining gradient directions; evidence for an additional benefit from functional mixing is inconclusive.

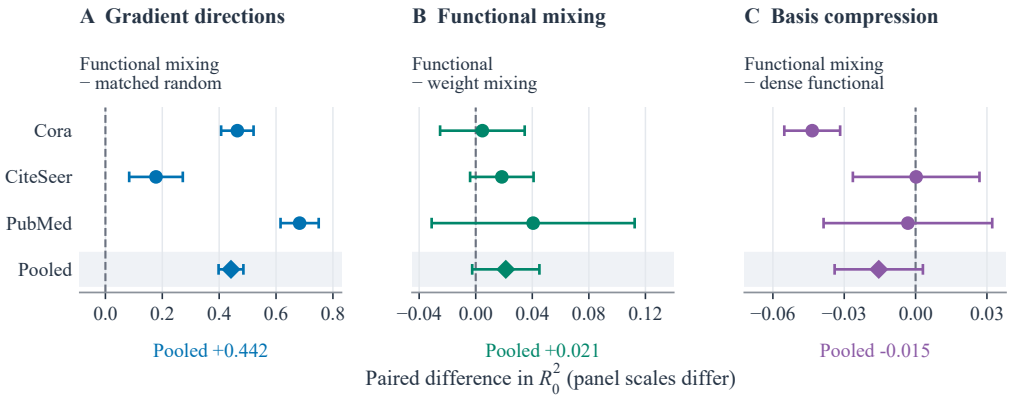

Figure 2: Paired Q2 contrasts in held-out reference-response reconstruction at $d = 8$ and factor rank $r = 2$. Compact maps with functional mixing: R_0^2 minus (A) matched random factors, (B) weight mixing, and (C) the dense functional map. Circles average ten anchors per dataset; diamonds average datasets within the same ten indices. Bars are paired 95% t intervals; horizontal scales differ across panels.

Compact maps reduce stored basis values by $14.70\times$ on Cora, $15.46\times$ on CiteSeer, and $12.97\times$ on PubMed. Their dataset-equal mean R_0^2 is 0.593, versus 0.608 for the dense functional reference. The paired difference is -0.015 ($[-0.034, 0.003]$; panel C), with the largest reduction on Cora (-0.043). The interval is not an equivalence test. Appendix Table 8 gives storage counts, absolute scores, and paired tests.

5.4 Label-free directions for few-shot adaptation (Q3)

We test label-free DGI and masked-feature directions against objective-matched random factors. Few-shot labels train the coordinates and classifier; Table 2 also reports separate reconstruction fits.

Table 2: Label-free few-shot adaptation in the primary Q3 cohort. Accuracies (%) average $k \in \{1, 5, 10\}$ within ten anchors per dataset; pooled weights datasets equally. Random controls match each objective. R_0^2 uses separate reconstruction fits with a frozen head and common radius. Cora counts exclude the shared anchor; frozen values are additional to trainable values. See text for paired 95% t intervals.

	Test accuracy (%)					Values (Cora)	
Method	Cora	CiteSeer	PubMed	Pooled	R_0^2	Trainable	Frozen
Head only	56.04	42.95	58.70	52.56	—	455	0
DGI random	56.12	43.01	58.76	52.63	0.070	463	52,128
DGI factors	59.74	45.44	60.45	55.21	0.379	463	52,128
Masked random	56.01	43.04	58.75	52.60	0.074	463	52,128
Masked factors	60.46	47.61	59.20	55.76	0.447	463	52,128

Few-shot accuracy. The pooled DGI gain over head-only adaptation is 2.65 percentage points ($[1.73, 3.57]$), and 2.58 over its random control ($[1.67, 3.49]$). Masked-feature factors improve by 3.19 and 3.16 points, respectively ($[2.14, 4.25]$ and $[2.10, 4.22]$); all four Holm-adjusted $p \leq 4.67 \times 10^{-6}$ (Figure 1D; per-shot results in Appendix C.10). The pooled benefit over both controls also holds for macro-F1 and balanced accuracy in a separate rerun (Appendix F.4).

Reconstruction with the same directions. At a common radius and fixed reference target, DGI factors exceed their random control by 0.309 in R_0^2 ($[0.195, 0.424]$), and masked-feature factors by 0.373 ($[0.303, 0.442]$). Supervised dense, supervised rank-two, and cyclic spans score 0.606, 0.583, and 0.220. Both label-free variants improve both endpoints, but their across-anchor gain correlations (-0.19 for DGI; -0.16 for masked reconstruction) do not establish a positive reconstruction–accuracy association (Appendix C.10).

In a separate paired comparison with the same fixed training budget, DGI and masked factors reach 56.00% and 56.27%, versus 53.00% for head-only, 58.37% for LoRA-$r = 4$, and 59.45% for full encoder tuning. On Cora, the map and head train 463 values but store 210,364 adapter bytes; LoRA-$r = 2$ trains 3,705 values and stores 14,820 bytes. The coordinate savings therefore trade against fixed storage, and LoRA-$r = 4, 8$ achieves better accuracy here. Appendix C.11 reports paired intervals, all ranks, and the separation from the primary Q3 cohort.

5.5 Other graphs and architectures (Q4)

On Chameleon and Squirrel, increasing d from 2 to 16 widens the gradient–cyclic gap by 0.422 under GCN ($[0.377, 0.468]$) and 0.262 under width-matched single-head GAT ($[0.200, 0.324]$). Cyclic reconstruction also improves, but gradient bases lead at every tested dimension (Appendix C.7).

A separate $d = 8$ classification study uses Chameleon, Squirrel, and Roman-empire with GCN, GraphSAGE (Hamilton et al., 2017), and multi-head GAT: 20 labels per class, a frozen fitted head, a common radius, and validation-selected learning rate and checkpoint (Appendix F.3). Across 90 graph–backbone–split combinations, the multi-gradient factor map exceeds frozen by 0.650 points ($[0.377, 0.923]$) and matched random by 0.581 ($[0.324, 0.838]$); all ten balanced split blocks used for paired intervals have positive gains. Grad-SVD uses eight shared directions from layerwise SVDs of one aggregate training gradient and exceeds multi-gradient by 0.538 points ($[0.169, 0.908]$), favoring the simpler construction. Table 3 also compares the frozen-basis controls VeRA (Kopiczko et al., 2024) and LoRA-XS (Bałazy et al., 2024).

Table 3: Accuracy (%) with a frozen fitted head and common radius across 90 graph–backbone–split combinations. Dataset columns average three backbones and ten splits; overall means weight the nine graph–backbone pairs equally. Bold marks column maxima, including ties at displayed precision.

Method	Chameleon	Squirrel	Roman-empire	Overall
Frozen fitted head	34.32	25.93	16.48	25.576
Matched random, $d = 8$	34.31	26.01	16.62	25.645
Multi-gradient, $d = 8$	34.61	**26.11**	17.96	26.226
Grad-SVD, $d = 8$	**35.57**	26.05	**18.68**	**26.764**
GiVA, $r = 4$	35.10	**26.11**	18.29	26.499
VeRA, $r = 16$	34.31	26.04	16.81	25.719
LoRA, $r = 1$	34.34	25.98	17.44	25.922
LoRA-XS, $r = 2$	34.11	26.10	16.59	25.600
Full encoder	34.71	25.94	17.92	26.186

On ogbn-arxiv (169,343 nodes), multi-gradient reaches 27.89% across five anchors, versus 24.81% for frozen and 25.64% for random; the gain intervals include zero. LoRA and full encoder average 30.84% and 31.35% (Appendix F.3).

6 Conclusion and Discussion

In this paper, we introduced MapPerturb, a framework for studying low-dimensional GNN adaptation through the weight directions controlled by shared coordinates. Our central finding is that coordinate count alone does not determine which prediction changes a map can realize, whereas the directions it spans determine the benefit of each added coordinate. Specifically, gradient-informed directions gain more from additional coordinates than random directions. Furthermore, the compact and label-free variants keep these gains while training only eight shared coordinates and remaining competitive with existing parameter-efficient and full fine-tuning methods. However, because weight-space retention and response reconstruction can disagree, a map is best evaluated at a specified update radius and response target, with task accuracy and stored basis size measured separately. A direction for future work is how to use reconstruction results to select a map rather than a fixed prior.

AI USE STATEMENT

Large language models assisted with drafting, language editing, manuscript organization, code review, and layout. Experimental results were obtained with the project's code. The authors are responsible for the scientific claims, numerical results, code, and final manuscript.

ETHICS STATEMENT

The reported experiments use public benchmark datasets. No new human-subject study or private-data collection is involved. The results concern controlled GNN adaptation; deployment in a specific application requires its own validation of accuracy, bias, and appropriate use.

REPRODUCIBILITY STATEMENT

Appendix A specifies the mappings, parameter conventions, factor storage, and rank-selection procedure. Appendix C gives the full experimental protocols, statistical units, hardware, and result tables. Complete conditional statements and proofs appear in Appendices B and E. Supplementary analyses identify exploratory and incomplete studies alongside their results. Appendix D describes numerical and reproducibility constraints. The unresolved cross-run anchor differences and the within-run pairing checks are documented in Appendix C.11. The accompanying review code includes core implementations, CPU checks, and a README with experiment entry points.

REFERENCES

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of ACL-IJCNLP*, pp. 7319–7328, 2021. doi: 10.18653/v1/2021.acl-long.568.

Klaudia Bałazy, Mohammadreza Banaei, Karl Aberer, and Jacek Tabor. LoRA-XS: Low-rank adaptation with extremely small number of parameters, 2024. URL `https://arxiv.org/abs/2405.17604`.

Peter L Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *JMLR*, 2002.

Ming Chen, Zhewei Wei, Zengfeng Huang, Bolin Ding, and Yaliang Li. Simple and deep graph convolutional networks. In *International Conference on Machine Learning*, 2020.

Kaize Ding, Jianling Wang, Jundong Li, Dingcheng Li, and Huan Liu. Few-shot graph learning: A comprehensive survey. *IEEE TPAMI*, 2022.

Taoran Fang, Yunchao Zhang, Yang Yang, Chunping Wang, and Lei Chen. Universal prompt tuning for graph neural networks. In *Advances in Neural Information Processing Systems*, 2023.

Zhi-Quan Feng, Ying-Jia Lin, and Hung-Yu Kao. Learning in the Fisher subspace: A guided initialization for LoRA fine-tuning, 2026.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. *ICML*, 2020.

Xingbo Fu, Yinhan He, and Jundong Li. Edge prompt tuning for graph neural networks. In *International Conference on Learning Representations*, 2025.

Neeraj Gangwar, Rishabh Deshmukh, Michael Shavlovsky, Hancao Li, Vivek Mittal, Lexing Ying, and Nickvash Kani. GiVA: Gradient-informed bases for vector-based adaptation. In *International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, 2026.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew G Wilson. Loss surfaces, mode connectivity, and fast ensembling of DNNs. *NeurIPS*, 2018.

Anchun Gui, Jinqiang Ye, and Han Xiao. G-Adapter: Towards structure-aware parameter-efficient transfer learning for graph transformer networks. In *AAAI Conference on Artificial Intelligence*, 2024.

David Ha, Andrew M Dai, and Quoc V Le. HyperNetworks. In *International Conference on Learning Representations*, 2017.

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

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. In *Advances in Neural Information Processing Systems*, 2020.

Damjan Kalajdzievski. A rank stabilization scaling factor for fine-tuning with LoRA, 2023. URL https://arxiv.org/abs/2312.03732.

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

Dawid J. Kopiczko, Tijmen Blankevoort, and Yuki M. Asano. VeRA: Vector-based random matrix adaptation. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.11454.

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. *ICLR*, 2018a.

Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *AAAI Conference on Artificial Intelligence*, 2018b.

Shengrui Li, Xueting Han, and Jing Bai. AdapterGNN: Parameter-efficient fine-tuning improves generalization in GNNs. In *AAAI Conference on Artificial Intelligence*, 2024.

Zemin Liu, Xingtong Yu, Yuan Fang, and Xinming Zhang. GraphPrompt: Unifying pre-training and downstream tasks for graph neural networks. In *Proceedings of the ACM Web Conference*, 2023.

Jialin Mao, Itay Griniasty, Han Kheng Teoh, Rahul Ramesh, Rubing Yang, Mark K Transtrum, James P Sethna, and Pratik Chaudhari. Microscaling: The training process of many deep networks explores the same low-dimensional manifold. *PNAS*, 2024.

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. PiSSA: Principal singular values and singular vectors adaptation of large language models. In *Advances in Neural Information Processing Systems*, 2024.

Christopher Morris, Nils M Kriege, Franka Bause, Kristian Kersting, Petra Mutzel, and Marion Neumann. TUDataset: A collection of benchmark datasets for learning with graphs. In *ICML Workshop on Graph Representation Learning*, 2020.

Hongbin Pei, Bingzhe Wei, Kevin Chen-Chuan Chang, Yu Lei, and Bo Yang. Geom-GCN: Geometric graph convolutional networks. In *International Conference on Learning Representations*, 2020.

Lord Sen and Shyamapada Mukherjee. Mapping networks, 2026. URL https://arxiv.org/abs/2602.19134.

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

Petar Veličković, William Fedus, William L Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. Deep graph infomax. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rklz9iAcKQ.

Shaowen Wang, Linxi Yu, and Jian Li. LoRA-GA: Low-rank adaptation with gradient approximation. In *Advances in Neural Information Processing Systems*, 2024.

Yiwei Wang, Wei Wang, Yuxuan Liang, Yujun Cai, and Bryan Hooi. Data augmentation for graph neural networks. In *AAAI Conference on Artificial Intelligence*, 2021.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019.

Zhe-Rui Yang, Jindong Han, Chang-Dong Wang, and Hao Liu. GraphLoRA: Structure-aware contrastive low-rank adaptation for cross-graph transfer learning. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2025.

Zhilin Yang, William W Cohen, and Ruslan Salakhutdinov. Revisiting semi-supervised learning with graph embeddings. In *International Conference on Machine Learning*, 2016.

Jingqi Ye, Haonan He, Minglei Li, Fujun Han, Tao Chen, and Peng Ye. Gradient intrinsic dimensionality alignment: Narrowing the gap between low-rank adaptation and full fine-tuning. In *International Conference on Learning Representations*, 2026.

Fan Zhou, Chengtai Cao, Kunpeng Zhang, Goce Trajcevski, Ting Zhong, and Ji Geng. Meta-GNN: On few-shot node classification in graph meta-learning. In *CIKM*, 2019.

SUPPLEMENTARY MATERIAL

Appendix A specifies the maps and resource counts; Appendices B and E give statements and proofs. Primary evidence appears in Appendix C.5 (Q1), Appendix C.9 (Q2), and Appendix C.10 (Q3). Appendices C.7 and F.3 extend the comparisons (Q4). The weight–response diagnostic is in Appendix F.1. Appendix F.4 evaluates a separate Q3 cohort with macro-F1 and balanced accuracy. Historical baselines (Appendix C.13) and selection audits (Appendices C.15 and C.16) use separate protocols; Appendix D summarizes their scope.

A DETAILED MAPPING CONSTRUCTIONS

MapPerturb first obtains an anchor through supervised or self-supervised pretraining (Phase 1), then freezes it and optimizes $\mathbf{z} \in \mathbb{R}^d$ through a fixed affine weight map (Phase 2). The map controls P_{w} selected weights with $d \ll P_{\mathrm{w}}$ coordinates; the GNN still executes at full width. This appendix details the cyclic and gradient-derived maps, their storage, and the classifier variants.

A.1 PROTOCOLS AND COUNTING CONVENTIONS

We write $\mathcal{D}_{\mathrm{pre}}$ for anchor data, $\mathcal{D}_{\mathrm{adapt}}(k)$ for the k-label-per-class adaptation subset, and $\mathcal{D}_{\mathrm{test}}$ for held-out evaluation labels. We study two distinct settings. In the *supervised-anchor diagnostic*, Phase 1 uses the public training labels of the target benchmark and Phase 2 uses a subset of those labels. This setting tests local navigability after a task-aligned solution has already been found; it is not a label-free few-shot transfer protocol. In the *self-supervised transfer* setting, Phase 1 optimizes DGI without a task loss, and adaptation labels come from the official training mask. Primary Q3 uses a fixed training budget followed by test evaluation (Appendix C.10); Q4 classification selects the learning rate and checkpoint on validation data before test evaluation (Appendix F.3).

We report three resource quantities separately. P_{w} is the number of modulated entries in the frozen weight matrices; P_{train} counts all trainable values, including heads and biases; and adapter payload includes both trainable values and any extra frozen adapter basis. Optimizer state, memory, and wall-clock time are reported independently because a small P_{train} does not imply a faster implementation. When an implementation does not reliably expose an extra frozen-basis field, we treat that quantity as unavailable rather than converting a missing measurement into zero.

For the mapped backbone, we define the optimization-dimension reduction factor as

$$\rho_{\mathrm{dim}} = \frac{P_{\mathrm{w}}}{d}. \tag{15}$$

This measures how many independently trainable weight coordinates are replaced by one trainable coordinate. It does not count a downstream head, nor does it claim compression of the frozen inference weights. With Adam, the first- and second-moment state associated with the mapped backbone likewise scales with d rather than P_{w}.

A.2 IMPLEMENTATION-NATIVE WEIGHT CONVENTION

PyTorch stores a linear weight as $\mathbf{W}^{(l)} \in \mathbb{R}^{d_l \times d_{l-1}}$, with output dimension first. We use this convention throughout. A GCN layer is therefore written as

$$\mathbf{H}^{(l)} = \sigma\Big(\hat{\mathbf{A}}\mathbf{H}^{(l-1)}(\mathbf{W}^{(l)})^\top\Big), \tag{16}$$

where $\hat{\mathbf{A}} = \tilde{\mathbf{D}}^{-1/2}(\mathbf{A}+\mathbf{I})\tilde{\mathbf{D}}^{-1/2}$. Thus $P_{\mathrm{w}} = \sum_l d_l d_{l-1}$. A two-layer, width-64 GCN on Cora has $P_{\mathrm{w}} = 92{,}160$ modulated entries and 92,231 total trainable parameters when trained conventionally.

For layer l, define the cyclic lift $\mathbf{v}_l : \mathbb{R}^d \to \mathbb{R}^{d_{l-1}}$ by

$$[\mathbf{v}_l(\mathbf{z})]_i = z_{i \bmod d}. \tag{17}$$

The input-shared cyclic operator is

$$\mathbf{W}_{\mathrm{eff}}^{(l)} = \mathbf{W}^{(l),*} + \beta\, \mathbf{1}_{d_l}\mathbf{v}_l(\mathbf{z})^\top. \tag{18}$$

It repeats the entries of $\mathbf{z}$ over input coordinates and shares the resulting update across output rows. We therefore call it the *input-shared cyclic mapping*, making the tensor orientation explicit. In the equivalent textbook convention $\widetilde{\mathbf{W}}^{(l)} = (\mathbf{W}^{(l)})^\top \in \mathbb{R}^{d_{l-1} \times d_l}$, the update is $\beta \mathbf{v}_l(\mathbf{z}) \mathbf{1}_{d_l}^\top$.

The same $\mathbf{z}$ is used in every selected layer, β is fixed, and all anchor parameters remain frozen. The implementation uses a stateless functional forward pass, so gradients flow to $\mathbf{z}$ without assigning tensors through ".data". Equation (18) is affine in $\mathbf{z}$ and can be vectorized as

$$\theta(\mathbf{z}) = g(\mathbf{z}) = \theta^* + \beta E \mathbf{z}, \tag{19}$$

for a fixed sparse expansion matrix E. Protocols with bounded coordinates additionally constrain $\mathbf{z} \in B_d(R)$; the main label-free transfer protocol does not. The trainable state is d scalar values, but the realized parameter update is not measured by $\|\mathbf{z}\|_2$ alone:

$$\|\Delta\theta\|_2 = \beta \|E\mathbf{z}\|_2 \le \beta \|E\|_2 \|\mathbf{z}\|_2. \tag{20}$$

For the cyclic map, if coordinate j is copied m_j times, then $\|E\mathbf{z}\|_2^2 = \sum_j m_j z_j^2$ and $\|E\|_2 = \sqrt{\max_j m_j}$. Equal coordinate dimension therefore does not imply equal update radius, equal stored payload, or equal materialization cost. We use this distinction when interpreting matched controls.

A.3 GNN Operator View of the Cyclic Mapping

Let the pre-activation of layer l be $\mathbf{S}_l = \hat{\mathbf{A}} \mathbf{H}^{(l-1)} (\mathbf{W}_{\text{eff}}^{(l)})^\top$. Substituting Eq. (18) gives

$$\Delta_{W_l} \mathbf{S}_l = \beta \big(\hat{\mathbf{A}} \mathbf{H}^{(l-1)} \mathbf{v}_l(\mathbf{z}) \big) \mathbf{1}_{d_l}^\top. \tag{21}$$

This is the direct change due to W_l at fixed incoming representation $\mathbf{H}^{(l-1)}$. If earlier layers also change, their contribution must be included in the total pre-activation difference. The direct term computes one graph-filtered scalar per node and broadcasts it across all output channels of that layer. In encoder layers the scalar can still change later representations after nonlinearities and subsequent message passing, but the immediate update is not channel-selective. At the final classifier head, the same structure becomes a common-logit shift, which motivates the softmax-invariance result and the class-aware head below. The output-head result is exact; the effect of cyclic encoder updates also depends on subsequent nonlinearities and message passing.

A.4 Eight-Dimensional Maps at a Common Weight-Update Radius

To isolate map orientation from coordinate dimension and realized weight radius, the functional-retention experiment represents every family by an orthonormal basis

$$\delta\theta = Q\mathbf{z}, \qquad Q^\top Q = I_d, \qquad \|\mathbf{z}\|_2 \le r_{\text{w}}. \tag{22}$$

The cyclic basis spans the implemented input-shared lift after column normalization; random bases are Gaussian matrices reduced by thin QR. The gradient basis contains the top eight right singular directions of the matrix of normalized encoder gradients from 16 class-balanced discovery microtasks. Displacement and trajectory bases analogously use final discovery-network displacements and anchor-centered discovery-network checkpoints. A source rank below eight invalidates the fold rather than being padded with random directions.

We separate direction discovery from response fitting by cross-fitting. Within each class, the 20 official Planetoid training nodes are split into two folds of 10. The common full linear head, basis, and radius use only D_{basis}; the head is then frozen for all maps. A separately fitted reference network, tangent coefficients, and mapped task optimizer use only D_{fit}. Official validation labels are used only to report endpoints, and the official test mask is not accessed. Swapping the folds and averaging within each anchor prevents discovery examples or random-basis replicates from becoming independent statistical units.

The common radius is $r_{\text{w}} = \min\{\|\Delta_{\text{probe}}\|_2, 0.05\|\theta_{\text{w}}^*\|_2\}$, determined from D_{basis} alone. Because Q is orthonormal, $\|Q\mathbf{z}\|_2 = \|\mathbf{z}\|_2$, so Eq. (22) matches the maximum weight-update norm exactly. This changes the coordinate scale of the raw cyclic operator but preserves its reachable span; the experiment is therefore an equal-radius span comparison, not a replay of the original fixed β setting.

A.5 COMPACT GRADIENT-ATOM MAP

The dense gradient basis establishes an orientation mechanism but stores $P_{\mathrm{w}}d$ values. Our development protocol compresses the discovery objects before choosing their top-d coordinates, avoiding dependence on an arbitrary rotation within the dense top-d span. Let $\{g_t\}_{t=1}^{M}$ be the $M = 16$ normalized D_{basis} microtask gradients. For a selected layer l of shape $o_l \times i_l$, reshape the corresponding block of g_t as $G_{t,l}$ and take a rank-r truncated SVD,

$$G_{t,l} \approx L_{t,l} R_{t,l}^{\top}, \qquad L_{t,l} \in \mathbb{R}^{o_l \times r}, \quad R_{t,l} \in \mathbb{R}^{i_l \times r}. \tag{23}$$

The retained singular values are split symmetrically between the two factors. Each reconstructed full atom is renormalized, with the normalization absorbed into its factors. Stack these atoms as $A_r = [a_1^{(r)}, \ldots, a_M^{(r)}] \in \mathbb{R}^{P_{\mathrm{w}} \times M}$. For a top-$d$ eigensystem with strictly positive eigenvalues, $A_r^{\top} A_r V_d = V_d \Lambda_d$ and $V_d^{\top} V_d = I_d$, define

$$K_r = V_d \Lambda_d^{-1/2}, \qquad Q_r = A_r K_r, \qquad Q_r^{\top} Q_r = I_d. \tag{24}$$

The executed map stores the factors and K_r, not dense Q_r. With $\alpha = K_r \mathbf{z}$, it forms

$$\Delta W_l(\mathbf{z}) = \sum_{t=1}^{M} \alpha_t L_{t,l} R_{t,l}^{\top}. \tag{25}$$

Consequently, adaptation still optimizes exactly $d = 8$ values and $\|Q_r \mathbf{z}\|_2 = \|\mathbf{z}\|_2$. Here r is the layer-matrix factor rank of each discovery atom; it is neither the coordinate dimension, a functional dimension, nor an intrinsic GNN dimension. Equation (25) is a sum of M rank-r terms, so the final layer update is not itself restricted to rank r.

The conservative stored payload is

$$N_{\mathrm{compact}}(r) = Mr \sum_{l} (o_l + i_l) + Md, \tag{26}$$

including every factor and the $M \times d$ mixing matrix. The matched structured-random control preserves each atom's retained layerwise singular spectrum and all factor shapes, while independently randomizing the left and right orthonormal frames. It therefore matches d, factor slots, stored payload, optimizer, and radius, but it is not label-free because its retained spectra come from D_{basis}. Construction still computes MP_{w} full-gradient values and the experiment transiently materializes Q_r for diagnostics; compact stored payload does not imply lower discovery cost, runtime, or peak memory.

The Euclidean Gram mixing in Eq. (24) retains high-energy weight directions, but it need not prioritize directions that change graph predictions. We therefore also study a functional mixing on the same atom factors. Let A denote a dense or factorized atom matrix and write $A^{\top} A = U \Lambda U^{\top}$ on its positive eigenspace. The complete weight-orthonormal atom span is

$$W = AU\Lambda^{-1/2}, \qquad W^{\top} W = I. \tag{27}$$

On discovery nodes D, form the normalized centered-logit tangent

$$F_D = \frac{1}{\sqrt{|D|c}} C_D J_{G,D} W, \qquad H_D = F_D^{\top} F_D. \tag{28}$$

If S_d contains the top d eigenvectors of H_D, the functional map is

$$Q_F = AK_F, \qquad K_F = U\Lambda^{-1/2} S_d, \qquad Q_F^{\top} Q_F = I_d. \tag{29}$$

Here c is the number of classes and C_D centers each node's logits. For $p = \mathrm{rank}(A) \geq d$ and $S \in \mathbb{R}^{p \times d}$ with $S^{\top} S = I_d$, total sensitivity is $\|F_D S\|_F^2 = \mathrm{tr}(S^{\top} H_D S)$. Rayleigh–Ritz maximizes this sum with the top d eigenvectors (Appendix E.2). It uses the unweighted centered-logit Jacobian, not a Fisher- or GGN-weighted metric. The SVD factors are unchanged, and the factor payload remains Eq. (26); only the $M \times d$ mixing matrix changes. The spectrum- and payload-matched random factors receive the same functional mixing, so their comparison with gradient factors isolates the retained factor-frame orientation rather than the functional eigensolver.

To select r without exposing the downstream endpoint, every outer discovery fold is split class-wise into two five-node halves. In both inner directions, one half independently fits the head, radius, atoms, and Eq. (29), while the other supplies only a centered-logit Jacobian score. For dense tangent T_0 and compact tangent T_r on the scoring half, the primary fidelity is

$$\mathcal{F}_{\mathrm{proc}}(T_0, T_r) = 1 - \frac{\min_{O^\top O = I_d} \|T_0 - T_r O\|_F^2}{\|T_0\|_F^2}. \tag{30}$$

Here $T_0, T_r \in \mathbb{R}^{m\times d}$, $O \in \mathbb{R}^{d\times d}$, and $\|T_0\|_F > 0$; a zero reference tangent cannot define this score. Unlike projection capture, this coordinate-invariant statistic penalizes a candidate with the correct column space but vanishing functional sensitivity. The two inner directions are averaged before rank selection.

Transfer normalization and optimizer. Algorithm 1 uses a different normalization from supervised reconstruction: each rank-truncated layer block is divided by its Frobenius norm, with a 10^{-12} floor. Its Gram eigenvalues are floored at 10^{-6}. Consequently, $\widetilde{A}_r K$ is exactly orthonormal only if none of the retained eigenvalues is floored; layer gains further change its norm. The transfer protocol uses $E = D_s \widetilde{A}_r K$ as executed. Functional comparisons orthonormalize its realized span before imposing the common radius. Discovery uses independent feature permutations for DGI or entrywise Bernoulli masks with rate 0.3 for masked reconstruction; auxiliary parameters remain fixed while collecting encoder gradients.

For $w_q = (z_q, \phi_q)$ and cross-entropy gradient h_q, the Adam step is

$$m_{q+1} = b_1 m_q + (1-b_1) h_q, \qquad v_{q+1} = b_2 v_q + (1-b_2) h_q^{\odot 2}, \tag{31}$$

$$w_{q+1} = w_q - \eta \, \frac{m_{q+1}/(1-b_1^{q+1})}{\sqrt{v_{q+1}/(1-b_2^{q+1})} + \epsilon_{\mathrm{A}}}. \tag{32}$$

Operations are coordinatewise, with $b_1 = 0.9$, $b_2 = 0.999$, $\epsilon_{\mathrm{A}} = 10^{-8}$, $\eta = 0.01$, zero initial moments, and zero weight decay. The 100-step transfer run jointly trains the eight encoder coordinates and the full classifier without projection.

A.6 Representational Boundary

The input-shared mapping is deliberately restrictive. At a C-class linear softmax head, Eq. (18) adds the same scalar to all class logits and therefore cannot change class probabilities directly (Proposition 1). Encoder updates can still change predictions through the hidden representations. To test whether a class-specific head makes these directions more useful for adaptation, we evaluate the augmentation below.

We pair the encoder map with a rank-budget-r factorized classifier

$$\mathbf{W}_{\mathrm{head}} = \mathbf{B}\mathbf{A}, \qquad \mathbf{A} \in \mathbb{R}^{r\times d_h}, \quad \mathbf{B} \in \mathbb{R}^{C\times r}, \tag{33}$$

and a trainable bias. The head has $r(d_h + C) + C$ trainable values and realized matrix rank at most $\min\{C, r, d_h\}$. We therefore use "rank budget" rather than claiming exact rank, which matters for three-class PubMed when $r = 4$. The confirmatory variant fixes the budget at $r = 4$, $d = 8$, $\beta = 0.01$, and projects $\mathbf{z}$ onto $\|\mathbf{z}\|_2 \le 2$ after each Adam step. Its paired control uses exactly the same factorized-head initialization and adaptation random stream but omits the eight-dimensional encoder map. Consequently, the pair differs by exactly eight trainable values and isolates the incremental contribution of MapPerturb.

A.7 Design Scope

The cyclic lift provides a simple fixed baseline. Full-matrix cyclic, block-constant, and fixed-random lifts perform similarly in the supervised structure ablation, and the coordinate-sharing ablation evaluates sharing within this benchmark family. A fixed affine lift permits explicit analysis of its Jacobian, parameter count, storage, and gradient path. Fixed-MLP mapping probes produced very small gradients with respect to $\mathbf{z}$ in our implementation; their behavior does not characterize learned mapping networks in general.

More broadly, the construction applies to any selected two-dimensional linear weight. For GAT, we adapt the projection matrices and keep attention vectors frozen; for GIN, we adapt linear layers inside the MLP. Architecture-specific success is an empirical question, and no equivariance or expressivity advantage follows from the lift alone.

A.8 Complexity

The input-shared cyclic map trains d backbone values instead of P_{w}, giving the exact reduction factor in Eq. (15); Adam allocates its mapped-backbone moment tensors for those d values. The rank-budget-4 hybrid trains $r(d_h + C) + C + d$ values, separating the low-dimensional encoder map from the class-aware output budget. Both variants retain the frozen backbone, and the fixed cyclic lift requires no stored dense basis. The aligned diagnostic maps still optimize only d coordinates, but their current dense bases require $P_{\mathrm{w}}d$ frozen values. Random bases can be regenerated from a seed; gradient, displacement, and trajectory bases are orientation diagnostics rather than equal-payload deployment methods. By contrast, the compact gradient-atom construction in Eqs. (23)–(26) reduces the stored basis description while preserving d trainable coordinates. The functional-fidelity criterion selects rank two for confirmation on new anchors (Section C.9); the separate Euclidean-overlap criterion selects none of the tested ranks. The transfer implementation stores these compact atom factors as task-specific frozen payload. Materializing the adapted weights requires $O(P_{\mathrm{w}})$ elementwise work, and activation memory depends on the implementation. Runtime and peak memory therefore require separate measurements.

B Complete Theoretical Statements

We analyze what is gained and what must be preserved when the optimization variables are reduced from P_{w} selected weights to d trainable coordinates. This map gives an exact reduction in trainable dimension by construction; performance additionally depends on its reachable weight directions and their graph-conditioned functional image. The results concern a fixed anchor; complete proofs appear in Appendix E.

B.1 Affine Weight Mapping and Exact Dimensional Reduction

Using the implementation-native convention from Eq. (16), MapPerturb applies

$$\mathbf{W}_{\mathrm{eff}}^{(l)} = \mathbf{W}^{(l),*} + \beta \mathbf{1}_{d_l} \mathbf{v}_l(\mathbf{z})^\top, \qquad g(\mathbf{z}) = \theta^* + \beta E \mathbf{z}. \tag{34}$$

Here $E \in \mathbb{R}^{P_{\mathrm{w}} \times d}$ is fixed. If coordinate j is copied m_j times over all selected entries, then the columns of E have disjoint support and

$$\mathrm{Lip}(g) = \beta \|E\|_2 = \beta \sqrt{\max_j m_j}. \tag{35}$$

When E has column rank d, the reachable mapped weights form a d-dimensional affine set even though they control P_{w} entries. Thus the number of independently optimized backbone variables, their gradients, and their first-order optimizer states is reduced from P_{w} to d. This is the basic optimization-state reduction statement. It is distinct from compressing the stored anchor and from proving that the reachable affine set is the optimal one for a task.

Proposition 1 (Common-logit invariance)**.** *Let a C-class linear classifier have logits $\mathbf{a}(\mathbf{h}) = \mathbf{W}\mathbf{h} + \mathbf{b}$, where $\mathbf{W} \in \mathbb{R}^{C \times d_h}$. For any $\mathbf{v} \in \mathbb{R}^{d_h}$, define $\mathbf{W}' = \mathbf{W} + \mathbf{1}_C \mathbf{v}^\top$. Then for every $\mathbf{h}$,*

$$\mathrm{softmax}(\mathbf{W}'\mathbf{h} + \mathbf{b}) = \mathrm{softmax}(\mathbf{W}\mathbf{h} + \mathbf{b}). \tag{36}$$

Consequently, the input-shared perturbation of the final classifier weight cannot by itself change cross-entropy probabilities or predictions.

Proof. $\mathbf{W}'\mathbf{h} + \mathbf{b} = \mathbf{W}\mathbf{h} + \mathbf{b} + (\mathbf{v}^\top \mathbf{h})\mathbf{1}_C$, and softmax is invariant to adding a common scalar to all logits. □

A shared $\mathbf{z}$ can still receive nonzero gradients through adapted encoder weights; the invariance concerns only the direct classifier-weight update.

B.2 Functional Dimension and Mapping Alignment

Reducing the trainable dimension is useful only if the mapped coordinates retain enough functional degrees of freedom. Let $G(\theta)$ collect the logits on the adaptation nodes or examples, and let C subtract the per-example mean logit so that common-logit shifts are removed. With $J_G = \nabla_\theta G(\theta^*)$ at a differentiable anchor, the first-order centered-logit variation is

$$C\big(G(\theta^* + \beta E\mathbf{z}) - G(\theta^*)\big) = \beta\, T_G\mathbf{z} + o(\|\mathbf{z}\|_2), \qquad T_G = CJ_GE. \tag{37}$$

The column space of T_G determines which first-order responses the map can produce. A small map can support useful adaptation when these responses align with the task, while a larger map can miss the relevant directions.

Here $\mathrm{col}(T_G)$ is a first-order object at the specified anchor and node set. A positive radius changes the map's bounded response ellipsoid and the fidelity of the Taylor approximation, not this column space itself. When comparing spectra across maps, we use a weight-orthonormal basis and one common weight radius; a mere rescaling of coordinates can otherwise change singular values without changing the span.

We call

$$d_{\mathrm{func}}(G, E) = \mathrm{rank}(T_G) \le \mathrm{rank}(E) \le d \tag{38}$$

the local functional dimension of the mapping on graph G. It separates the nominal optimization budget d from the number of first-order, class-relevant output directions that the budget actually realizes. A low value can indicate redundant mapped coordinates, an invariant direction such as the common-logit shift, or graph-conditioned collapse such as oversmoothing. It is therefore not automatically evidence of beneficial compression.

The same point appears in a local quadratic approximation. Let

$$q(\delta) = \mathcal{L}(\theta^*) + g^\top\delta + \tfrac{1}{2}\delta^\top H\delta, \tag{39}$$

where $g = \nabla_\theta\mathcal{L}(\theta^*)$ and $H \succ 0$ is a symmetric positive-definite curvature model. An actual Hessian need not satisfy this assumption. Write $\bar{E} = \beta E$. If $\delta^\star$ minimizes q over the ambient space and $\bar{E}z^\star$ minimizes it over the mapped subspace, then

$$q(\bar{E}z^\star) - q(\delta^\star) = \frac{1}{2}\left\|(I - P_{H^{1/2}\bar{E}})H^{-1/2}g\right\|_2^2, \tag{40}$$

where $P_{H^{1/2}\bar{E}}$ is the Euclidean projector onto $\mathrm{col}(H^{1/2}\bar{E})$. Appendix E.1 derives the identity and states the additional range condition for a positive-semidefinite H. The gap is zero only when the preconditioned descent direction $H^{-1/2}g$ is covered after Hessian scaling. The identity separates the coordinate budget from alignment with a specified descent direction. Our reconstruction experiments examine this distinction through the graph-conditioned tangent CJ_GE, using different nodes for map construction, coefficient fitting, and evaluation (Section C.4). The quadratic identity assumes a positive-definite curvature model; the empirical reconstruction score instead measures a finite nonlinear response.

Weight-space overlap and functional retention can also differ after factorization. In the Cora development study, the strongest tested compact rank has mean squared-cosine overlap 0.762 with the dense discovery span, below the prespecified 0.80 threshold, while its held-out functional R_0^2 is 0.677, compared with 0.691 for the dense reference. This motivates the discovery-tangent fidelity criterion in Eq. (11). The subsequent confirmation supports the gradient factors' advantage over matched random directions; the smaller functional-versus-weight mixing contrast remains unresolved (Section C.9). Maximizing sensitivity within an atom span and aligning that span with a particular target are distinct objectives, as the next calculation shows.

B.3 Target-specific approximation and nested spaces

The following projection calculation quantifies how well a fixed functional image approximates a specified target.

Proposition 2 (Target-specific least-squares approximation)**.** *Let $T \in \mathbb{R}^{m\times d}$ be a fixed centered-logit tangent and $y \in \mathbb{R}^m$ a target response. Let $P_T = TT^\dagger$ be the orthogonal projector onto its column space, with $T^\dagger$ the Moore–Penrose inverse. Then*

$$\min_{z\in\mathbb{R}^d} \|y - Tz\|_2^2 = \|(I - P_T)y\|_2^2. \tag{41}$$

The minimum under $\|z\|_2 \le r$ is at least this value, with equality exactly when $\|T^\dagger y\|_2 \le r$.

Proof. For every z, $y - Tz = (I - P_T)y + (P_T y - Tz)$ is an orthogonal decomposition. The second term vanishes at $z = T^\dagger y$, proving the first statement. All least-squares solutions are $T^\dagger y + v$ with $v \in \ker T$; the two summands are orthogonal, so $T^\dagger y$ has the smallest coefficient norm. This proves the radius condition as well. □

For a direct counterexample to sensitivity as target coverage, take an orthonormal basis e_1, e_2 in a two-dimensional centered-logit space, $y = e_2$, $T_1 = ae_1$, and $T_2 = e_2$. For $a > 1$, the first map has larger Frobenius norm and the same rank, but its minimum squared residual is one; the second attains zero with radius at least one. Replacing the unweighted norm by a positive semidefinite metric does not remove the need for target alignment. For example, at uniform class probabilities the categorical Fisher metric is a scalar multiple of the identity on the centered-logit space, and the same counterexample applies to a Fisher-weighted trace. Neither a large trace nor a large leading singular value alone ensures downstream classification accuracy.

Now fix a weight radius r and let $\mathcal{S}_1 \subseteq \mathcal{S}_2$ be nested weight spans. Their feasible updates satisfy $\mathcal{S}_1 \cap B(0, r) \subseteq \mathcal{S}_2 \cap B(0, r)$. The minimum of any fixed reconstruction loss over the larger set cannot increase; this set-inclusion argument holds even for the exact nonlinear network. For orthonormal bases, the coordinate ball represents precisely this weight ball. Ridge fitting followed by radial rescaling need not attain this constrained minimum. Even direct constrained fitting of a tangent need not yield monotone nonlinear reconstruction on different validation nodes. Q1 measures this achieved held-out response, and its cyclic decline under the original estimator reverses in the constrained rerun (Appendix C.6). Random draws at different dimensions in Q1 are not nested in the first place.

B.4 Conditional Local Covering

Theorem 1 (Tube-based local covering number)**.** *Let θ^* be an anchor and let*

$$\mathcal{M}_{\text{local}} = \{\theta \in B(\theta^*, r) : \mathcal{L}(\theta) \le \mathcal{L}(\theta^*) + \varepsilon_0\}. \tag{42}$$

Assume that there are an orthonormal matrix $\mathbf{U} \in \mathbb{R}^{P \times q}$ and $\tau \ge 0$ such that every $\theta \in \mathcal{M}_{\text{local}}$ admits an $\mathbf{a} \in \mathbb{R}^q$ with $\|\mathbf{a}\|_2 \le r$ and

$$\|\theta - \theta^* - \mathbf{U}\mathbf{a}\|_2 \le \tau. \tag{43}$$

Let $\mathcal{N}_{\text{ext}}$ denote the external covering number of the induced function set, so cover centers may lie in the ambient function class. If the model-output map is L_f-Lipschitz ($L_f > 0$) in parameters under the chosen functional metric, then for every $\varepsilon > L_f\tau$,

$$\mathcal{N}_{\text{ext,func}}(\mathcal{M}_{\text{local}}, \varepsilon) \le \left(1 + \frac{2rL_f}{\varepsilon - L_f\tau}\right)^q, \tag{44}$$

The result is a direct consequence of Eq. (43); it does not derive that condition. In particular, counting Hessian eigenvalues above a threshold neither implies tube containment nor identifies the directions of low-loss displacement. The trajectory PCA/SVD experiment measures concentration of observed optimization paths, which is weaker than measuring the geometry of the entire sublevel set.

B.5 Cold-Start Random-Offset Bound

Proposition 3 (Random-offset chart intersection bound)**.** *Let $1 \le d < P$, $R > 0$, and $\Omega \subseteq B_P(0, R)$ be measurable and satisfy*

$$\text{Vol}(\Omega) \ge c_\Omega \,\text{Vol}(B_P(0, R)), \qquad c_\Omega > 0. \tag{45}$$

Conditionally on any $\mathbf{U}$ with $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_d$, let the chart be $\mathbf{a} + \mathbf{U}B_d(s)$ with $s \le R$, where the random offset $\mathbf{a}$ has a density bounded by $c_0/\,\text{Vol}(\Omega)$ and supported on Ω. If one target basin is contained in a radius-r_b ball, then the probability over $\mathbf{a}$ satisfies

$$\Pr[(\mathbf{a} + \mathbf{U}B_d(s)) \cap \mathcal{B} \ne \varnothing] \le C'_{P,d} \left(\frac{s + r_b}{R}\right)^d \left(\frac{r_b}{R}\right)^{P-d}, \tag{46}$$

where $C'_{P,d}$ contains the density, volume-ratio, and c_Ω constants. For N_{basins} such basins, a union bound applies. If $r_b/R \le e^{-c}$ for fixed $c > 0$, the bound decays exponentially in $P - d$ along a sequence of problems whenever, for a fixed $\eta \in (0, 1)$,

$$\log N_{\text{basins}} + \log C'_{P,d} + d \log 2 \le (1 - \eta)c(P - d). \tag{47}$$

The proposition describes a sufficient geometric mechanism for cold-start failure under its random-offset assumptions. The implemented chart does not follow this offset model, so the bound does not quantify its failure probability.

B.6 Capacity of the Class Parameterized by Reduced Coordinates

Theorem 2 (Conditional Rademacher bound)**.** *Fix the anchor, lift, and radius independently of an i.i.d. evaluation sample $S = \{(x_i, y_i)\}_{i=1}^m$ (or condition on independent discovery data). Define*

$$\mathcal{F}_R = \left\{(x, y) \mapsto \ell(f_{g(\mathbf{z})}(x), y) : \|\mathbf{z}\|_2 \le R\right\}. \tag{48}$$

Assume the loss is bounded in $[0, B]$ and, under the normalized empirical ℓ_2 metric, the sample loss vector is L_S-Lipschitz in $\mathbf{z}$. Then

$$\log \mathcal{N}(\mathcal{F}_R|_S, \varepsilon) \le d \log\left(1 + \frac{2RL_S}{\varepsilon}\right) \tag{49}$$

and

$$\widehat{\mathfrak{R}}_S(\mathcal{F}_R) \le C \frac{RL_S\sqrt{d}}{\sqrt{m}} \tag{50}$$

for a universal constant C. Hence, with probability at least $1 - \delta$, uniformly over $\mathcal{F}_R$,

$$L(\mathbf{z}) \le \widehat{L}_S(\mathbf{z}) + 2C \frac{RL_S\sqrt{d}}{\sqrt{m}} + 3B\sqrt{\frac{\log(2/\delta)}{2m}}. \tag{51}$$

If the loss is ρ-Lipschitz in model output and the model-output map is L_f-Lipschitz in parameters under compatible metrics, one may take $L_S \le \rho L_f \beta \|E\|_2$.

An analogous calculation for a P-dimensional parameter ball replaces d by P, but the two upper bounds are not comparable from dimension alone: their radii and Lipschitz constants generally differ. For the cyclic lift, every selected weight entry copies exactly one coordinate, so $\sum_j m_j = P_{\text{w}}$ and $\|E\|_2 = \sqrt{\max_j m_j}$. In a balanced cyclic copy $\sqrt{d}\,\|E\|_2 \approx \sqrt{P_{\text{w}}}$, so the nominal coordinate dimension alone does not yield a complexity improvement. Planetoid node classification is transductive; the i.i.d. theorem therefore does not directly apply to those experiments.

B.7 Conditional Phase-2 Convergence

Theorem 3 (Linear convergence under a coordinate-space PL condition)**.** *Let $\widetilde{\mathcal{L}}(\mathbf{z}) = \mathcal{L}(\theta^* + \beta E\mathbf{z})$ and $\widetilde{\mathcal{L}}^* = \inf_{\mathbf{z}} \widetilde{\mathcal{L}}(\mathbf{z}) > -\infty$. Suppose the iterates remain in a region containing their update segments, where $\widetilde{\mathcal{L}}$ is L_z-smooth and, for $0 < \mu_z \le L_z$, satisfies*

$$\tfrac{1}{2}\|\nabla \widetilde{\mathcal{L}}(\mathbf{z})\|_2^2 \ge \mu_z\big(\widetilde{\mathcal{L}}(\mathbf{z}) - \widetilde{\mathcal{L}}^*\big). \tag{52}$$

Gradient descent with step size $1/L_z$ then satisfies

$$\widetilde{\mathcal{L}}(\mathbf{z}_t) - \widetilde{\mathcal{L}}^* \le \left(1 - \frac{\mu_z}{L_z}\right)^t \big(\widetilde{\mathcal{L}}(\mathbf{z}_0) - \widetilde{\mathcal{L}}^*\big). \tag{53}$$

If a convex region $\mathcal{Z}$ in coordinate space contains the iterates and their gradient updates, and $\|\nabla^2_\theta \mathcal{L}(\theta)\|_2 \le L_\theta$ throughout $g(\mathcal{Z})$, then one valid smoothness constant on $\mathcal{Z}$ is

$$\bar{L}_z = \beta^2 \|E\|_2^2 L_\theta, \tag{54}$$

and the same conclusion holds with step size $1/\bar{L}_z$ when $\bar{L}_z > 0$ and $\mu_z \le \bar{L}_z$. The Hessian statement assumes twice continuous differentiability on this region.

The pullback identity $\nabla^2_{\mathbf{z}}\widetilde{\mathcal{L}} = \beta^2 E^\top(\nabla^2_\theta \mathcal{L})E$ is exact because g is affine. The measured β^2 scaling at the fixed anchor is therefore an implementation check of this identity, not evidence that the PL condition holds. The experiments use Adam, and the confirmatory hybrid additionally projects onto a coordinate ball, so the theorem should not be read as an optimizer-specific convergence guarantee.

Adam applied to the coordinates preserves the reachable weight span: for any coordinate update u_t, including an Adam update, $\theta_{t+1} - \theta_t = \beta E u_t \in \mathrm{col}(E)$. Adaptive coordinate scaling can change optimization trajectories without changing this invariant. If an ambient stochastic gradient has covariance Σ, its gradient with respect to $\mathbf{z}$ has covariance $\beta^2 E^\top \Sigma E$. Under orthonormal columns and the additional isotropy assumption $\Sigma = \sigma^2 I$, its trace is $\beta^2 d\sigma^2$; in general it depends on orientation and covariance, not dimension alone. The covariance identity describes gradients with respect to $\mathbf{z}$, not Adam's adaptively rescaled updates.

C Full Experimental Protocols and Results

The confirmatory Planetoid studies use PyTorch 1.13.0, PyG 2.5.3, and an NVIDIA RTX 3060 Laptop GPU with 6 GB memory. The 4090D stress tests use PyTorch 2.5.1 and an NVIDIA RTX 4090D. Timing and memory are therefore compared only within a run environment. Code, result files, and exact split indices are provided in the artifact.

C.1 Datasets, Models, and Counting Conventions

Datasets. Node experiments use the official Planetoid graphs: Cora has 2,708 nodes and 7 classes; CiteSeer has 3,327 nodes and 6 classes; PubMed has 19,717 nodes and 3 classes (Yang et al., 2016). Secondary graph-level diagnostics use PROTEINS, MUTAG, and NCI1 from TUDataset (Morris et al., 2020). The heterophily stress tests add Actor and the WebKB graphs Cornell, Texas, and Wisconsin using their 10 official PyG splits (Pei et al., 2020). Additional studies use ogbn-arxiv, PPI, and the official ogbg-molhiv scaffold split (Hu et al., 2020). Appendix F.3 specifies the fixed external classification protocol; Appendix C.19 distinguishes the earlier pilots.

Models. The main studies use a two-layer width-64 GCN. Additional settings use GAT (Veličković et al., 2018), GIN (Xu et al., 2019), GraphSAGE, and deeper GCNs. Primary Q3 uses the self-supervised and adaptation settings in Appendix C.10. Supervised-anchor diagnostics instead train on public training labels with Adam, learning rate 0.01, weight decay 5×10^{-4}, and up to 200 epochs with validation early stopping; their 100-step adaptation retains that weight decay.

Resources. We use the weight, trainable-value, and payload definitions in Appendix A.1. Adam state, peak incremental CUDA allocation, and wall-clock time are measured separately; unavailable resource fields are not assigned zero.

C.2 Protocol Separation and Statistical Analysis

Q1 and Q2 are common-radius reconstruction studies; primary Q3 is the objective-matched few-shot cohort in Appendix C.10. The earlier baseline cohort is reported separately in Appendix C.13. All use the full graph's features and edges, with label access determined by the experiment. Q1/Q2 use disjoint official training folds for discovery and fitting, report validation endpoints, and do not evaluate the test mask. Q3 uses support labels only during adaptation and evaluates the test endpoint after its fixed training budget. Selection studies explicitly distinguish validation-based choices from subsequent test evaluation.

Supervised-anchor diagnostics allow public training labels during pretraining and study local preservation of that trained solution. The separate class-aware-head and cold-start studies use their own head, projection, and selection rules.

For the class-aware-head confirmation (P3), the candidate rank budget, coordinate dimension, scale, radius, and optimizer were frozen before the 10-seed run. The run uses seeds 210000–210009 on all three datasets and all three shot levels. One DGI anchor is reused across the three shot levels, so the primary independent unit is a dataset–seed pair after averaging over k. We report paired means, 95% confidence intervals, exact or Monte-Carlo sign-flip tests as appropriate, Wilcoxon tests, effect sizes, and a ± 0.5-percentage-point region of practical equivalence (ROPE). This band was

not independently preregistered, so its interpretation is descriptive. The nine dataset–shot secondary contrasts use Holm correction. The P4 confirmation uses 10 new seeds 270000–270009.

C.3 Preserving task-supervised anchors

We compare eight-coordinate cyclic adaptation with its task-supervised frozen anchor over 90 paired configurations, averaging the three shot levels within each dataset–seed anchor. Table 4 shows a mean MP-8 minus Frozen increment of -0.07 points (95% CI $[-0.20, +0.06]$). At the individual configuration level, MP-8 wins 30 times, ties 23 times, and loses 37 times; 78/90 differences lie inside the ± 0.5-point descriptive band. Adaptation therefore largely preserves the supervised anchor's accuracy, with no resolved mean improvement.

Table 4: MP-8 versus its task-supervised frozen anchor. Clustered Δ averages the three shot levels within each of 10 anchor seeds. W/T/L counts wins, ties, and losses over 30 configurations per dataset. Band B/E/W counts better, within-band, and worse differences using a descriptive ± 0.5-point band, not a formal equivalence test.

Dataset	Clustered Δ [95% CI]	W/T/L	Band B/E/W
Cora	-0.27 [-0.57,+0.02]	6/4/20	0/22/8
CiteSeer	-0.03 [-0.27,+0.21]	11/5/14	2/26/2
PubMed	+0.09 [-0.01,+0.18]	13/14/3	0/30/0
Overall	-0.07 [-0.20,+0.06]	30/23/37	2/78/10

The trajectory analysis gives complementary geometric evidence. Across 10 supervised GCN seeds, late-stage displacements explaining 95% of observed weight-trajectory variance use 17 dimensions on Cora and 22 on CiteSeer, compared with $P_{\mathrm{w}} = 92{,}160$ and $237{,}376$. These are ranks of sampled trajectories, not Hessian ranks or dimensions of an entire sublevel set. Dimension saturation, depth scaling, and the exact operator orientation are reported in the appendix.

C.4 Response Reconstruction in the d-Dimensional Search Space

To measure which target responses the eight-coordinate maps retain, we evaluate reconstruction through $T_G = C J_G E$ under a protocol fixed before the main run. For each of 10 label-free DGI anchors per dataset, the 20 official training nodes in every class are split into two 10-node folds. In each cross-fit direction, D_{basis} alone fits the common linear head, calibrates the weight radius, and constructs the map; D_{fit} fits a separate encoder-only reference network from the same anchor and fits tangent-reconstruction coefficients; official validation nodes are used only for evaluation. The official test mask is never accessed. Random bases are averaged within fold, then the two folds are averaged within each dataset–anchor unit.

All families use a thin orthonormal basis $Q \in \mathbb{R}^{P_{\mathrm{w}} \times 8}$ and the same realized weight radius. The selected DGI encoder matrices contain $95{,}808$, $241{,}088$, and $36{,}096$ entries on Cora, CiteSeer, and PubMed, so the optimization-dimension reductions are respectively $11{,}976\times$, $30{,}136\times$, and $4{,}512\times$. These counts differ from the supervised Cora example because the DGI diagnostic maps two encoder matrices rather than an input matrix and a seven-class output matrix. We compare the implicit cyclic map, five Gaussian random orthonormal maps per fold, and a gradient basis formed from 16 class-balanced D_{basis} microtasks. Displacement and trajectory bases are secondary diagnostics reported in the appendix.

Recorded reconstruction estimator. Let y_{fit} be the reference network's centered-logit change on fitting nodes and let T_{fit} approximate $C_{\mathrm{fit}} J_{G,\mathrm{fit}} Q$. Each column uses the central difference

$$(T_S)_{:,j} = \frac{C_S f_{G,S}(\theta^* + \epsilon_{\mathrm{fd}} Q_{:,j}) - C_S f_{G,S}(\theta^* - \epsilon_{\mathrm{fd}} Q_{:,j})}{2\epsilon_{\mathrm{fd}}}, \qquad \epsilon_{\mathrm{fd}} = \max(10^{-3}, 0.01 r_{\mathrm{w}}). \tag{55}$$

The implementation fits ridge coefficients and then rescales them radially:

$$\begin{aligned} &a = (T_{\mathrm{fit}}^\top T_{\mathrm{fit}} + \lambda I_d)^{-1} T_{\mathrm{fit}}^\top y_{\mathrm{fit}}, \qquad z = a \min\{1, r_{\mathrm{w}}/\|a\|_2\}, \\ &\lambda = \max\{10^{-12}, 10^{-6}\,\mathrm{tr}(T_{\mathrm{fit}}^\top T_{\mathrm{fit}})/d\}. \end{aligned} \tag{56}$$

Table 5: Held-out response reconstruction with $d = 8$ DGI-GCN encoder coordinates. Each entry first averages five random bases when applicable and the two cross-fit folds, then averages 10 anchor seeds. Nonlinear R_0^2 reconstructs the separately fitted D_{fit} reference network's centered-logit change on the official validation nodes; gradient capture is measured from a gradient computed on D_{fit}. Sensitivity is $\|C_{\mathcal{V}} J_{G,\mathcal{V}} Q\|_F / \sqrt{|\mathcal{V}|c}$ for validation nodes $\mathcal{V}$ and c classes. Validation accuracy uses separate task-loss optimization, not reconstruction coefficients. The official test mask is never accessed.

Dataset	Basis	Nonlinear R_0^2	Grad. capture	Sensitivity	Part. rank	Val. acc. (%)
Cora	Random	0.085	0.00009	0.057	6.943	66.9
	Cyclic	0.418	0.00842	0.213	1.228	67.9
	Gradient	0.695	0.34363	1.654	4.911	70.0
CiteSeer	Random	0.092	0.00003	0.047	6.868	51.7
	Cyclic	0.106	0.00024	0.163	1.228	51.8
	Gradient	0.397	0.12488	2.389	4.443	52.2
PubMed	Random	0.057	0.00023	0.023	5.223	63.2
	Cyclic	0.138	0.00859	0.175	1.127	63.0
	Gradient	0.769	0.37223	0.727	3.299	64.8
Macro	Random	0.078	0.00012	0.042	6.345	60.6
	Cyclic	0.221	0.00575	0.184	1.195	60.9
	Gradient	0.621	0.28024	1.590	4.218	62.3

We set $z = 0$ when $a = 0$. The reported nonlinear response is evaluated at $\theta^* + Qz$ on validation nodes. This ridge-and-rescaling estimator is distinct from the exact radius-constrained least-squares optimum in Proposition 2. For an orthonormal basis, capture of a nonzero weight vector u is $\|Q^\top u\|_2^2 / \|u\|_2^2$; we use this for gradients on D_{fit} and displacements of the reference network. If σ_j are the validation tangent singular values, participation rank is $(\sum_j \sigma_j^2)^2 / \sum_j \sigma_j^4$, defined as zero at zero tangent energy. It summarizes spectral concentration and is not an algebraic rank.

The primary endpoint is nonlinear reconstruction of the separately fitted reference network's centered-logit change on validation nodes after tangent coefficients are fit on D_{fit}. The cyclic map retains nonzero functional signal on all three datasets: its R_0^2 is 0.418 on Cora, 0.106 on CiteSeer, and 0.138 on PubMed, with all dataset-level 95% intervals above zero. The dataset-equal mean is 0.221, compared with 0.078 for random maps. The paired cyclic-minus-random contrast is $+0.143$ (95% CI $[0.122, 0.164]$); this advantage is clear on Cora and PubMed, whereas the CiteSeer contrast includes zero.

At this fixed dimension, map orientation has a large effect. The gradient-aligned basis reaches dataset-equal nonlinear $R_0^2 = 0.621$. Its paired advantage is $+0.400$ over cyclic (95% CI $[0.380, 0.420]$) and $+0.543$ over random (95% CI $[0.514, 0.571]$); both Holm-adjusted paired t-test values are 1.3×10^{-11}. All three dataset-specific intervals are positive. Capture of the D_{fit} gradient rises from 0.006 for cyclic and approximately zero for random to 0.280 for the gradient basis, while functional sensitivity rises from 0.184 and 0.042 to 1.590.

Weight-space overlap gives a different ordering. The gradient basis captures 3.7% of the reference network's displacement in Euclidean weight energy, compared with 7.9% for the cyclic basis, despite its higher reconstruction score. Displacement and trajectory bases capture about 43% of reference-network weight energy without a commensurate advantage in functional R_0^2. These results motivate evaluating directions through the graph-conditioned image CJ_GE. The measurements concern local responses around the anchor and leave global weight-space geometry unresolved.

Under the same d, radius, and optimization budget, gradient alignment also improves dataset-equal validation accuracy by 1.45 percentage points over cyclic (95% CI $[0.06, 2.84]$) and 1.75 points over random ($[0.38, 3.13]$), with lower validation loss in both comparisons. Cora and PubMed support the gradient-versus-cyclic accuracy contrast; the CiteSeer interval includes zero. These gains require storing $P_{\mathrm{w}} d$ frozen basis values, whereas cyclic and random bases can be regenerated from structure or a seed. The comparison controls optimization dimension and weight radius while allowing these storage requirements to differ.

C.5 Q1: DIMENSION AND ORIENTATION

The crossed study fixes the anchor, head, gradient bank, separately fitted reference network, validation nodes, and optimization budget within each dataset–seed–fold while varying $d \in \{2, 4, 8, 16\}$. Seeds for the five random bases are independent across dimensions; rank-deficient banks invalidate a cell rather than trigger random completion. All cells were valid and the official test mask remained sealed.

Gradient bases are nested leading singular directions of the shared bank. The input-shared cyclic bases are nested for the tested doubling dimensions: a period-d pattern is reproduced in period $2d$ by repeating its coordinate vector. Orthonormalization preserves these spans and the common weight-ball feasible sets. Random seeds include d, so random bases are drawn independently for each dimension. The random-family interaction must not be interpreted as adding coordinates to a single fixed random basis.

Table 6: Dimension and orientation under ridge fitting with radial rescaling. Entries are dataset-equal R_0^2 after fold and random-replicate averaging within each of ten anchor seeds.

Basis	$d = 2$	$d = 4$	$d = 8$	$d = 16$
Random	.0297	.0526	.0768	.0995
Cyclic	.2398	.2309	.2056	.1618
Gradient	.3925	.5270	.6165	.6003

The eight predeclared orientation contrasts all survive one Holm family. Gradient minus cyclic is .153 [.077, .229], .296 [.247, .345], .411 [.375, .447], and .438 [.383, .494] at increasing d; gradient minus random is .363 [.297, .428], .474 [.434, .515], .540 [.513, .566], and .501 [.456, .546]. Each contrast wins on all ten seed-level dataset-equal pairs. The corresponding largest Holm-adjusted p value is .00139.

The predeclared $d = 16$ minus $d = 2$ effects are $+0.208$ ($[0.137, 0.278]$) for gradient, -0.078 ($[-0.118, -0.038]$) for cyclic, and $+0.070$ ($[0.062, 0.078]$) for random. The gradient–cyclic gap therefore grows by 0.286 ($[0.194, 0.377]$), and the gradient–random gap by 0.138 ($[0.068, 0.208]$). Secondary cross-cell comparisons place gradient $d = 2$ above cyclic $d = 16$ by 0.231 ($[0.155, 0.306]$) and random $d = 16$ by 0.293 ($[0.227, 0.359]$). The separate $d = 8$ analysis supports the orientation comparison with independent seeds; varying d additionally shows how orientation changes the benefit of dimension.

Every basis satisfies the 10^{-5} orthonormality tolerance.



Figure 3: Task-loss optimization on the original Q1 spans at $d = 8$ with a fixed head. Each fitting-node cross-entropy trajectory is divided by its initial value, then averaged over folds and random repeats, anchors, and datasets. Lines are descriptive means; losses are recorded before each Adam update. This is a separate optimization from response reconstruction.

C.6 ESTIMATOR AND RADIUS SENSITIVITY

We repeat the Q1 dimension grid to separate the reconstruction estimator from the evaluated span. Within this rerun, all estimators share anchors, folds, reference responses, bases, and finite-difference tangents at each radius setting. The 25,200 fitted records are repeated measurements, not independent

samples. Folds and five random-basis draws are averaged within each anchor, then datasets within each of ten seed indices ($n = 10$). These rerun estimates are reported separately from Table 6.

For orthonormal weight coordinates, the constrained estimators solve

$$\min_{\|z\|_2 \le \rho} \|T_{\mathrm{fit}} z - y_{\mathrm{fit}}\|_2^2 + \lambda \|z\|_2^2, \tag{57}$$

where T_{fit} is the finite-difference centered-logit tangent and y_{fit} is the reference response on fitting nodes. Constrained ridge uses the ridge coefficient in Equation (56); constrained least squares uses $\lambda = 0$. With $T_{\mathrm{fit}} = U\Sigma V^\top$, the minimum-norm solution has coordinates $V \operatorname{diag}_j[\sigma_j/(\sigma_j^2 + \lambda + \mu)]U^\top y_{\mathrm{fit}}$, with zero contributions from null singular directions. The multiplier is $\mu = 0$ if the unconstrained solution is feasible, and otherwise is chosen so that $\|z\|_2 = \rho$. This solves the stated linear fitting problem, not the nonlinear held-out reconstruction problem. It differs from radially rescaling an unconstrained ridge solution.

Let R denote the original fold-specific weight radius. Under *matched reference/candidate update budgets*, the reference network is refit at radius aR and the candidate receives the same radius, for $a \in \{0.5, 1, 2\}$. Under a *fixed reference target*, the response fitted at R is kept fixed while only the candidate budget changes. The reference logits are never multiplied by a. The finite-difference step is held at its base-R value throughout this grid.

Table 7: Q1 estimator and budget sensitivity. Entries give paired means and 95% t intervals across ten dataset-balanced seed indices. Panel A uses reference/candidate radii R/R. Panel B uses constrained ridge and reports the change in a gradient–control gap from $d = 2$ to 16.

A. Dimension gain: $R_0^2(d = 16) - R_0^2(d = 2)$ Estimator	Gradient	Cyclic	Random
Ridge + rescaling	.2135[.1599, .2672]	−.0876[−.1230, −.0522]	.0710[.0623, .0798]
Constrained ridge	.2353[.1804, .2901]	.0802[.0618, .0987]	.0933[.0808, .1057]
Constrained least squares	.2353[.1804, .2901]	.0802[.0618, .0987]	.0933[.0808, .1057]
B. Interaction: gap at $d = 16$ minus gap at $d = 2$			
Reference/candidate radii	Gradient–cyclic	Gradient–random	Largest Holm p
$0.5R/0.5R$	.1813[.1162, .2464]	.1680[.1133, .2226]	.00226
R/R	.1550[.0928, .2173]	.1420[.0927, .1914]	.00447
$2R/2R$	.1148[.0529, .1767]	.1047[.0583, .1510]	.01863
$R/0.5R$ (fixed target)	.2130[.1570, .2691]	.2071[.1550, .2591]	.000310
$R/2R$ (fixed target)	.0813[−.0072, .1699]	.0179[−.0425, .0782]	1.000

At R/R, constrained fitting reverses the cyclic dimension effect (paired change $+.1678$, $[.1476, .1880]$). The gradient advantage at each tested dimension survives the eight-contrast Holm family for every estimator/radius setting. Interaction estimates remain positive with intervals excluding zero throughout the matched-budget grid. With a fixed target and doubled candidate radius, both interaction intervals include zero; this is not itself a test of a difference between the two protocols. Panel B reports the larger adjusted p of its two contrasts from the secondary sensitivity family. The agreement between constrained ridge and least squares at displayed precision concerns this small ridge coefficient, not an unrestricted regularization sweep.

C.7 Scope extension beyond citation graphs and GCN

To test whether the interaction extends beyond citation graphs and GCN, we use Chameleon and Squirrel with a width-matched two-layer GCN and single-head GAT. The same cross-fit, common-radius, orientation-bank, and held-out R_0^2 protocol is used on the first official split; test masks remain sealed.

Both backbones reproduce the interaction reported on the citation graphs. Pooling the two graphs per backbone, the $d = 16$ minus $d = 2$ effect on R_0^2 is $+0.465$ ($[0.411, 0.519]$) for gradient, $+0.043$ ($[0.007, 0.079]$) for cyclic, and $+0.209$ ($[0.184, 0.235]$) for random under GCN; under the single-head GAT the effects are $+0.323$ ($[0.252, 0.395]$), $+0.062$ ($[0.043, 0.080]$), and $+0.113$ ($[0.097, 0.130]$). The gradient–cyclic difference-in-differences is $+0.422$ ($[0.377, 0.468]$) for GCN and $+0.262$ ($[0.200, 0.324]$) for GAT, each excluding zero. Gradient bases lead both controls at every tested d on both backbones. Cyclic coverage also increases on these graphs, so the replicated

interaction is the larger increase for gradient bases. This study uses a width-matched single-head GAT; the external classification comparison uses multi-head GAT.

C.8 Rank Selection by Euclidean Subspace Overlap

Euclidean subspace overlap provides one criterion for choosing the factor rank in Section A.5. On three Cora anchors and both cross-fit directions, ranks $r \in \{1, 2, 4\}$ were compared with spectrum-matched random orientations using only D_{basis} geometry. A rank had to achieve mean and minimum squared-cosine overlaps of at least 0.80 and 0.70 with the uncompressed discovery-gradient subspace, together with at least $6\times$ compression of stored basis values. Validation reconstruction and accuracy were excluded from rank selection.

No tested rank meets all selection criteria. The mean and minimum overlaps are 0.364 and 0.341 at $r = 1$, 0.545 and 0.497 at $r = 2$, and 0.762 and 0.711 at $r = 4$. The corresponding Cora storage reductions are $29.33\times$, $14.70\times$, and $7.36\times$. Rank four meets the minimum-overlap and storage conditions but falls below the required mean overlap. No rank is selected for confirmation under this rule.

Reconstruction and accuracy provide secondary diagnostics of the tested ranks. Compact-gradient R_0^2 is 0.618, 0.635, and 0.677 for ranks one, two, and four. Rank four exceeds its matched structured-random orientation by 0.480 (95% CI $[0.418, 0.542]$) and differs from the uncompressed reference by -0.014 ($[-0.035, 0.007]$). Thus useful prediction responses can survive compression even when weight-space overlap falls below the selection threshold. These diagnostics were excluded from the selection rule, so rank four remains unselected. The compact-minus-random validation-accuracy interval includes zero at every rank.

C.9 Q2: rank selection and confirmation

Functional fidelity selects the rank through an inner class-balanced cross-fit in D_{basis} at fixed $d = 8$. Five nodes per class fit the atom map and disjoint nodes supply the Jacobian score; D_{fit} and official validation endpoints were excluded. Cora and CiteSeer used three new anchors, both outer folds, ranks $r \in \{1, 2, 4\}$, and spectrum-matched random replicates.

The main-text Figure 2 reports the paired confirmation contrasts, while Table 8 gives the complete absolute scores and paired comparisons.

Table 8: Reconstruction and basis storage of compact gradient maps. (a) Held-out dataset means of nonlinear functional R_0^2 over ten new anchors. (b) Paired contrasts balance datasets within the shared anchor seeds; Holm adjustment applies to the two primary functional contrasts. Rank selection by amplitude-sensitive orthogonal-Procrustes fidelity is reported in the text.

(a) Held-out dataset means

Map	Cora	CiteSeer	PubMed
Cyclic	0.399	0.127	0.098
Matched random $r = 2$	0.195	0.169	0.090
Weight mixing $r = 2$	0.654	0.328	0.732
Functional mixing $r = 2$	0.659	0.346	0.772
Dense functional	0.702	0.346	0.776
Stored factors and mixing coefficients			
Frozen compact values	52,128	124,768	22,272
Dense / compact storage	$14.70\times$	$15.46\times$	$12.97\times$

(b) Paired contrasts

Contrast	mean [95% CI]	Holm p	role
Functional–random	0.442 [0.398, 0.485]	5.48e-09	primary
Functional–weight	0.021 [-0.002, 0.045]	0.0736	primary
Compact–dense	-0.015 [-0.034, 0.003]	–	descriptive
Val. acc.–random	1.791 [0.949, 2.633] pp	–	secondary
Val. acc.–weight	-0.110 [-0.656, 0.436] pp	–	secondary

Rank one has mean and minimum fidelity of 0.762 and 0.643, below the fixed thresholds of 0.80 and 0.70, despite $29.33\times$ storage reduction. Rank four meets both thresholds (0.977 and 0.952) with $7.36\times$ reduction. Rank two is the smallest qualifying rank, with fidelity 0.928 and 0.867 and $14.70\times$ reduction relative to a dense $P_{\mathrm{w}}d$ basis. We fix this rank before downstream confirmation.

Confirmation rebuilds only $r = 2$ from the complete D_{basis} on Cora, CiteSeer, and PubMed using 10 new DGI anchors per dataset. Disjoint D_{fit} nodes train the reference network and fit the tangent coefficients; official validation supplies the held-out nonlinear endpoints. Five random-frame maps per fold preserve the layerwise spectra, factor shapes, payload, optimizer, and radius. A second control uses the same gradient factors with weight-Gram rather than functional mixing.

Compact maps with functional mixing exceed matched random factors by 0.442 in R_0^2 ($[0.398, 0.485]$), while their difference from the dense reference is -0.015 ($[-0.034, 0.003]$). No equivalence margin was specified. Functional mixing adds 0.021 over weight mixing on the same factors ($[-0.002, 0.045]$; Holm $p = 0.074$). Thus the prespecified requirement that both primary comparisons show a benefit is not met: confirmation supports the gradient-factor orientation, while the extra benefit of functional mixing remains unresolved.

C.10 Primary Q3: objective-matched transfer and reconstruction

Primary Q3 uses DGI anchors $s = 120000, \ldots, 120009$, support-sampling seeds $121000+s$ (241000–241009), and optimization seeds $123000 + s + 1000k$. The two-layer width-64 GCN is pretrained for 200 Adam steps at learning rate 0.01 and weight decay 10^{-4}, with dropout 0.5. The masked decoder is fitted on frozen encoder features for 100 steps at learning rate 0.01, zero weight decay, and mask rate 0.3. Discovery forms 16 atoms, with rank two and eight shared coordinates; each objective supplies its own spectrum-matched random control.

All five transfer methods reuse the same in-memory encoder and initial full linear classifier, support indices, and optimizer random stream. Only the support labels enter the 100-step Adam adaptation (lr $= 0.01$, zero weight decay, $\beta = 0.01$, $z_0 = 0$, no coordinate projection). No checkpoint is selected on validation outcomes; the test endpoint is read after the fixed training budget. Discovery uses all graph features and edges. The shared pretrainer reads the global label maximum to size the classifier; all classes also occur in the official training mask. This metadata access does not enter the DGI or masked-feature loss.

Atom seeds start at $122000 + s$ for DGI and $123000 + s$ for masked features; random-frame seeds are $125000 + s$ and $126000 + s$, respectively. The decoder seed is $124000 + s$. Functional cross-fitting uses the deterministic seed rule in the released runner with offset 126000. Its head is refitted on one training fold, the separately fitted reference network and reconstruction coefficients use the other, and validation nodes supply the reported response. The realized encoder spans are orthonormalized for the common-radius comparison. Test endpoints are not used in this fit. Transfer averages the three shot levels within 30 dataset–anchor units; reconstruction averages two folds within those same units.

DGI factors improve accuracy by 2.65 percentage points over head-only adaptation ($[1.73, 3.57]$) and by 2.58 over their random control ($[1.67, 3.49]$). Masked-feature factors improve by 3.19 and 3.16 points, respectively ($[2.14, 4.25]$ and $[2.10, 4.22]$). All four comparisons belong to the prespecified Holm family; the largest adjusted p is 4.67×10^{-6}. On Cora, CiteSeer, and PubMed, DGI gains over head-only adaptation are 3.71, 2.49, and 1.75 points. Masked-feature gains are 4.43, 4.65, and 0.50 points, with the largest improvements on Cora and CiteSeer.

The factor columns include the published layer gains and Gram mixing and are orthonormalized before analysis, preserving their weight span while imposing a common weight-space radius. DGI minus its random control is $+.309$ R_0^2 ($[.195, .424]$; Holm $p = 5.78 \times 10^{-6}$); masked minus its own control is $+.373$ ($[.303, .442]$; Holm $p = 1.38 \times 10^{-11}$). Masked minus DGI is exploratory and crosses zero ($+.067$; $[-.029, .163]$). At the dataset–seed unit, correlations between the functional and transfer aligned-minus-random deltas are $-.187$ for DGI and $-.156$ for masked. These estimates do not establish a positive association in this sample.

Reconstruction–accuracy association. Figure 4 plots all 30 paired dataset–seed effects for each map construction. Transfer first averages the three shot levels; reconstruction first averages the two folds. A post-hoc paired bootstrap resamples the ten anchors independently within each of the three

Table 9: Primary Q3 cohort (anchors 120000–120009): two endpoints for the same label-free encoder spans. (a) Transfer accuracy averages ten anchors and three shot levels; the last column gives pooled target-specific coverage. (b) The common functional protocol reports nonlinear validation R_0^2 for one reference-network displacement target by dataset. It measures target-specific coverage, not functional-image dimension. (c) Each shot-level accuracy averages the same ten anchors; this descriptive breakdown preserves all methods and all three shot levels.

(a) Transfer and pooled reconstruction

Method / span	Cora	CiteSeer	PubMed	Pooled	R_0^2
Head only	56.04	42.95	58.70	52.56	—
DGI-matched random	56.12	43.01	58.76	52.63	.070
DGI-gradient factors	59.74	45.44	60.45	55.21	.379
Masked-matched random	56.01	43.04	58.75	52.60	.074
Masked-gradient factors	60.46	47.61	59.20	55.76	.447

(b) Reference-response reconstruction by dataset

Weight span	Cora	CiteSeer	PubMed	Pooled
Cyclic	.422	.116	.122	.220
Dense task gradient	.718	.346	.755	.606
Rank-two task factors	.701	.318	.731	.583
DGI-matched random	.082	.069	.059	.070
DGI-gradient factors	.332	.195	.612	.379
Masked-matched random	.086	.071	.066	.074
Masked-gradient factors	.567	.278	.495	.447

(c) Transfer accuracy (%) at each shot level

	$k = 1$			$k = 5$			$k = 10$		
Method	Cora	CiteSeer	PubMed	Cora	CiteSeer	PubMed	Cora	CiteSeer	PubMed
Head only	41.76	33.54	47.73	61.95	47.11	62.36	64.40	48.21	66.01
DGI random	41.96	33.56	47.74	61.98	47.11	62.44	64.43	48.37	66.09
DGI factors	44.30	35.17	49.36	66.07	49.20	64.47	68.86	51.96	67.53
Masked random	41.71	33.63	47.81	62.01	47.11	62.41	64.30	48.37	66.03
Masked factors	45.36	36.73	48.01	66.63	51.69	63.52	69.40	54.40	66.07

fixed datasets, preserving both endpoints of each sampled unit. With 10,000 resamples and seed 20260920, the percentile 95% intervals for the pooled Pearson correlations are $[-.543, .164]$ (DGI) and $[-.506, .237]$ (masked). Within-dataset descriptive correlations are $-.325/.356/-.239$ for DGI and $.137/-.206/.142$ for masked (Cora/CiteSeer/PubMed, ten units each). Signs vary across datasets. These exploratory intervals condition on the three datasets and do not test prediction on new graphs.

All evaluated factor spans have rank eight; reconstruction uses official validation nodes and no test outcomes.

C.11 Fixed-budget comparison with LoRA and full encoder tuning

We compare ten methods across three datasets, ten anchors, and three shot levels (900 adaptation runs). Inference averages shots within each of the 30 dataset–anchor units. Every method reused the same in-memory anchor, support indices, initial full linear head, adaptation seed, 100-step Adam budget ($\mathrm{lr} = 0.01$, zero weight decay), and terminal test endpoint; there was no validation checkpoint selection. LoRA updates both encoder weight matrices with ranks $r \in \{1, 2, 4, 8\}$ and scale $\alpha/r = 1$. Its A factors use Kaiming-uniform initialization, B factors start at zero, and the full linear head is trained. Encoder biases stay frozen for the factor methods and LoRA; full encoder tuning also trains these biases. This is a fixed-step controlled comparison, not a method-specific hyperparameter-optimization study.

This rerun reuses the primary Q3 nominal seed schedule and support sets, but its saved anchor weights differ from the earlier run. The cause of this cross-run difference has not been resolved. Within the rerun, anchor and initial-head equality across methods was verified. We therefore make all

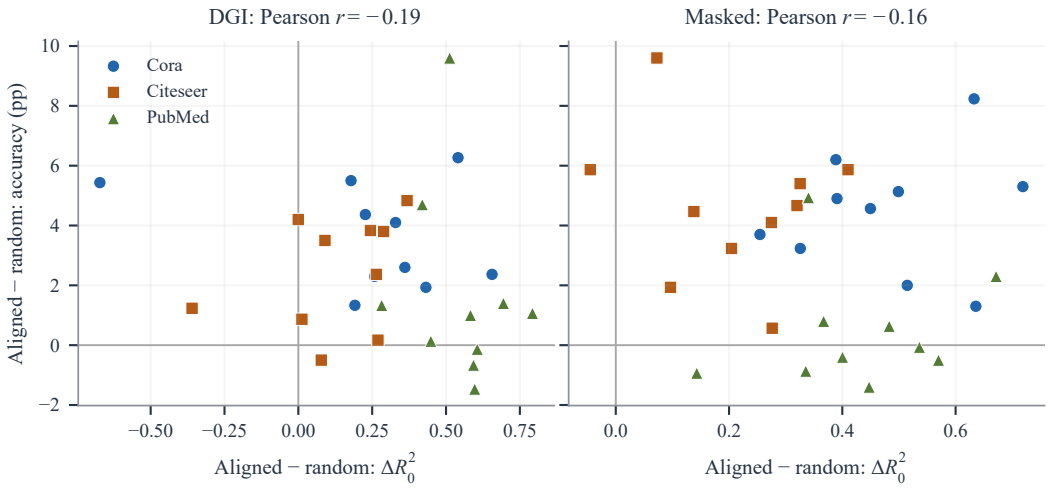

Figure 4: Reconstruction and accuracy gains of label-free gradient factors over their objective-matched random controls. Each panel contains all 30 dataset–anchor units and uses the same axis limits. Accuracy gains are in percentage points (pp). Titles give Pearson correlations with exploratory 95% percentile intervals from 10,000 bootstrap samples stratified by dataset. Positive average gains at both endpoints do not establish a positive association between their per-unit magnitudes.

comparisons within each run: the new LoRA rows are not paired with the old head-only rows, and these accuracies are not substituted into the earlier reconstruction–accuracy correlation.

Table 10: Fixed-budget few-shot comparison with LoRA and full encoder tuning (30 dataset–anchor units; accuracies average the three shot levels). All accuracy columns are percentages; dataset means average ten anchors and pooled means weight datasets equally. Trainable values and float32 payload bytes refer to Cora and include the full linear head; payload includes all trainable values and method-specific frozen factors, but excludes the shared anchor.

Method	Cora	CiteSeer	PubMed	Pooled (%)	Trainable	Bytes
Head only	56.553	43.847	58.613	53.004	455	1,820
DGI matched random	56.757	43.943	58.660	53.120	463	210,364
DGI factors	59.753	47.603	60.643	56.000	463	210,364
Masked matched random	56.560	43.933	58.617	53.037	463	210,364
Masked factors	60.393	48.947	59.477	56.272	463	210,364
LoRA-$r = 1$	61.013	46.267	61.750	56.343	2,080	8,320
LoRA-$r = 2$	62.240	48.840	62.757	57.946	3,705	14,820
LoRA-$r = 4$	62.420	50.190	62.500	58.370	6,955	27,820
LoRA-$r = 8$	62.523	51.453	63.277	59.084	13,455	53,820
Full encoder	63.150	50.863	64.347	59.453	96,391	385,564

The four MapPerturb contrasts are paired within the 30 dataset–anchor units. DGI and masked factors improve over head-only by $+2.996$ ($[1.903, 4.088]$) and $+3.268$ ($[2.119, 4.417]$) percentage points, and over their objective-matched random controls by $+2.880$ ($[1.834, 3.926]$) and $+3.236$ ($[2.092, 4.379]$), respectively; all four Holm-adjusted values are $p = 1.05 \times 10^{-5}$. LoRA-$r = 4$, LoRA-$r = 8$, and full encoder tuning exceed both factor variants in the separate 15-contrast external family (each external method against head-only and the two factor variants). The LoRA-$r = 1$ and $r = 2$ contrasts against the factor variants do not survive this correction; non-rejection does not establish equivalence. These fixed-rank results do not evaluate a rank-selection policy. VeRA and GiVA appear in the separate Q4 protocol, not this cohort.

C.12 Metric choices and comparison scope

The zero baseline in R_0^2 is the no-update response, rather than the sample-mean baseline of conventional R^2. Reconstruction measures target-specific response coverage; classification accuracy measures task performance. Sensitivity, weight capture, and participation rank describe different aspects of local geometry. Paired inference uses anchor or split blocks, not individual folds, shots, or

random-basis draws. Appendix F.4 complements accuracy with class-balanced task metrics from a separate, fully paired cohort. Following the distinction between optimization and storage costs in PEFT (Hu et al., 2022; Kopiczko et al., 2024), our payload counts include method-specific frozen factors and the learned head but exclude the shared anchor. These counts do not measure peak memory or end-to-end runtime.

C.13 Broader parameter-efficient baselines

This earlier cohort provides within-cohort baseline and resource comparisons. Its anchors 90000–90009 differ from the primary Q3 anchors 120000–120009, so its accuracy estimates are not pooled with Table 2. Support sampling uses seed $91000 + s$ for anchor seed s, giving 181000–181009. Optimization uses $93000 + s + 1000k$. All methods share their cohort's anchor, support nodes, full linear head, and 100 Adam steps at learning rate 0.01 and zero weight decay. The secondary weight-decay runs use 5×10^{-4}. Unlike primary Q3, the original random factor control uses DGI spectra for both objectives. The prompt and bottleneck baselines implement the specified parameterizations on this GCN; they are not reproductions of complete GraphLoRA, G-Adapter, or AdapterGNN systems.

Table 11: Historical baseline cohort (anchors 90000–90009), distinct from primary Q3. (a) Dataset means average ten anchors and three shot levels; pooled gives datasets equal weight. (b) Each contrast treats one dataset–seed average as an independent unit. The first four rows formed the prespecified Holm family; the original masked–random row is not treated as an orientation control because it used DGI rather than masked-atom spectra.

(a) Accuracy by dataset

Method	Cora	CiteSeer	PubMed	Pooled
Head only	58.09	44.33	61.32	54.58
Cyclic $d = 8$	62.49	47.87	61.01	57.12
Matched random factors	58.20	44.42	61.35	54.66
Task-gradient factors	61.81	48.06	62.27	57.38
DGI-gradient factors	62.30	48.30	62.29	57.63
Masked-gradient factors	63.87	49.38	61.61	58.29
GPF feature prompt	63.02	44.45	53.67	53.71
Bottleneck $b = 16$	60.88	44.69	63.18	56.25
LoRA $r = 2$	62.52	47.17	64.42	58.03
LoRA $r = 4$	64.04	51.11	65.04	60.06
LoRA $r = 8$	62.30	50.43	65.06	59.26
Full fine-tuning	66.06	51.78	65.88	61.24

(b) Clustered contrasts

Contrast	Δ points	95% CI	p_t
DGI − head	+3.05	[1.82,4.28]	2.04×10^{-5}
DGI − random	+2.97	[1.75,4.20]	2.75×10^{-5}
Masked − head	+3.71	[2.31,5.11]	8.05×10^{-6}
Masked − random	+3.63	[2.23,5.03]	1.06×10^{-5}
DGI − cyclic	+.51	[−.39,1.41]	.259
Task − DGI	−.25	[−.89,.39]	.432
LoRA-$r = 2$ − DGI	+.40	[−1.21,2.01]	.613
Full FT − DGI	+3.61	[1.98,5.24]	9.37×10^{-5}

The four originally specified Holm-adjusted values are at most 4.08×10^{-5}. The DGI–random comparison is a valid DGI-spectrum orientation control. The masked–random comparison matches trainable count, factor shape, rank, and payload but not masked-atom spectra. The objective-matched study in Appendix C.10 addresses this difference. Masked–head remains a valid task comparison. Post-report secondary contrasts place masked factors $+.66$ points above DGI ($[-.14, 1.46]$), $+1.17$ above cyclic ($[.47, 1.86]$; three-test Holm $p = .0056$), and $+.26$ above LoRA-$r = 2$ ($[-1.56, 2.08]$). Weight-decay sensitivity is negligible in this seed set: zero minus 5×10^{-4} is $+.018$ points for both head-only and DGI factors, with both intervals crossing zero.

For Cora, DGI/masked factors train 463 values, store 52,128 extra frozen values, and use 210,364 adapter bytes in float32. LoRA-$r = 2$ trains 3,705 values and uses 14,820 adapter bytes with no extra frozen payload. Mean adaptation time is .55 seconds for gradient-factor adaptation and .48

Table 12: Baseline comparability within the historical cohort. Counts are for Cora; frozen values exclude the common anchor. All rows use the same support split and full head, Adam with 100 steps, learning rate 0.01, and zero weight decay, followed by terminal test evaluation. The task-factor row uses support labels during discovery. Primary Q3 repeats the head and factor parameterizations on new anchors and adds the masked-spectrum random control. GPF, bottleneck, and LoRA are direct implementations of their parameterizations; graph-specific systems cited in Related Work supply no numerical rows.

Method	Adaptation location	Discovery / initialization	Trainable	Extra frozen
Head only	Head	None	455	0
Cyclic $d = 8$	Encoder + head	Fixed rule	463	0
DGI factors	Encoder + head	DGI	463	52,128
Masked factors	Encoder + head	Masked features	463	52,128
Random factors	Encoder + head	DGI spectra	463	52,128
Task factors	Encoder + head	Support labels	463	52,128
GPF prompt	Input + head	Zero prompt	1,888	0
Bottleneck $b = 16$	Hidden + head	Identity update	4,711	0
LoRA $r = 2$	Encoder + head	Zero update	3,705	0
LoRA $r = 4$	Encoder + head	Zero update	6,955	0
LoRA $r = 8$	Encoder + head	Zero update	13,455	0
Full tuning	Encoder + head	Pretrained weights	96,391	0

for LoRA-$r = 2$; mean incremental allocation is about 36.6–36.9 MB and 39.2 MB, respectively. Discovery time is additional. These measurements distinguish trainable-state reduction from total adapter storage and execution cost.

C.14 Cyclic updates with a class-aware head

We test whether eight cyclic encoder coordinates improve adaptation from a DGI anchor when paired with a class-aware classifier. The classifier uses rank-four factors; its head-only control shares the initialization, random stream, and labels, so the models differ by eight trainable values. Its realized rank is at most $\min(4, C)$, hence at most three on PubMed. This supplies output directions that the cyclic head cannot represent (Proposition 1).

The primary contrast is MapPerturb+rank-budget-4 head minus its matched rank-budget-4 head. Across 30 independent dataset–seed clusters, the mean increment is $+1.75$ points. The 95% CI runs from $+0.95$ to $+2.55$; Monte Carlo sign-flip $p = 0.00013$ using 100,000 Monte Carlo permutations, Wilcoxon $p = 9.9 \times 10^{-5}$, and Cohen's $d_z = 0.82$. The hybrid wins 26 clusters and loses four; 19 cluster differences exceed $+0.5$ points, seven lie within the descriptive ± 0.5-point band, and four fall below it. Dataset-level increments are $+3.28$ points on Cora (95% CI from $+2.00$ to $+4.56$), $+1.89$ on CiteSeer (from $+0.24$ to $+3.55$), and $+0.07$ on PubMed (from -0.41 to $+0.55$). After Holm correction, only Cora at $k = 5, 10$ and CiteSeer at $k = 10$ are significant among the nine secondary contrasts.

Table 13: Confirmatory DGI few-shot results on 10 disjoint seeds. Columns report test accuracy (%) with 95% t intervals for the rank-budget-4 head, the same head plus MP-8, a full head, and validation-selected ScaleShift; the paired difference is MapPerturb+rank-budget-4 head minus its identically initialized rank-budget-4 control, with Holm-adjusted p and win/tie/loss counts.

Dataset	k	Rank-budget-4	+MP-8	Full head	ScaleShift-Val	Paired Δ	p_{Holm}	W/T/L
Cora	1	41.6 [38.7,44.5]	43.3 [39.0,47.7]	47.8 [44.0,51.6]	54.3 [49.4,59.2]	+1.74 [-0.66,+4.14]	0.562	6/1/3
Cora	5	51.8 [46.5,57.1]	55.2 [50.1,60.3]	59.7 [57.1,62.3]	71.4 [69.1,73.6]	+3.38 [+2.24,+4.52]	0.018	10/0/0
Cora	10	54.7 [49.9,59.5]	59.4 [55.5,63.2]	63.9 [61.9,65.9]	73.4 [70.9,75.9]	+4.72 [+2.44,+7.00]	0.018	10/0/0
CiteSeer	1	34.0 [27.9,40.1]	34.8 [29.4,40.2]	39.1 [34.3,43.9]	44.4 [39.7,49.2]	+0.74 [-1.29,+2.77]	0.609	7/0/3
CiteSeer	5	43.4 [38.5,48.3]	45.8 [41.5,50.2]	47.7 [43.1,52.3]	53.5 [50.4,56.6]	+2.41 [+0.07,+4.75]	0.293	7/0/3
CiteSeer	10	48.0 [44.1,51.9]	50.5 [47.9,53.2]	50.7 [46.5,54.9]	57.2 [53.0,61.4]	+2.53 [+0.51,+4.55]	0.041	9/0/1
PubMed	1	53.0 [44.9,61.2]	52.5 [44.4,60.7]	50.9 [41.3,60.5]	55.6 [46.4,64.7]	-0.50 [-1.27,+0.27]	0.609	4/0/6
PubMed	5	59.6 [53.8,65.4]	59.9 [54.3,65.5]	61.4 [57.1,65.8]	69.6 [67.3,72.0]	+0.34 [-0.26,+0.94]	0.609	7/1/2
PubMed	10	64.1 [60.2,68.0]	64.5 [60.8,68.2]	64.5 [60.7,68.3]	72.8 [71.2,74.3]	+0.37 [-0.07,+0.81]	0.488	6/0/4

A validation-only control matches rank-four initialization, trainable values, fixed-basis payload, optimizer, and class-aware head capacity. Across 30 dataset–seed clusters, the mean difference

Table 14: Class-aware-head performance and resource accounting. Each row averages the 90 selected dataset–shot–seed runs; P_{train} spans the trainable-value counts across configurations. Payload includes trainable and extra frozen adapter values; memory and time remain implementation- and hardware-dependent.

Method	P_{train}	Payload (KiB)	Adam state (KiB)	ΔVRAM (MiB)	Time (s)	Macro acc. (%)
Rank-budget-4 head	271–291	1.10	2.22	29.1	0.436	50.0
+MP-8	279–299	1.14	2.29	40.3	0.616	51.8
Full head	195–455	1.35	2.72	29.2	0.378	54.0
ScaleShift-Val	451–711	2.35	4.73	35.8	0.583	61.3
LoRA-Val	887–15970	25.95	51.93	36.4	0.572	63.1
rsLoRA-Val	887–15970	28.75	57.52	36.3	0.586	63.1
Full tuning	36419–241606	487.52	975.07	34.0	0.527	61.9

is $+0.002$ points with 95% CI $[-0.563, +0.568]$ and paired t-test $p = 0.994$ and sign-flip $p = 0.997$. This comparison does not establish a benefit from allocating the matched budget to encoder coordinates.

The primary contrast supports an encoder contribution conditional on the small class-aware head. In the same 90 selected configurations, the macro-average accuracies are 51.8% for the hybrid, 54.0% for a full head, 61.3% for ScaleShift, and about 63.1% for both validation-selected LoRA and rsLoRA. MP-8 alone has a macro-average of 19.2%, close to the 19.1% Frozen DGI anchor. Encoder coordinates alone therefore do not replace a class-aware head in this protocol. The hybrid coordinate vector reaches its $\|\mathbf{z}\|_2 \leq 2$ boundary in 60/90 runs.

C.15 Retrospective map selection

We next ask whether reconstruction can select an adapter at a fixed anchor. The analysis averages each family's two folds in the 420 archived functional records, leaving 30 dataset–anchor units. The mapped models and reference network use official-training labels. Coverage uses the reference network's responses on validation nodes without their labels, while selection quality is measured on those same nodes using validation loss and accuracy. The analysis is therefore retrospective. We report regret relative to each candidate pool's per-anchor oracle and top-1 rates that count ties as correct selections.

The seven-family diagnostic pool includes dense task-gradient directions and two objective-matched random controls. Mean within-anchor Spearman correlation of R_0^2 with validation loss is -0.654 (30 units), but R_0^2 selects the minimum-loss map on only 0.400 of anchors and has mean loss regret 0.0196; always choosing the dense gradient basis has top-1 0.467 and regret 0.0189. The gradient-capture selector has top-1 0.500 and regret 0.0173, but selects dense on 29 of 30 anchors. A correlation with candidate ordering therefore does not by itself establish an anchor-adaptive or cost-effective selection policy. Dense bases store a large $P_{\text{w}}d$ payload; matched-random spans are mechanistic controls, not deployment choices.

The smaller pool {cyclic, DGI factor, masked-feature factor} instead compares three potential label-free *construction* policies at fixed $d = 8$ and per-layer factor rank $r = 2$ under one orthonormalized dense-basis evaluation protocol. In this *post-hoc* subset, R_0^2 chooses masked factor on 18 anchors, DGI factor on 11, and cyclic on one. Table 15 compares selection with the in-sample best fixed family, same-dataset leave-one-anchor-out prior, and leave-one-dataset-out prior. The R_0^2 selector's mean loss regret is 0.0047, versus 0.0096 for always-masked; the paired mean loss difference is -0.0049 with a dataset-stratified anchor-bootstrap 95% interval $[-0.0077, -0.0024]$. This is exploratory: the candidate subset was chosen after inspecting the earlier development results, the best-fixed comparator is itself in-sample, and these are validation rather than test outcomes. The Cora R_0^2 policy always chooses masked, so its advantage over that fixed family is zero there; the differences arise on CiteSeer and PubMed. The accuracy top-1 rate is 0.367, so the loss result does not establish reliable accuracy-based selection. Neither score is label-free at *selection* time: the fitted reference network and fit-task gradient use training labels.

Coverage requires a full-weight reference network, finite-difference tangents, and ridge fitting. Comparable policy-level timings are unavailable, so the analysis does not establish a selection-cost advantage. Exact calculations and policy frequencies are recorded in `analyze_p0_policy_audit.py`. The prospective study below tests selection on new anchors while retaining $d = 8$ and $r = 2$.

Table 15: Retrospective map selection from a post-hoc three-family compact candidate pool. The unit is one dataset–anchor after averaging the two training cross-fit folds. Loss regret is selected minus minimum validation loss; accuracy regret is maximum minus selected validation accuracy. Top-1 counts ties as hits. The best fixed family is selected in-sample and is not a prospective baseline. All endpoints are validation-only.

Policy	Loss top-1	Loss regret	Accuracy top-1	Accuracy regret
R_0^2 selector	0.733	0.0047	0.367	0.0065
Gradient-capture selector	0.600	0.0060	0.400	0.0062
Best fixed family (in sample)	0.367	0.0096	0.267	0.0077
Same-dataset leave-one-anchor-out	0.500	0.0085	0.400	0.0076
Leave-one-dataset-out prior	0.267	0.0146	0.367	0.0064

C.16 Prospective selection of compact maps on new anchors

We compare five selection policies on ten new DGI anchors per dataset (seeds 140000–140009). The candidate maps—cyclic, DGI factor, and masked-feature factor—and the values $d = 8$ and $r = 2$ were fixed before evaluation. The R_0^2 policy uses a reference network fitted to training labels and validation-node logits without validation labels. Gradient capture uses a training-fold task gradient; train-all selection fits all three maps and chooses by validation loss. Two policies use the earlier development results: a fixed prior always chooses masked factor, while a leave-one-dataset-out prior chooses DGI on Cora and masked factor elsewhere.

In each cross-fit direction, map construction uses no task labels. One half of the official training set fits the head and radius; the other fits the reference network and map coordinates and supplies the task gradient. Policies average the two folds and fix all 30 dataset–anchor choices before any candidate test outcome is evaluated. The test oracle is computed afterward to measure regret. The pretrainer obtains the class count from the full label vector, although the official training mask already contains every class on all three datasets. Thus selection includes this metadata access but uses no test outcomes.

Table 16: Selection of compact maps on new anchors. Each policy fixes one family per dataset–anchor before the official test endpoint is released. Loss regret is selected minus per-anchor minimum test cross-entropy; accuracy regret is per-anchor maximum minus selected test accuracy. Top-1 counts ties. Entries average 30 dataset–anchor units with equal dataset weight. The test oracle is a reporting reference, not a selectable policy.

Policy	Loss top-1	Loss regret	Accuracy top-1	Accuracy regret
R_0^2	.800	.00634	.667	.00488
Fit-gradient capture	.767	.00270	.633	.00617
Fixed development prior	.467	.01009	.400	.00955
Leave-one-dataset-out prior	.267	.01925	.400	.00663
Train all; choose by validation loss	.833	.00076	.700	.00448

The primary comparison tests whether R_0^2 selection reduces test loss relative to the fixed development prior. Its mean difference is -0.00375, with a dataset-stratified bootstrap 95% interval of $[-0.00988, 0.00340]$, which does not meet the prespecified criterion of an interval below zero. The difference is $+0.01389$ on Cora, -0.00363 on CiteSeer, and -0.02152 on PubMed. The R_0^2 policy is also worse than train-all validation-loss selection by $+0.00558$ test-loss units, with a dataset-stratified interval $[+0.00049, +0.01177]$. Its top-1 success can therefore coexist with a worse *mean* regret when mistaken choices have larger loss gaps. The gradient-capture policy was a specified secondary comparator: its test-loss difference from the fixed prior is -0.00739 with an unadjusted interval $[-0.01116, -0.00360]$ from a separate secondary analysis. It chooses DGI on nine anchors and masked factor on 21; its Cora choices all match the fixed prior. This secondary comparison does not establish an advantage over train-all selection.

The full comparison takes a mean 7.79 seconds per anchor and peaks at 457.3 MB of allocated CUDA memory. This includes constructing all three candidates, evaluating the reference network and finite-difference probes, and fitting every map's coordinates for the oracle comparison; it does not measure individual policy costs. The R_0^2 score requires a full-weight reference network, whereas gradient capture does not. To impose the same weight-radius constraint, we evaluate each map

through a dense orthonormal basis of its encoder span with a common projected optimizer. The study therefore compares span orientation at fixed $d = 8$ and $r = 2$; it does not measure the execution cost of factorized adapters or select their dimension and factor rank.

C.17 Official-Mask Cold-Start Boundary

We test bounded fixed cyclic charts around random GCN weights on Cora, CiteSeer, and PubMed at seven coordinate dimensions from 4 to 256, with full-matrix controls at 8, 32, and 128. All 90 dataset–shot–seed configurations use the official masks, and method selection excludes candidate test outcomes.

Table 17: Official-mask cold-start confirmation. Each entry is official-test accuracy (%) with a 95% t interval over 10 independent seeds after averaging $k \in \{1, 5, 10\}$ within seed. IS and FM denote input-shared and full-matrix fixed cyclic charts. Random-head and Scratch controls show that the labels are learnable under the same adaptation budget.

Method	Cora	CiteSeer	PubMed
Chance	14.3	16.7	33.3
Random frozen	14.5 [10.6,18.3]	15.8 [14.0,17.7]	32.3 [27.7,36.8]
Cold IS-8	14.8 [11.3,18.4]	16.3 [14.4,18.1]	32.0 [26.8,37.1]
Cold FM-8	15.0 [11.6,18.3]	16.4 [14.8,18.0]	31.9 [26.7,37.1]
Random head	45.4 [42.8,48.0]	32.1 [28.9,35.3]	51.2 [48.0,54.3]
Scratch	63.9 [61.0,66.9]	52.0 [48.7,55.2]	66.6 [64.4,68.7]
Two-phase IS-8	80.8 [80.1,81.4]	68.7 [67.8,69.5]	78.9 [78.6,79.1]
Supervised frozen	80.9 [80.3,81.4]	69.2 [68.6,69.7]	78.9 [78.6,79.2]

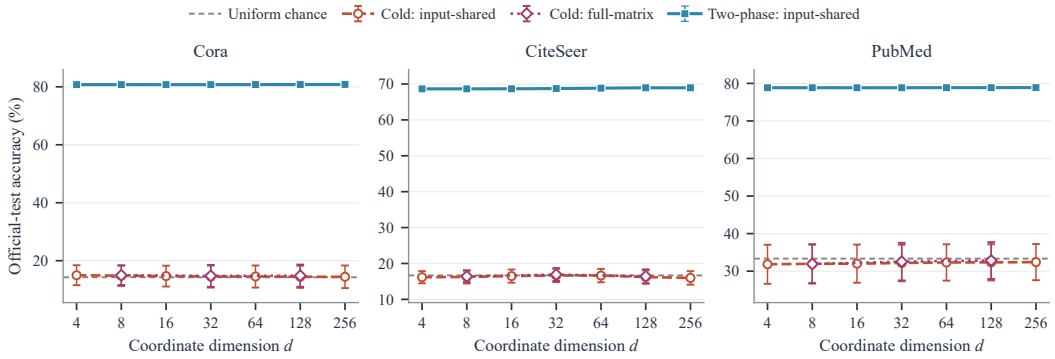

Figure 5: Cold-start boundary across the tested coordinate counts. Both fixed cyclic charts remain near uniform-chance accuracy, whereas the two-phase input-shared chart recovers the supervised-anchor level. Error bars are 95% t intervals over ten seeds after averaging the three shot levels within seed. This is a boundary result for the tested fixed charts, not a claim about learned or data-dependent generators.

At $d = 8$, cold input-shared accuracy is 14.8%, 16.3%, and 32.0% on Cora, CiteSeer, and PubMed, against chance levels 14.3%, 16.7%, and 33.3%. The full-matrix control is similarly close to chance. In contrast, the random last-layer control reaches 45.4%, 32.1%, and 51.2%, while Scratch reaches 63.9%, 52.0%, and 66.6% when averaged over the same shot levels. These controls rule out the explanation that the official labels or optimizer are intrinsically unusable. The two-phase $d = 8$ chart reaches 80.8%, 68.7%, and 78.9%, matching the corresponding supervised Frozen anchors.

Across the 1,800 chart runs, 1,083 (60.2%) finish on the coordinate-ball boundary, mostly at larger dimensions. Every run has a nonzero initial gradient and update in coordinate space, so near-chance performance persists under active optimization. This result concerns the tested bounded fixed charts; learned generators and data-dependent bases are outside its scope.

C.18 Heterophily Stress Test

The heterophily run evaluates the original input-shared MP-8 chart, a wider block-constant chart, last-layer tuning, and LoRA on all 10 official splits of Actor, Cornell, Texas, and Wisconsin. Each

split receives a separately trained three-layer GCN anchor. Adapters train on the official training mask; the official validation mask selects an epoch from $\{0, 1, 10, \ldots, 100\}$, where epoch 0 is an explicit no-adaptation option. The adapter winner is selected without test metrics, but the shared anchor-training helper logged test performance at evaluation checkpoints; this is therefore a validation-selected diagnostic rather than a sealed-test confirmation. The original MP-8 chart changes mean accuracy by only $+0.08$, 0.00, 0.00, and -0.20 points on Actor, Cornell, Texas, and Wisconsin, respectively; every 95% interval includes zero. Recorded test accuracy on Cornell and Texas is unchanged on all 10 splits despite nonzero training gradients, which is consistent with a practical plasticity limit rather than a completely inactive optimization path. The wider block-constant chart (Block-128) also has intervals containing zero. Last-layer tuning and LoRA show larger positive means on Wisconsin and Texas, but their intervals remain wide in this one-seed-per-split analysis. These intervals do not establish equivalence to the frozen anchor.

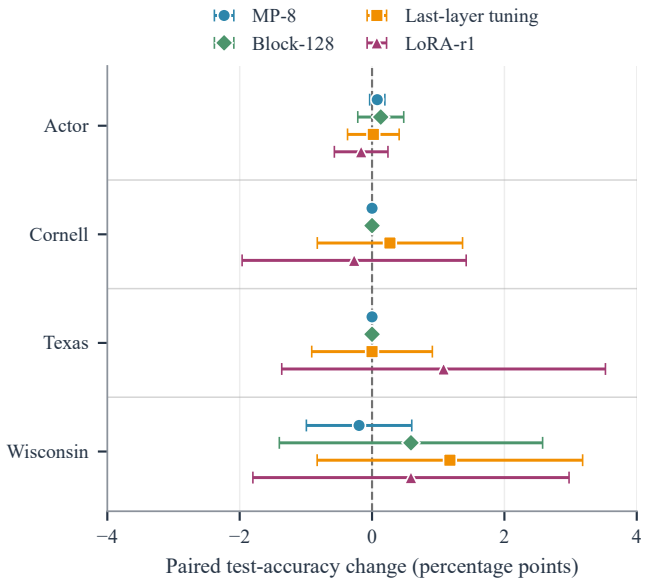

Figure 6: Validation-selected heterophily stress test over ten official splits. Points show paired test-accuracy changes from each split's frozen anchor, with 95% t intervals. The anchor-training routine also logged test accuracy during training, so test access was not restricted to the final evaluation.

C.19 Secondary Diagnostics and Scope

Across GCN, GAT, and GIN, the fixed chart gives small mixed deltas relative to Frozen. On PROTEINS, MUTAG, and NCI1, repeated-split graph-classification differences from head-only tuning include zero in their paired intervals. Full deep pretraining allows a shared chart to track the tested GCN depths. Expansion-structure ablations favor no single pattern over full-matrix, block, or fixed-random maps.

Earlier large-graph and graph-level pilots do not establish an adaptation advantage. The initial ogbn-arxiv sweep places MP-8 below its supervised anchor; its test-selected $d_{\text{eff}} = 8$ is not an estimate of intrinsic dimension. In the five-seed MolHIV validation-only study, all paired intervals against Frozen include zero. The PPI validation rule selects the unadapted model for every method. The products study ended in host-memory exhaustion and contributes no accuracy evidence.

The beta-squared Hessian plot checks the affine pullback identity. A separate controlled SBM fit has slope -1.66 for $\log(\text{stability})$ versus $L \log(1 - \lambda_1)$ ($R^2 = 0.52$); because a simple $(1 - \lambda_1)^L$ decay law predicts a positive slope, this experiment does *not* validate that law. Appendix F.2 shows the dimension, depth, architecture, Hessian, and spectral diagnostics.

D LIMITATIONS AND SCOPE

Local response and selection. The tangent and quadratic analysis describe behavior around a fixed anchor. The covering and convergence results additionally assume a low-dimensional tube and coordinate-space PL conditions, respectively. Neither the sampled trajectories nor the Hessian diagnostic verifies those assumptions for GNN training. Reconstruction measures a specified response of the reference network; primary Q3 does not establish a positive association between its gains and accuracy gains. The prospective selection study did not resolve an advantage over its fixed development prior, and computing coverage requires a supervised reference network and finite-difference evaluations (Appendix C.16).

Data access and evaluation range. Label-free direction discovery uses a transductive DGI anchor and the full graph's features and edges. Adaptation still requires support labels. The shared pretrainer also reads class-count metadata from the global label vector, as detailed in Appendix C.10. Supervised-anchor diagnostics are separate from this transfer setting. The dimension sweep and rank-two construction cover the tested graphs and budgets. External graph families had been seen during development, and the five-anchor ogbn-arxiv gain intervals include zero. These evaluations do not establish an inductive or universal adapter-selection result.

Resources and comparisons. Eight encoder coordinates reduce trainable state while retaining the anchor and frozen factors. Construction still forms full-weight gradients, and the executed update sums 16 rank-two terms per layer. Basis storage savings over dense maps do not imply lower total storage, runtime, or activation memory than LoRA or head-only adaptation. The historical baseline table reports comparisons within its own cohort; the prompt, bottleneck, and low-rank controls implement their parameterizations on a common GCN (Table 12). The additional gain from functional mixing remains unresolved.

Numerical reproducibility. CUDA/PyG sparse operations are seeded but not bitwise deterministic, and timing and memory depend on hardware. A same-budget Cora repeat changes fold-averaged nonlinear R_0^2 by 0.036 in mean absolute value while preserving the gradient-above-cyclic-above-random ordering.

E COMPLETE PROOFS

E.1 QUADRATIC TARGET-ALIGNMENT IDENTITY

Proof. Let $H \in \mathbb{R}^{P\times P}$ be symmetric positive definite, $g \in \mathbb{R}^P$, and $\bar{E} \in \mathbb{R}^{P\times d}$. The unrestricted minimizer is $\delta^\star = -H^{-1}g$, and completing the square gives

$$q(\delta) = q(0) - \tfrac{1}{2}g^\top H^{-1}g + \tfrac{1}{2}\|H^{1/2}\delta + H^{-1/2}g\|_2^2. \tag{58}$$

For mapped updates, $H^{1/2}\delta = H^{1/2}\bar{E}z$ ranges over $\mathrm{col}(H^{1/2}\bar{E})$. Its best approximation to $-H^{-1/2}g$ is the orthogonal projection onto that column space, whether or not $\bar{E}$ has full column rank. The residual norm is therefore $\|(I - P_{H^{1/2}\bar{E}})H^{-1/2}g\|_2$. Subtracting $q(\delta^\star)$ proves Eqs. (13) and (40). □

If $H \succeq 0$, a finite unrestricted minimum additionally requires $g \in \mathrm{range}(H)$. Under that condition the same argument uses $\delta^\star = -H^\dagger g$ and $(H^\dagger)^{1/2}g$ in place of $-H^{-1}g$ and $H^{-1/2}g$. Null-space displacements do not change the quadratic value. If the range condition fails, moving along a null vector with a nonzero gradient component makes q unbounded below; the finite-gap identity is then inapplicable. A radius-constrained mapped problem can only have a larger minimum than the unrestricted mapped problem.

E.2 FUNCTIONAL MIXING AND ORTHOGONAL TANGENT FIDELITY

Proof. Let $A \in \mathbb{R}^{P_w\times M}$ have rank $p \geq d$. On its positive Gram eigenspace, let $U \in \mathbb{R}^{M\times p}$ and $\Lambda \in \mathbb{R}^{p\times p}$. Then $W = AU\Lambda^{-1/2}$ satisfies $W^\top W = I_p$. Every orthonormal d-frame in $\mathrm{col}(A)$ can

be written as WS with $S^\top S = I_d$. Its discovery sensitivity is

$$\|F_D S\|_F^2 = \operatorname{tr}(S^\top H_D S) \le \sum_{j=1}^{d} \lambda_j(H_D), \tag{59}$$

where eigenvalues are ordered from largest to smallest. To see the bound, diagonalize $H_D = V \operatorname{diag}(\lambda_j) V^\top$ and set $B = V^\top S$. The row norms $w_j = \|B_{j,:}\|_2^2$ satisfy $0 \le w_j \le 1$ and $\sum_j w_j = d$, so $\sum_j \lambda_j w_j \le \sum_{j=1}^{d} \lambda_j$. Choosing the leading eigenvectors attains equality. Thus $Q_F = WS_d$ is optimal for this total-sensitivity criterion. The criterion contains no target response and does not itself minimize target reconstruction error.

For $T_0, T_r \in \mathbb{R}^{m \times d}$, expand the Procrustes residual:

$$\min_{O^\top O = I_d} \|T_0 - T_r O\|_F^2 = \|T_0\|_F^2 + \|T_r\|_F^2 - 2\|T_r^\top T_0\|_*. \tag{60}$$

Indeed, if $T_r^\top T_0 = U_c \Sigma_c V_c^\top$, the trace term is maximized by $O = U_c V_c^\top$ and equals $\operatorname{tr}(\Sigma_c)$. The value is invariant to independent orthogonal changes of tangent coordinates. When $T_r = aT_0$ with $a \ge 0$, fidelity equals $1 - (1-a)^2$, so amplitude mismatch remains visible. For a nonzero reference the score is at most one, but can be negative; it is not a probability. □

E.3 Proof of Theorem 1

Proof. Let d_{func} be the functional metric under which the model-output map is L_f-Lipschitz in parameters. Fix $\varepsilon > L_f \tau$ and set $\alpha = \varepsilon / L_f - \tau > 0$. The Euclidean ball $B_q(0, r)$ has an α-net $\{\mathbf{a}_j\}_{j=1}^N$ with

$$N \le \left(1 + \frac{2r}{\alpha}\right)^q = \left(1 + \frac{2rL_f}{\varepsilon - L_f \tau}\right)^q. \tag{61}$$

For any $\theta \in \mathcal{M}_{\text{local}}$, the tube assumption provides an $\mathbf{a}(\theta)$ with $\|\mathbf{a}(\theta)\|_2 \le r$ and $\|\theta - \theta^* - \mathbf{U}\mathbf{a}(\theta)\|_2 \le \tau$. Choose $\mathbf{a}_j$ such that $\|\mathbf{a}(\theta) - \mathbf{a}_j\|_2 \le \alpha$. Since $\mathbf{U}^\top \mathbf{U} = \mathbf{I}_q$,

$$\|\theta - (\theta^* + \mathbf{U}\mathbf{a}_j)\|_2 \le \|\theta - \theta^* - \mathbf{U}\mathbf{a}(\theta)\|_2 + \|\mathbf{U}(\mathbf{a}(\theta) - \mathbf{a}_j)\|_2 \tag{62}$$

$$\le \tau + \alpha = \frac{\varepsilon}{L_f}. \tag{63}$$

Therefore $d_{\text{func}}(f_\theta, f_{\theta^* + \mathbf{U}\mathbf{a}_j}) \le \varepsilon$. The functions indexed by the coefficient net form the required cover, and Eq. (61) proves the claim with an explicit constant. No Hessian property is used. □

Remark 1 (Scope of the covering assumption). A low-rank optimization trajectory shows that sampled iterates concentrate; it does not imply that every point in a sublevel set lies near the same subspace. Likewise, a thresholded Hessian rank does not imply Eq. (43): curvature eigendirections and low-loss displacement directions need not coincide.

E.4 Exact Norm of the Affine Lift

Lemma 1 (Expansion-operator norm). *Let $E \in \mathbb{R}^{P_{\text{w}} \times d}$ represent the input-shared cyclic lift and let m_j be the number of selected weight entries assigned to coordinate j. Then*

$$E^\top E = \operatorname{diag}(m_1, \dots, m_d), \qquad \operatorname{Lip}(g) = \beta \|E\|_2 = \beta \sqrt{\max_j m_j}. \tag{64}$$

Proof. Every row of E contains exactly one unit entry because each selected weight entry receives one cyclic coordinate. Distinct columns therefore have disjoint supports, while column j contains m_j ones. This gives the stated Gram matrix and spectral norm. Since $g(\mathbf{z}) - g(\mathbf{z}') = \beta E(\mathbf{z} - \mathbf{z}')$, the upper bound is attained by a top right singular vector of E. □

E.5 Proof of Proposition 3

Proof. Write $V_j = \mathrm{Vol}(B_j(0,1))$. For a fixed orientation $\mathbf{U}$, complete its columns to an orthogonal matrix $[\mathbf{U}, \mathbf{U}_\perp]$. If the chart $\mathbf{a} + \mathbf{U}B_d(s)$ intersects a basin contained in $B_P(\theta_b, r_b)$, then $\mathbf{a} \in B_P(\theta_b, r_b) - \mathbf{U}B_d(s)$. In the orthogonal coordinates induced by $[\mathbf{U}, \mathbf{U}_\perp]$, the parallel component of this Minkowski difference has norm at most $s + r_b$ and the orthogonal component has norm at most r_b. Hence its volume is at most $V_d V_{P-d}(s + r_b)^d r_b^{P-d}$. The offset-density and volume assumptions then give

$$\Pr[(\mathbf{a} + \mathbf{U}B_d(s)) \cap \mathcal{B} \neq \varnothing] \leq \frac{c_0 V_d V_{P-d}}{c_\Omega V_P} \left(\frac{s + r_b}{R}\right)^d \left(\frac{r_b}{R}\right)^{P-d}. \tag{65}$$

This proves Eq. (46) with $C'_{P,d} = c_0 V_d V_{P-d}/(c_\Omega V_P)$. A union bound multiplies this quantity by N_{basins}. If $s \leq R$ and $r_b/R \leq e^{-c}$, the logarithm of that bound is at most $\log N_{\mathrm{basins}} + \log C'_{P,d} + d\log 2 - c(P-d)$, which tends to $-\infty$ at rate at least $\eta c(P-d)$ under the stated condition. □

E.6 Proof of Theorem 2

Proof. On a fixed sample S, equip loss vectors with the normalized empirical metric

$$d_S(\mathbf{z}, \mathbf{z}') = \left(\frac{1}{m}\sum_{i=1}^{m} |\ell(f_{g(\mathbf{z})}(x_i), y_i) - \ell(f_{g(\mathbf{z}')}(x_i), y_i)|^2\right)^{1/2}. \tag{66}$$

If $RL_S = 0$, all restricted loss vectors coincide, so the covering number is one and the empirical Rademacher complexity is zero. For $RL_S > 0$, the assumption gives $d_S(\mathbf{z}, \mathbf{z}') \leq L_S\|\mathbf{z} - \mathbf{z}'\|_2$. An ε/L_S-net of $B_d(0, R)$ therefore induces an ε-net of the restricted loss class, and the Euclidean volumetric bound gives

$$\log \mathcal{N}(\mathcal{F}_R|_S, \varepsilon, d_S) \leq d\log\left(1 + \frac{2RL_S}{\varepsilon}\right). \tag{67}$$

Subtract the fixed loss vector at $\mathbf{z} = 0$; this leaves the expected Rademacher supremum unchanged and gives a class containing zero. The d_S-diameter is at most $2RL_S$, and Dudley's entropy integral yields

$$\widehat{\mathfrak{R}}_S(\mathcal{F}_R) \leq \frac{C_0}{\sqrt{m}} \int_0^{2RL_S} \sqrt{d\log\left(1 + \frac{2RL_S}{\varepsilon}\right)}\, d\varepsilon \tag{68}$$

$$\leq C\frac{RL_S\sqrt{d}}{\sqrt{m}}, \tag{69}$$

because the rescaled integral is finite. The standard bounded-loss Rademacher inequality (Bartlett & Mendelson, 2002) gives the stated uniform generalization bound. Under compatible metrics, composition with a ρ-Lipschitz loss and an L_f-Lipschitz model map gives $L_S \leq \rho L_f \beta \|E\|_2$ by the affine-map norm identity (Lemma 1 gives its cyclic specialization). □

Remark 2 (Dimension is not enough). For the cyclic lift, $\sum_j m_j = P_{\mathrm{w}}$. If the copied coordinates are roughly balanced, then $\sqrt{d}\,\|E\|_2 \approx \sqrt{P_{\mathrm{w}}}$. Thus the reduced-class Rademacher expression cannot be compared with a full-parameter ball by dimension alone; the radius, metric, expansion norm, and transductive nature of Planetoid evaluation all matter.

E.7 Proof of Theorem 3

Proof. The affine map has constant Jacobian $\mathbf{J}_g = \beta E$ and zero second derivative. The chain rule gives the exact pullback Hessian

$$\nabla^2_{\mathbf{z}} \widetilde{\mathcal{L}}(\mathbf{z}) = \beta^2 E^\top \nabla^2_\theta \mathcal{L}(\theta^* + \beta E\mathbf{z})E. \tag{70}$$

Let $\mathcal{Z}$ be the convex region in coordinate space from the theorem. Because g is affine, the line segment between any two points of $\mathcal{Z}$ maps into $g(\mathcal{Z})$. If the full-parameter Hessian norm is at most L_θ

throughout $g(\mathcal{Z})$, then $\|\nabla^2_{\mathbf{z}}\widetilde{\mathcal{L}}(\mathbf{z})\|_2 \le \beta^2\|E\|_2^2 L_\theta$ on every such segment. The mean-value formula for the gradient therefore gives the valid smoothness constant $\bar{L}_z = \beta^2\|E\|_2^2 L_\theta$ on $\mathcal{Z}$. Smoothness gives

$$\widetilde{\mathcal{L}}(\mathbf{z}_{t+1}) \le \widetilde{\mathcal{L}}(\mathbf{z}_t) - \frac{1}{2\bar{L}_z}\|\nabla\widetilde{\mathcal{L}}(\mathbf{z}_t)\|_2^2 \tag{71}$$

for a step of size $1/\bar{L}_z$. The PL inequality therefore implies

$$\widetilde{\mathcal{L}}(\mathbf{z}_{t+1}) - \widetilde{\mathcal{L}}^* \le \left(1 - \frac{\mu_z}{\bar{L}_z}\right)(\widetilde{\mathcal{L}}(\mathbf{z}_t) - \widetilde{\mathcal{L}}^*). \tag{72}$$

Iteration proves the corresponding bound with $\bar{L}_z$; when a tighter valid L_z is known, Eq. (53) follows by using that constant. All iterates must remain in the assumed region. □

F Additional Experimental Results

F.1 Complete Functional-Alignment Diagnostics

Table 18: Dataset-equal diagnostics for all five $d = 8$ basis families. Oracle R_0^2 projects the reference network's displacement in weight space before evaluating its nonlinear validation-logit change. Linear and nonlinear R_0^2 fit coefficients on D_{fit} and evaluate on validation nodes. Validation accuracy comes from separate task-loss optimization, not the reconstruction coefficients. Dense aligned bases require $P_{\text{w}}d$ frozen values; this comparison matches coordinate count and radius, not storage.

Basis	Weight capture	Oracle R_0^2	Linear R_0^2	Nonlinear R_0^2	Grad. capture	Val. acc. (%)
Random	0.0001	0.000	0.077	0.078	0.00012	60.6
Cyclic	0.0792	0.182	0.219	0.221	0.00575	60.9
Gradient	0.0367	0.509	0.624	0.621	0.28024	62.3
Displacement	0.4308	0.379	0.599	0.605	0.03538	62.2
Trajectory	0.4333	0.391	0.625	0.633	0.03748	62.2

The aligned families reveal why weight-space concentration is not sufficient. At the dataset-equal level, the gradient basis captures only 0.037 of the reference network's weight energy, compared with 0.431 for displacement and 0.433 for trajectory, yet their nonlinear held-out functional R_0^2 values are similar (0.621, 0.605, and 0.633). The gradient basis instead captures 0.280 of the gradient computed on D_{fit}, versus 0.035–0.037 for the displacement families. These quantities distinguish alignment with a task derivative from proximity to one realized optimizer displacement.

The nominal value $d = 8$ also overstates the balanced functional degrees of freedom of the cyclic map. Its dataset-equal participation rank is 1.195, versus 6.345 for random and 4.218 for gradient alignment. Effective rank is interpreted jointly with sensitivity: random directions span more balanced but very weak changes, whereas the gradient basis combines higher participation rank with sensitivity 1.590. Finite-difference halving errors are below 0.021 across all 60 folds.

The common radius is an active trust region. The unrestricted D_{basis} probe exceeds the cap in every fold, so $\rho = 0.05\|\theta^*_{\text{w}}\|_2$ throughout; every projected reference network and 524/540 mapped optimizations finish numerically on that radius. Cross-fit fold differences remain substantial for some anchors, which is why inference first averages the two folds and never treats them as independent samples. The run is transductive: DGI and message passing use the full graph's features and edges, but validation labels are not used for the anchor, head, basis, radius, coefficient fit, optimizer, or method choice, and official test data are never evaluated.

Dense gradient, displacement, and trajectory bases store $P_{\text{w}}d$ frozen values: 766,464 on Cora, 1,928,704 on CiteSeer, and 288,768 on PubMed (approximately 3.07, 7.71, and 1.16 MB in float32). The compact study in Section C.9 tests how much of this reconstruction quality survives factorization.

F.2 Dimension, depth, and architecture diagnostics

We examine dimension, depth, architecture, and modulation scale around task-supervised anchors; Section C.19 summarizes the results.

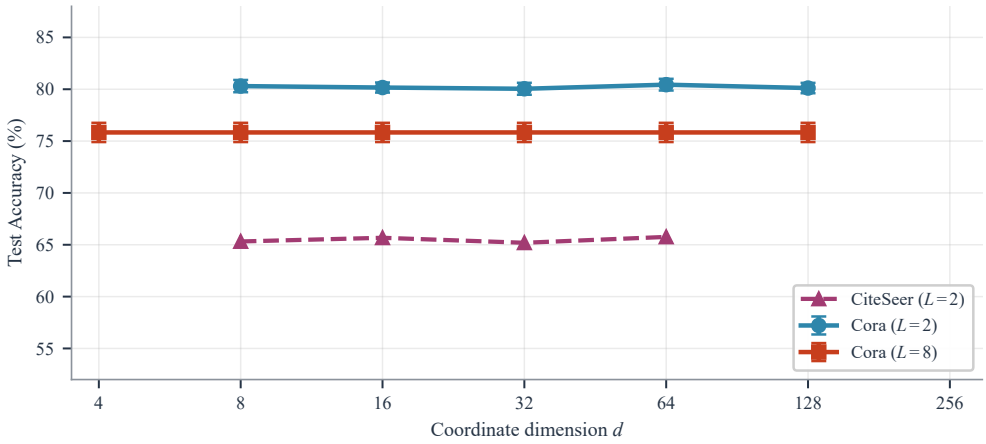

Figure 7: Observed dimension sweeps around task-supervised anchors at $k = 1$. Curves show mean accuracy; Cora error bars show one standard deviation. The nearly flat curves describe these measured settings; no functional rank or sufficient coordinate dimension is estimated.

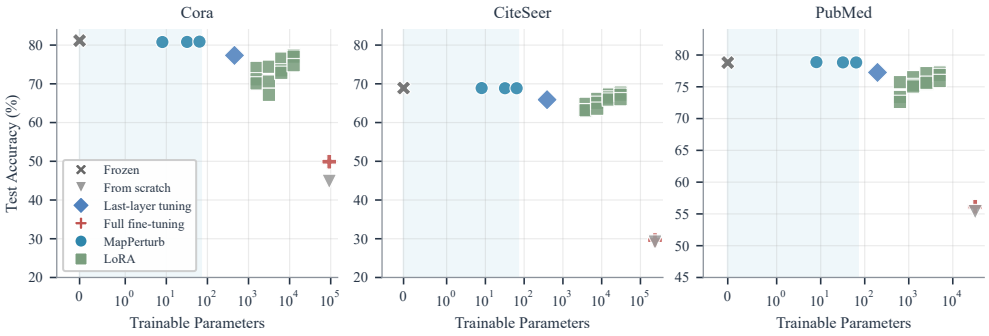

Figure 8: Trainable parameter count and mean test accuracy with task-supervised anchors ($k = 1$, ten paired seeds). Symbols distinguish method families; shaded regions highlight the small-coordinate settings. The count excludes extra frozen adapter storage. The horizontal axis is symmetric-log with a linear region up to one parameter, allowing the zero-parameter Frozen point. Dataset panels use different vertical ranges.

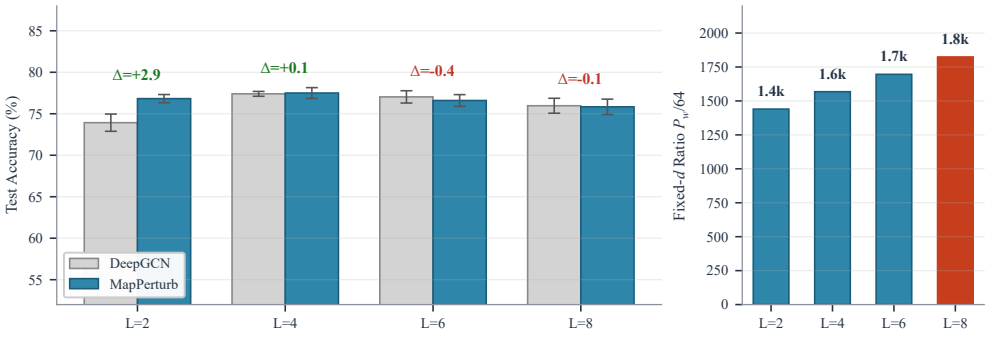

Figure 9: Depth diagnostic on Cora with task-supervised deep anchors. Left: bars show mean accuracy with one-standard-deviation error bars; annotations give MapPerturb-minus-DeepGCN changes in percentage points. Right: mapped weights per trainable coordinate, $P_w/64$, at each depth. This coordinate ratio does not measure stored model compression.

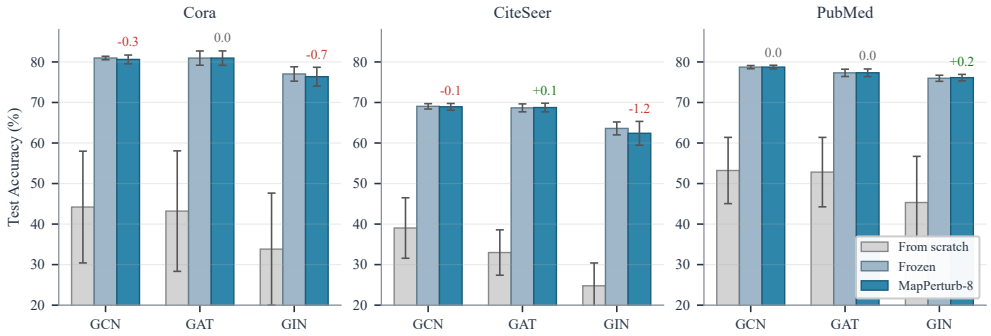

Figure 10: Cross-architecture supervised-anchor diagnostic on Cora, CiteSeer, and PubMed ($k = 1$, $d = 8$, ten seeds). Bars show mean accuracy and error bars show one standard deviation for training from scratch, the frozen anchor, and MapPerturb. Labels give MapPerturb-minus-frozen mean changes in percentage points; these are not paired confidence intervals.

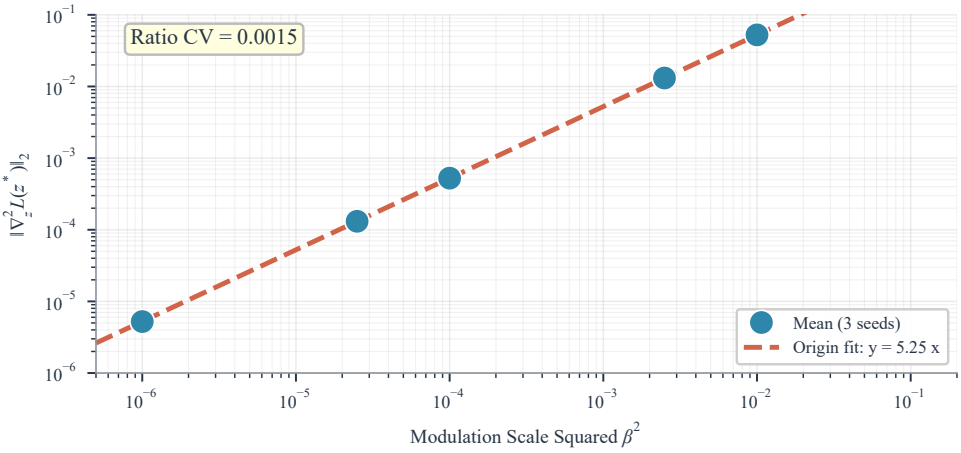

Figure 11: Implementation check for the scaling of the Hessian with respect to $\mathbf{z}$ under modulation scale at a fixed anchor. For $g(z) = \theta^* + \beta E z$, the pullback Hessian is $\beta^2 E^\top \nabla_\theta^2 \mathcal{L} E$. The horizontal axis is β^2, so proportional scaling has log–log slope one. The fitted line is empirical; ratio CV is the coefficient of variation of the mean Hessian norm divided by β^2. This checks the identity, not the PL condition.

Across 72 two-class SBM graphs ($n = 300$, varied within/between-community edge probabilities), spectral gap and MapPerturb-minus-scratch accuracy have a descriptive negative association (Spearman $\rho = -0.409$, $p = 0.0004$). Most accuracy differences lie near zero, with several negative outliers. Difficulty and saturation may confound this exploratory association, which establishes neither causality nor a $(1 - \lambda)^L$ rate.

F.3 Classification across graphs and backbones

The classification comparison covers nine methods on three graphs, three backbones, and ten supplied splits. Chameleon and Squirrel use the cached unfiltered graphs. Each context samples 20 training nodes per class and trains a common width-64 DGI encoder for 200 Adam steps (lr $= 0.01$, weight decay 10^{-4}). A linear head is fitted for 100 steps (lr $= 0.01$, weight decay 5×10^{-4}) and then frozen. The multi-gradient map uses 16 subsets of five labels per class from this same training pool, with rank-two factors per layer and eight shared coordinates. Grad-SVD instead computes one gradient on the entire training pool. For each coordinate $j = 1, \ldots, 8$, it concatenates the jth singular-value-weighted rank-one component from each layer's gradient SVD and normalizes that concatenated direction. Its eight directions are fixed during adaptation. Features and edges are visible transductively; labels outside the pool are hidden during construction and fitting.

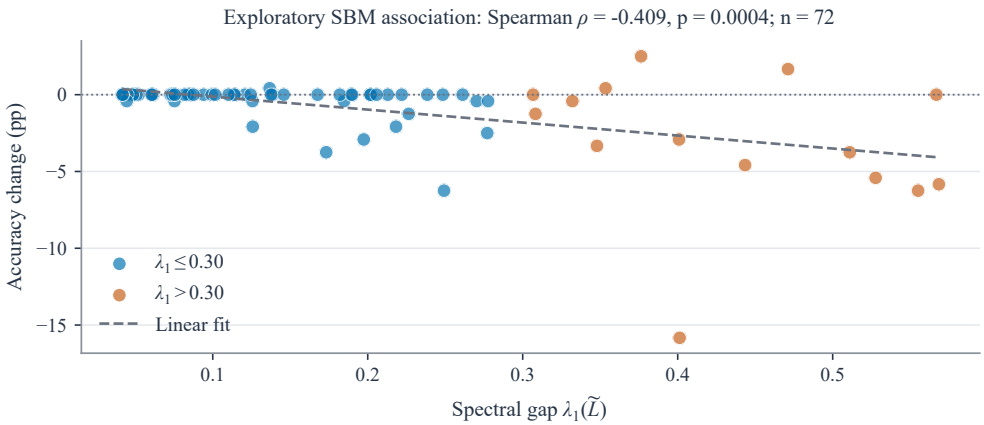

Figure 12: Spectral gap and MapPerturb-minus-scratch accuracy on 72 SBM graphs. Accuracy differences are in percentage points. Blue and orange distinguish $\lambda_1 \leq 0.30$ and $\lambda_1 > 0.30$ for visual reference; the split is not an estimated transition. The dashed line is a linear fit; the title reports the Spearman rank correlation.

All encoder updates, including PEFT and full-encoder controls, share a Frobenius radius set by a 20-step training-label probe and capped at 5% of the mapped anchor norm. Adaptation uses Adam with zero weight decay and 100 steps for each learning rate 0.005 and 0.05. Validation selects among the zero update and checkpoints 25, 50, and 100, ordered by accuracy, loss, earlier step, and lower learning rate. Selected updates are fixed before one test evaluation. Paired 95% t intervals first average the nine graph–backbone cells within each of ten split blocks. These graph families informed earlier development, so this is a fixed comparison on familiar benchmarks. The implementations compare the named PEFT parameterizations under this shared encoder-update protocol.

Full classification results appear in Table 3; the resource counts below use the same implementations.

Table 19: Resource accounting for the Q4 implementations on Chameleon with GCN. Every row shares the anchor, frozen fitted head, training labels, radius, optimizer budget, and validation rule described above. Counts exclude the shared anchor and head. Frozen execution values describe the basis held during adaptation, including materialized buffers; they are not an additional persistent payload to sum with the preceding columns. Regenerable values can be reconstructed from a seed or the anchor.

Method	Trainable	Task-specific frozen	Regenerable	Frozen execution
Frozen fitted head	0	0	0	0
Matched random	8	80,672	0	80,672
Multi-gradient	8	80,672	0	80,672
Grad-SVD	8	20,152	0	1,223,168
GiVA $r = 4$	136	10,068	0	10,068
VeRA $r = 16$	160	0	38,224	38,224
LoRA $r = 1$	2,517	0	0	0
LoRA-XS $r = 2$	8	0	5,034	5,034
Full encoder	152,896	0	0	0

Balanced split-block inference gives multi-gradient minus frozen $+.650$ $[+.377, +.923]$ and minus random $+.581$ $[+.324, +.838]$, with ten wins in each comparison. Multi-gradient minus Grad-SVD is $-.538$ $[-.908, -.169]$ with nine losses; minus GiVA is $-.273$ $[-.631, +.086]$. Grad-SVD minus GiVA is $+.265$ $[+.054, +.477]$. Grad-SVD therefore gives the strongest overall result in this comparison.

The ogbn-arxiv evaluation uses the same 20-label-per-class, frozen-head, radius, and validation-selection rules with five GCN anchors (202712000–202712004) and seven methods. Intervals pair the five anchors. Means are 24.806 (frozen), 25.638 (random), 27.892 (multi-gradient), 28.482 (Grad-SVD), 29.080 (GiVA), 30.838 (LoRA), and 31.347 (full encoder). Multi-gradient minus frozen is $+3.086$ points and minus random is $+2.254$, with four wins and one tie, but five-pair intervals

cross zero. The study demonstrates execution on the larger graph under its official chronological split; the accuracy gains remain uncertain.

The earlier pilots in Appendix C.19 are separate from this fixed external comparison.

F.4 Class-balanced evaluation of Q3

Test-class sizes vary by factors of 4.98, 3.00, and 2.29 on Cora, CiteSeer, and PubMed. We complement accuracy with macro-F1 and balanced accuracy, the unweighted means of per-class F1 and recall, respectively (undefined F1 is zero).

We rerun all ten methods under the fixed protocol in Appendix C.11, sharing anchors, supports, and initial heads within each configuration. This separate cohort uses refitted anchors and the earlier support sets; all metrics come from its 900 terminal predictions, without metric-based tuning or checkpoint selection. Paired inference averages shots within 30 dataset–anchor units.

Table 20: Class-balanced evaluation in the supplementary Q3 cohort: (a) mean scores (%); (b) paired differences (pp) with nominal 95% t intervals over 30 dataset–anchor units. All 12 core contrasts have Holm-adjusted $p < 10^{-5}$ across metrics. The three metrics share predictions.

(a) All methods

Method	Accuracy	Macro-F1	Balanced accuracy
Head only	52.76	50.25	52.54
DGI matched random	52.81	50.34	52.59
DGI factors	55.72	53.58	55.81
Masked matched random	52.79	50.32	52.58
Masked factors	56.21	54.12	56.28
LoRA-$r = 1$	56.83	54.45	57.17
LoRA-$r = 2$	57.05	54.69	57.52
LoRA-$r = 4$	58.40	55.99	58.77
LoRA-$r = 8$	58.14	55.68	58.65
Full encoder	60.01	57.08	60.33

(b) Core paired contrasts

Contrast	Accuracy	Macro-F1	Balanced accuracy
DGI − head only	+2.95 [1.87, 4.04]	+3.33 [2.22, 4.43]	+3.27 [2.21, 4.33]
Masked − head only	+3.45 [2.34, 4.56]	+3.87 [2.71, 5.02]	+3.74 [2.62, 4.86]
DGI − matched random	+2.91 [1.85, 3.97]	+3.24 [2.16, 4.33]	+3.22 [2.17, 4.27]
Masked − matched random	+3.42 [2.33, 4.51]	+3.80 [2.67, 4.94]	+3.70 [2.59, 4.80]

DGI and masked factors improve macro-F1 over head-only by 3.33 and 3.87 points. Gains persist against matched random controls and under balanced accuracy (Table 20). All three metrics give the same pooled method ordering, with full encoder tuning highest.

Table 21: Macro-F1 (%) in the same supplementary cohort. Graph columns average ten anchors and three shot levels; shot columns average 30 dataset–anchor units.

	By graph			Labels per class		
Method	Cora	CiteSeer	PubMed	1	5	10
Head only	53.08	39.88	57.80	38.88	54.20	57.68
DGI matched random	53.16	39.97	57.89	38.91	54.35	57.76
DGI factors	57.15	43.66	59.93	41.25	58.10	61.39
Masked matched random	53.16	40.02	57.77	38.91	54.32	57.72
Masked factors	58.33	45.36	58.66	41.64	59.03	61.68
LoRA-$r = 1$	60.82	42.31	60.22	39.63	59.96	63.77
LoRA-$r = 2$	60.03	42.77	61.26	38.89	60.15	65.02
LoRA-$r = 4$	61.45	46.07	60.45	39.84	62.20	65.92
LoRA-$r = 8$	59.76	46.32	60.97	37.93	62.01	67.12
Full encoder	62.32	46.10	62.83	37.49	64.48	69.28

Factor maps have higher mean macro-F1 than LoRA at one label per class; LoRA scores higher at five and ten (Table 21). The exploratory one-shot contrasts do not establish superiority after

Holm correction across 45 external comparisons; masked minus LoRA-$r = 4$ gives $+1.80$ points ($[-1.62, 5.23]$).

PubMed macro-F1 gains over head-only are $+2.12$ ($[-0.23, 4.47]$) for DGI and $+0.86$ ($[-0.64, 2.36]$) for masked factors. Cora class 5 has negative mean recall changes (-2.88 and -1.26 points; both intervals include zero). Pooled gains therefore do not imply uniform class or graph improvements. Complete metric and class breakdowns accompany the saved predictions.