Beyond Pairwise: Diagnosing Higher-Order Merge Failures via Hodge Decomposition

Dongzhe Zheng [1] **Christine Allen-Blanchette** [1]

Abstract

Model merging typically chooses a reference model, aligns all other models to it, and interpolates their weights. This star-shaped pipeline treats mergeability as pairwise: if each model can be merged with the reference, then larger collections are assumed to be mergeable as well. We show that this assumption can fail even in a controlled setting. For 24 weight-matched PlainCNNs trained on CIFAR-10, more than 90% of triples whose pairs are individually mergeable are not jointly mergeable. We diagnose this failure by representing mergeability as a simplicial complex and decomposing merge barriers into gradient, curl, and harmonic components. These components capture reference-dependent effects, local inconsistencies around triples, and global obstructions that persist under fixed alignment. The mergeability complex exposes substantial higher-order obstruction: its first Betti number, counting cycles of pairwise-mergeable models that do not bound jointly mergeable triples, reaches 122. Under standard star-topology alignment, 25.5% of barrier energy remains harmonic, a component not directly targeted by existing merging methods. Re-aligning models along a minimum spanning tree weighted by harmonic energy reduces this fraction to 1.8% and decreases the first Betti number from 75 to 5, though the total merge barrier increases by 15%. Thus, alignment topology is not merely an implementation detail: it determines which merge barriers are visible, removable, or hidden as global obstructions.

[1]Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ, USA. Correspondence to: Dongzhe Zheng <dz5992@princeton.edu>, Christine Allen-Blanchette <ca15@princeton.edu>.

Workshop on Weight-Space Symmetries, held in conjunction with the 43rd International Conference on Machine Learning, Seoul, South Korea. 2026. Copyright 2026 by the author(s).

1. Introduction

Model merging methods such as model soups (Wortsman et al., 2022), federated averaging (McMahan et al., 2017), and task arithmetic (Ilharco et al., 2023) combine neural networks directly in weight space, requiring compatible regions of the loss landscape. A common geometric explanation is *mode connectivity*: after permutation alignment (Ainsworth et al., 2023; Entezari et al., 2022), independently trained networks can often be linearly interpolated without performance degradation (Frankle et al., 2020).

Most analyses of this compatibility are pairwise. Yet pairwise mergeability need not imply joint mergeability: three models may have flat interpolation paths along every edge while the loss rises inside their convex hull. We observe this failure even in a controlled setting: among 24 weight-matched PlainCNNs trained on CIFAR-10, over 90% of pairwise-mergeable triples fail as three-way merges, and the resulting mergeability complex reaches first Betti number $\beta_1 = 122$. The standard response to merge barriers is better alignment. Methods such as Git Re-Basin (Ainsworth et al., 2023) typically use a *star topology*, aligning every model to a single reference. We show that this topology shapes not only barrier size, but barrier type. Hodge decomposition separates pairwise merge barriers into gradient, curl, and harmonic components, corresponding to per-model effects, triangle-local inconsistencies, and global obstructions. Under star alignment, 25.5% of barrier energy remains harmonic, a component that cannot be removed by reweighting merges within the fixed alignment.

To test whether this harmonic residual depends on alignment topology, we construct a minimum spanning tree (MST) weighted by harmonic energy and re-align models along it. Harmonic energy drops by 93% ($25.5\% \rightarrow 1.8\%$), and β_1 collapses from 75 to 5. Although the total barrier rises by 15%, the increase shifts into gradient and curl components, which are local obstructions that merging strategies can reduce. Thus, changing alignment topology can convert global obstructions into more accessible local ones.

Contributions.

1. We introduce the mergeability complex $\mathcal{K}_\tau$ and use

Hodge decomposition to separate merge barriers into gradient, curl, and harmonic components.

2. On 24 CIFAR-10 PlainCNNs, we show that over 90% of pairwise-mergeable triples fail as three-way merges, with β_1 reaching 122; initialization strongly shapes this topology.
3. We propose Hodge-guided MST alignment, which reduces harmonic energy by 93% and collapses β_1 from 75 to 5 while shifting barrier mass into gradient and curl components, despite a 15% increase in total barrier.

2. Mergeability Complex and Hodge Analysis

2.1. Merge Barrier and Mergeability Complex

Let $\mathcal{M} = \{\theta_1, \ldots, \theta_n\}$ be a set of aligned models and let $\mathcal{L}$ denote the task loss, cross-entropy in our experiments. For a subset $\sigma \subseteq \mathcal{M}$, define its *merge barrier*

$$b(\sigma) = \sup_{\alpha \in \Delta^{|\sigma|-1}} \mathcal{L}\big(\textstyle\sum_{i\in\sigma} \alpha_i \theta_i\big) - \min_{i\in\sigma} \mathcal{L}(\theta_i), \qquad (1)$$

where $\Delta^{|\sigma|-1}$ is the probability simplex. This measures the worst loss increase over convex merges of σ, relative to the best endpoint. The sublevel set $\mathcal{K}_\tau = \{\sigma : b(\sigma) \le \tau\}$ is an abstract simplicial complex: for $\sigma' \subseteq \sigma$, the supremum can only decrease (the face $\Delta^{|\sigma'|-1}$ embeds in $\Delta^{|\sigma|-1}$) and $\min_{\sigma'} \mathcal{L} \ge \min_\sigma \mathcal{L}$, so $b(\sigma') \le b(\sigma)$. Thus every subset of a mergeable set is mergeable.

Importantly, $\mathcal{K}_\tau$ is *not* a Vietoris–Rips complex: a 3-clique of low-barrier pairs need not form a 2-simplex. These unfilled triangles are the higher-order merge failures we aim to diagnose.

2.2. Hodge Decomposition

Hodge decomposition separates an edge signal into three kinds of structure. In our setting, the signal is the pairwise merge barrier on mergeable edges. The *gradient* component captures barriers explained by per-model potentials: some models are globally easier or harder to merge. The *curl* component captures local inconsistency around filled triangles, where all three models are jointly mergeable but the three pairwise barriers do not agree with any scalar potential. The *harmonic* component captures circulation around unfilled cycles: collections of pairwise-mergeable models whose higher-order interiors are not jointly mergeable. Thus, harmonic energy is the part of the barrier signal tied to global merge obstructions rather than endpoint effects or triangle-local inconsistencies.

Let V, E, and T denote the vertex, edge, and triangle sets of $\mathcal{K}_\tau$, respectively. The boundary operators $B_1 \in \mathbb{R}^{|V|\times|E|}$ and $B_2 \in \mathbb{R}^{|E|\times|T|}$ define the Hodge Laplacian $\Delta_1 = B_1^\top B_1 + B_2 B_2^\top$ on edges. Its kernel dimension $\beta_1 = \dim\ker\Delta_1$ counts independent edge-cycles. The 1-cochain space $C^1 = \mathbb{R}^{|E|}$ decomposes orthogonally (Lim, 2020):

$$C^1 = \underbrace{\mathrm{im}(B_1^\top)}_{\text{gradient}} \oplus \underbrace{\mathrm{im}(B_2)}_{\text{curl}} \oplus \underbrace{\ker(\Delta_1)}_{\text{harmonic}}. \qquad (2)$$

Assigning each mergeable edge $e = (i, j)$ its pairwise barrier value $b(\{i, j\})$ yields an edge signal $\mathbf{b} \in C^1$, which decomposes as $\mathbf{b} = \mathbf{b}_{\text{grad}} + \mathbf{b}_{\text{curl}} + \mathbf{b}_{\text{harm}}$. Since the three subspaces are mutually orthogonal, $\|\mathbf{b}\|^2 = \|\mathbf{b}_{\text{grad}}\|^2 + \|\mathbf{b}_{\text{curl}}\|^2 + \|\mathbf{b}_{\text{harm}}\|^2$. We report the corresponding energy fractions and define the per-edge harmonic energy as $h_e^2 = (\mathbf{b}_{\text{harm}})_e^2$.

Formally, $\mathbf{b}_{\text{grad}} = B_1^\top \mathbf{s}$ for a scalar merge potential $\mathbf{s}$ on models, and $\mathbf{b}_{\text{curl}} = B_2 \mathbf{r}$ for a triangle cochain $\mathbf{r}$. The harmonic component, $\mathbf{b}_{\text{harm}}$ lies in $\ker\Delta_1$; within a fixed alignment and complex, it cannot be explained by model potentials or local triangle effects. We fix the reference orientation (i, j) with $i < j$; component energies, β_1, and $\mathbf{s}$ up to a global constant are invariant to this choice; see Section E.

2.3. Hodge-Guided Alignment

Standard alignment uses a *star topology*: every model is permutation-aligned to a single reference θ_0 by layer-wise weight matching, using the weight-matching variant of Git Re-Basin (Ainsworth et al., 2023). Thus each pair (i, j) is implicitly compared through θ_0. We hold the pairwise aligner fixed and vary only the topology through which alignments are composed; stronger aligners may change absolute barriers, but the star-versus-MST contrast is measured under a single fixed aligner.

We replace the star with a minimum spanning tree weighted by per-edge harmonic energy. Starting from the star alignment, we compute pairwise barriers, build $\mathcal{K}_\tau$, and Hodge-decompose the edge signal to obtain h_e^2. We then construct the MST with weights $H_{ij} = h^2_{(i,j)}$, root it at the model with smallest total incident harmonic energy $\sum_j h^2_{(i,j)}$, and re-align models from the raw checkpoints along the rooted tree using BFS. This routes each alignment through harmonically compatible neighbors.

3. Experiments

3.1. Setup

We train 24 PlainCNNs ($\sim$1.2M parameters) on CIFAR-10 from eight random seeds and three learning rates $\{0.01, 0.05, 0.1\}$ using SGD with momentum, weight decay 10^{-4}, cosine annealing, and 100 epochs.[1] Test accuracies

[1] Code and trained checkpoints are available at https://github.com/ContinuumCoder/

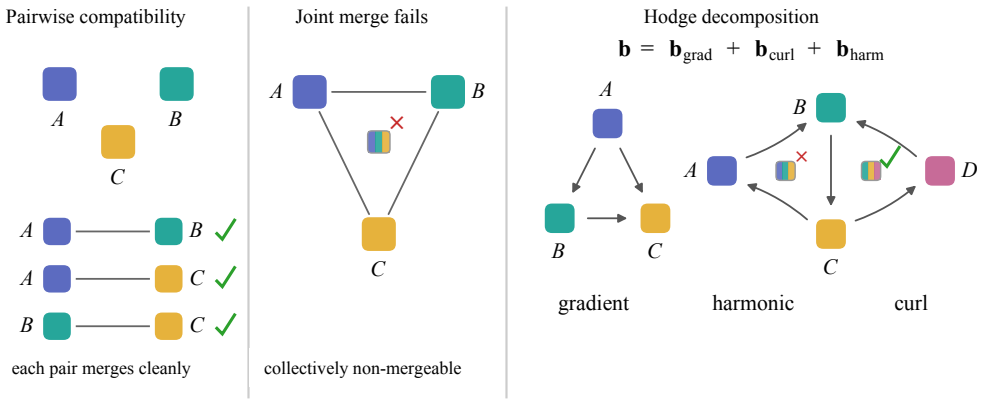

Figure 1. **Higher-order merge failure and its Hodge signature.** Models A, B, C may merge cleanly pairwise while their joint three-way merge fails. Hodge decomposition separates the edge-barrier signal into gradient, curl, and harmonic components. Gradient barriers follow per-model potentials; curl barriers circulate around filled, jointly mergeable triangles; harmonic barriers circulate around unfilled cycles, giving topological obstructions counted by β_1.

range from 87.8% to 90.3%. The architecture has no skip connections, so layer-wise Hungarian weight matching following Git Re-Basin (Ainsworth et al., 2023) is exact.

We compute all $\binom{24}{2} = 276$ pairwise barriers and approximate triplet barriers on simplex grids with batch normalization recalibrated at every interpolation point. Triplets are evaluated for the 884 3-cliques at the 75th-percentile cutoff, while the filtration analysis sweeps τ. Full architecture, training, grid, and enumeration details are in Section A.

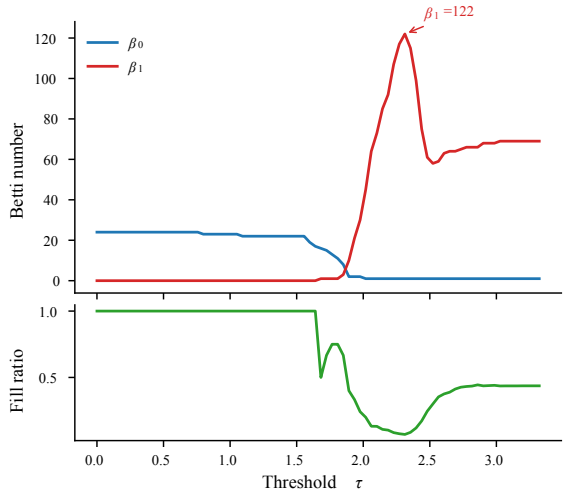

Figure 2. Filtration over τ. Top: β_0, β_1; bottom green: fill ratio (jointly/pairwise mergeable triples). Fill ratio < 0.1 at peak $\beta_1 = 122$, showing higher-order merge failure.

Table 1. Hodge energy decomposition across scales. Gradient and harmonic energy remain substantial; curl spikes when triangles enter the complex.

τ	Edges	Tri	β_1	Grad%	Curl%	Harm%
1.99	69	16	31	79.8	2.1	18.1
2.11	138	40	75	71.4	3.1	25.5
2.24	193	60	110	69.1	3.7	27.3
2.31	221	73	125	69.9	3.7	26.3
2.38	234	131	106	69.3	12.5	18.3

3.2. Mergeability Complex Structure

Figure 2 shows the filtration of $\mathcal{K}_\tau$. The first Betti number peaks at $\beta_1 = 122$ near $\tau \approx 2.31$: at this scale, the complex contains 221 mergeable edges but only 76 filled triangles, so 93% of pairwise-mergeable triples fail as three-way merges. As τ increases, triangles fill in and β_1 declines, but persistent unfilled cycles remain.

3.3. Hodge Decomposition and Node Potential

At $\tau = 2.11$, the median pairwise barrier, Hodge decomposition yields 71.4% gradient, 3.1% curl, and 25.5% harmonic energy. The gradient–harmonic balance is stable across scales, while curl increases when many triangles enter the complex (Table 1); high-threshold behavior is summarized in Section C.

The gradient component $\mathbf{b}_{\text{grad}} = B_1^\top \mathbf{s}$ defines a merge potential $\mathbf{s}$ on models. This potential aligns strongly with training seed (Spearman $\rho = 0.96$, $p < 10^{-12}$), but only weakly with learning rate ($\rho = 0.13$, $p = 0.56$) or test accuracy ($\rho = 0.26$, $p = 0.23$). As shown in Figure 3, models cluster by seed along the potential axis, regardless of learning rate. Together with the per-seed harmonic fractions

`Higher-Order-Model-Merge`.

in Figure 4, this suggests that initialization strongly shapes the topology of $\mathcal{K}_\tau$.

3.4. Anomaly Detection via Harmonic Fraction

Per-model harmonic fraction separates outliers that have similar mean pairwise barriers but different obstruction types (Figure 4). Seed 0 models have low harmonic fraction (10–15%) and much lower within-seed than cross-seed barriers ($\sim$1.0 vs. $\sim$2.0), indicating that their incompatibility is mostly gradient-like. In contrast, seeds 3–5 have high harmonic fraction (40–48%) and similar within- and cross-seed barriers, indicating incompatibility that remains global under the alignment family considered here. Thus, models with similar mean pairwise barrier ($\approx$2.1) can be distinguished by harmonic fraction; see Section D for further interpretation.

3.5. Hodge-Guided MST Alignment

We apply the procedure from Section 2.3. Figure 5 and Table 2 compare star and MST alignment.

MST alignment reduces harmonic energy by 93% and collapses β_1 from 75 to 5, showing that most harmonic obstruction in this setting depends on the star topology. The tradeoff is a 15% increase in mean barrier, but this increase is concentrated in curl rather than harmonic energy. This distinction matters: reweighting within a fixed alignment cannot remove harmonic energy, whereas curl is locally supported and, as Section 3.6 shows, reducible by post-merge optimization. The five cycles that persist under MST may reflect deeper incompatibilities or limitations of the alignment family. Computational cost is summarized in Section F.

3.6. Downstream Validation: Post-Merge Fine-Tuning

The shift from harmonic to curl energy in Table 2 suggests a practical consequence: once global obstruction is converted into local inconsistency, post-merge optimization should be more effective. We test this on three model pairs spanning the regimes in Figure 4: two same-seed pairs with low star-aligned harmonic fraction, where the predicted gain is small, and one cross-seed pair with high harmonic fraction, where the predicted gain is large.

For each pair, we merge at $\alpha = 0.5$ under star and MST alignment, recalibrate batch normalization on 10,000 training samples, and fine-tune for 10 epochs using SGD with $\eta = 0.005$, cosine schedule, and weight decay 10^{-4}. Table 3 reports CIFAR-10 test accuracy after fine-tuning.

MST-aligned merges outperform star-aligned merges in all three cases, with a mean gain of $+0.74$ points. The largest gain occurs for the cross-seed pair (s_0–s_2, $+1.84$ points), where star alignment has the strongest harmonic obstruction. The gains ($+0.20$, $+0.17$, $+1.84$) follow the ordering predicted by the star-aligned harmonic fractions, supporting the claim that MST alignment makes downstream repair easier

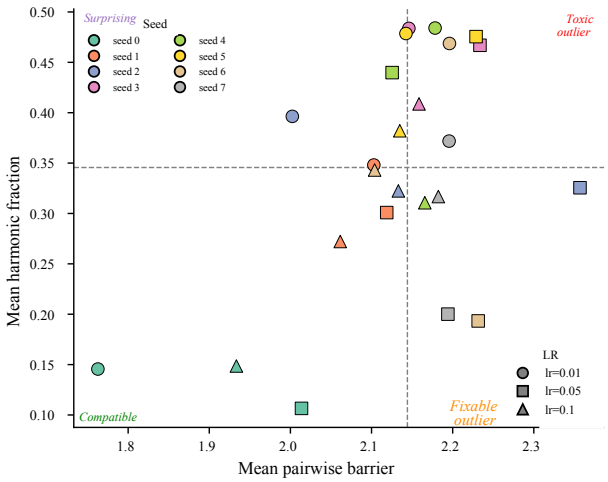

Figure 4. Model anomaly profile: mean barrier versus harmonic fraction, colored by seed. Harmonic fraction separates low-harmonic and high-harmonic outliers that have similar pairwise barriers.

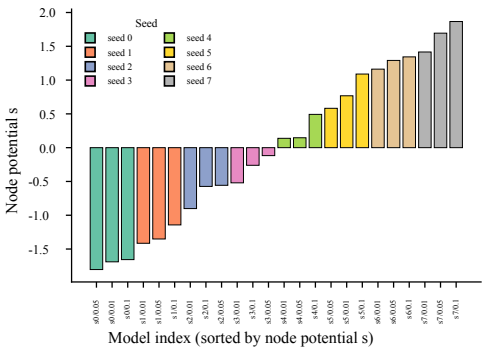

Figure 3. Merge potential **s** from the Hodge gradient component, sorted by value and colored by seed. Models cluster tightly by seed (Spearman $\rho = 0.96$).

Table 2. Star vs. MST alignment at $\tau = 2.11$. Bold marks the row-wise preferred value: lower for barriers, β_1, and harmonic energy; higher for gradient and curl energy. MST sharply reduces harmonic energy and β_1, while increasing total barrier.

Metric	Star	MST	Δ
Barrier mean	**2.13**	2.45	+15%
Barrier max	**3.02**	3.40	+13%
β_1	75	**5**	−93%
Gradient %	**71.4**	66.7	−4.7
Curl %	3.1	**31.5**	+28.4
Harmonic %	25.5	**1.8**	−23.7

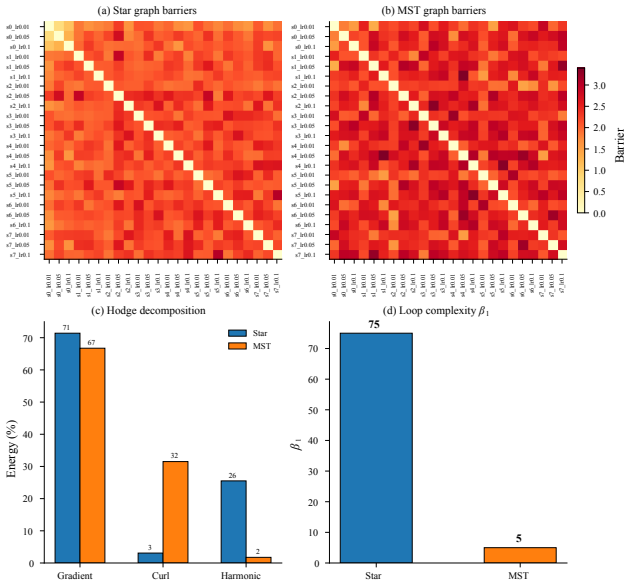

Figure 5. Star vs. MST alignment. **(a,b)**: Barrier matrices (same colorscale). MST barriers are slightly higher but more uniform. **(c)**: Hodge energy redistribution; harmonic drops from 25.5% to 1.8%, largely replaced by curl. **(d)**: β_1 drops from 75 to 5.

Table 3. Post-merge fine-tuning: star vs. MST alignment. Entries are CIFAR-10 test accuracy (%) after 10 epochs; bold marks the higher accuracy in each row. MST improves all three pairs, with the largest gain on the high-harmonic cross-seed pair.

Pair	**Star**	**MST**	Δ
s_0(lr.05) + s_0(lr.1)	87.42	**87.62**	+0.20
s_1(lr.05) + s_1(lr.1)	87.53	**87.70**	+0.17
s_0(lr.05) + s_2(lr.05)	85.62	**87.46**	+1.84
Mean	86.86	**87.59**	+0.74

by shifting obstruction from harmonic to curl components.

4. Related Work

Linear mode connectivity (Frankle et al., 2020) and permutation symmetries (Entezari et al., 2022) explain why independently trained networks can sometimes be connected by low-loss linear paths after alignment, motivating methods such as Git Re-Basin (Ainsworth et al., 2023). Model soups (Wortsman et al., 2022), REPAIR (Jordan et al., 2023), task arithmetic (Ilharco et al., 2023), and TIES-Merging (Yadav et al., 2023) then operate on already-aligned models. In contrast, we study when pairwise alignment fails to extend to higher-order mergeability, and how the topology of the alignment graph changes the resulting obstruction.

Our analysis uses the combinatorial Hodge Laplacian (Lim, 2020) and ideas from topological signal processing (Barbarossa & Sardellitti, 2020; Schaub et al., 2021). Related work on higher-order network interactions (Benson et al., 2016; Battiston et al., 2020) motivates moving beyond pairwise structure, while persistent homology has been applied to neural representations (Rieck et al., 2019) and loss landscapes (Draxler et al., 2018). We differ by applying these tools to model merging, where missing simplices correspond directly to failed higher-order merges.

5. Discussion

Each Hodge component suggests a different mitigation: gradient calls for potential-based grouping, curl for local per-triple correction, and harmonic for optimizing the alignment topology itself (which reduced harmonic energy by 93% in our setting).

The large curl transfer under MST alignment (3.1%→31.5%) reflects the longer, less symmetric paths of the tree, which create triangle-local circulations that the curl component captures; a concrete follow-up is to localize this curl energy to specific triples and correct them individually.

The gradient-derived merge potential $\mathbf{s}$ correlates with seed ($\rho = 0.96$), and the per-seed harmonic fraction separates fixable from toxic outliers (Figure 4). Initialization appears to shape the higher-order mergeability structure in our setting: which 1-cycles appear, which triangles fill, and how much harmonic energy remains after alignment. If this association holds more broadly, initialization-aware grouping could complement alignment-topology optimization.

Beyond merging proper, the same diagnostic applies to *MoE expert compression*: the harmonic fraction of pairwise barriers flags which expert groups will lose capability regardless of merge weights.

Limitations. We study 24 small CNNs on CIFAR-10. Post-merge fine-tuning (Section 3.6) covers three model pairs, with the substantial gain concentrated in the single cross-seed pair, matching the harmonic-energy prediction; a statistically powered evaluation, and the extension to k-way soups and task arithmetic, remain open. Open questions include whether the component balance and the seed–potential correlation persist at larger architectures (e.g., ViTs, LLMs).

Impact Statement

Model merging supports federated learning, community model soups, and on-device personalization. Our diagnostic flags harmonic obstructions so practitioners do not spend compute masking degraded merges in sensitive deployments.

References

Ainsworth, S. K., Hayase, J., and Srinivasa, S. Git re-basin: Merging models modulo permutation symmetries. In *International Conference on Learning Representations*, 2023.

Barbarossa, S. and Sardellitti, S. Topological signal processing over simplicial complexes. *IEEE Transactions on Signal Processing*, 68:2992–3007, 2020.

Battiston, F., Cencetti, G., Iacopini, I., Latora, V., Lucas, M., Patania, A., Young, J.-G., and Petri, G. Networks beyond pairwise interactions: structure and dynamics. *Physics Reports*, 874:1–92, 2020.

Benson, A. R., Gleich, D. F., and Leskovec, J. Higher-order organization of complex networks. *Science*, 353(6295): 163–166, 2016.

Draxler, F., Veschgini, K., Salmhofer, M., and Hamprecht, F. Essentially no barriers in neural network energy landscape. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *PMLR*, pp. 1309–1318, 2018.

Entezari, R., Sedghi, H., Saukh, O., and Neyshabur, B. The role of permutation invariance in linear mode connectivity of neural networks. In *International Conference on Learning Representations*, 2022.

Frankle, J., Dziugaite, G. K., Roy, D. M., and Carbin, M. Linear mode connectivity and the lottery ticket hypothesis. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *PMLR*, pp. 3259–3269, 2020.

Ilharco, G., Ribeiro, M. T., Wortsman, M., Gururangan, S., Schmidt, L., Hajishirzi, H., and Farhadi, A. Editing models with task arithmetic. In *International Conference on Learning Representations*, 2023.

Jordan, K., Sedghi, H., Saukh, O., Entezari, R., and Neyshabur, B. REPAIR: REnormalizing permuted activations for interpolation repair. In *International Conference on Learning Representations*, 2023.

Lim, L.-H. Hodge Laplacians on graphs. *SIAM Review*, 62 (3):685–715, 2020.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and Arcas, B. A. y. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *PMLR*, pp. 1273–1282, 2017.

Rieck, B., Togninalli, M., Bock, C., Moor, M., Horn, M., Gumbsch, T., and Borgwardt, K. Neural persistence: A complexity measure for deep neural networks using algebraic topology. In *International Conference on Learning Representations*, 2019.

Schaub, M. T., Zhu, Y., Seby, J.-B., Roddenberry, T. M., and Segarra, S. Signal processing on higher-order networks: Livin' on the edge... and beyond. *Signal Processing*, 187: 108149, 2021.

Wortsman, M., Ilharco, G., Gadre, S. Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A. S., Namkoong, H., Farhadi, A., Carmon, Y., Kornblith, S., and Schmidt, L. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *PMLR*, pp. 23965–23998, 2022.

Yadav, P., Tam, D., Choshen, L., Raffel, C. A., and Bansal, M. TIES-merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

A. Experimental Details

Architecture. The PlainCNN is a VGG-style feedforward network: three blocks of [Conv3×3–BN–ReLU–Conv3×3–BN–ReLU–MaxPool2×2] with 32, 64, and 128 channels respectively (all convolutions bias-free, padding 1), followed by FC(2048, 256)–ReLU–FC(256, 10), totaling ∼1.2M parameters. Convolutions use Kaiming-normal initialization; BN scales and shifts are initialized to 1 and 0. Training uses random crop (padding 4), horizontal flip, and per-channel normalization. The absence of skip connections leaves every hidden channel group freely permutable, so layer-wise weight matching covers the full permutation symmetry group of the architecture.

Weight matching. For each permutable channel group, a single linear assignment minimizes $\|W_a - PW_b\|_F^2$ over permutations P, with the cost matrix accumulating inner products of both the layer's output weights (including the associated BN parameters) and the next layer's input weights; one greedy pass over the layers is used, without alternating iterations.

Barrier evaluation. The supremum in Eq. (1) is approximated on a grid over the simplex: for pairs, 11 uniformly spaced points on $[0, 1]$ (the 9 interior points are evaluated; the endpoints define the baseline), and for triples, the interior points of a barycentric grid of resolution 5 (6 points with all coordinates positive; the boundary faces are covered by the pairwise evaluations). Pairwise barriers use the first 5,000 CIFAR-10 training samples for both BN recalibration

and loss evaluation; triplet barriers use 1,000 samples for recalibration and 2,000 for loss evaluation. Barrier evaluations use batch size 256; BN recalibration runs the network in training mode for one full pass over the calibration subset, then evaluation proceeds in inference mode. Betti numbers are computed as the number of eigenvalues of the corresponding Hodge Laplacian below a tolerance of 10^{-8}.

Triplet enumeration. We compute all $\binom{24}{2} = 276$ pairwise barriers. Triplet barriers are computed for all 884 3-cliques present at the 75th percentile of pairwise barriers ($\tau \approx 2.45$). This cutoff bounds only the set of triangles we evaluate; the analysis itself sweeps a range of τ (Table 1). Every 3-clique above the cutoff contains an edge whose pairwise barrier already exceeds the 75th percentile, so the corresponding triple fails at the pairwise level before any higher-order effect can arise.

Sources of randomness. The random seed controls both parameter initialization and data ordering in our setup, so the seed–potential correlation ($\rho = 0.96$) aggregates the two sources of randomness. Training models from identical initializations with different data orderings, and conversely, would isolate the contribution of each and reveal whether the cluster structure in Figure 3 originates in the initialization basin or in SGD noise; we leave this to future work.

B. Additional Hodge Interpretation

The Hodge decomposition of edge signals (Figure 1) is the discrete analogue of the Helmholtz decomposition of vector fields, which splits a field into a curl-free gradient part, a divergence-free rotational part, and a residual that is both. In a resistive electrical network, the gradient component consists of the edge voltages induced by node potentials and is fully determined by one scalar per node. For steady incompressible flow on an annular domain, the circulation around the central hole has zero vorticity yet is induced by no single-valued potential; this is the harmonic component, and its dimension equals the number of holes in the domain. On a simplicial complex the same structure arises with a function on oriented edges in place of the vector field: the gradient part is induced by a scalar function on vertices, the curl part by circulations around filled triangles, and the harmonic part is supported on edge cycles that bound no triangle. In our setting the edge signal is the pairwise merge barrier, vertices are models, and triangles are jointly mergeable triples; the harmonic component measures exactly the barrier structure carried by pairwise-mergeable cycles whose interiors fail to merge.

C. Filtration Details

Beyond $\tau \approx 2.38$, additional triangles fill in, the curl fraction grows at the expense of the harmonic fraction, and β_1 declines from its peak of 122 to 69, where it stabilizes for the remainder of the evaluated range (up to the 75th-percentile cutoff $\tau \approx 2.45$). The cycles we describe as persistent are the 69 that survive at the upper end of this range.

D. Anomaly Interpretation

For seed 0, a gradient-dominated barrier profile is explained by a scalar potential at the node, so the excess barrier follows from the position of the model relative to its merge partners and can be reduced by routing the alignment through low-potential neighbors. The within/cross-seed gap of seed 0 confirms that low-barrier paths to these models exist. For seeds 3–5, harmonic energy admits no explanation by node potentials or by corrections on filled triangles, so any choice of reference model or of merge weights within a fixed alignment leaves it unchanged. Consistent with this, the barriers of seeds 3–5 are equally high toward all partners. The label *toxic* characterizes behavior under the alignment family considered here (weight matching over star and MST topologies); whether richer alignment families reduce this component is an open question.

E. Orientation Invariance

Reversing the orientation of an edge negates the corresponding coordinate of C^1 together with the corresponding column of B_1 and row of B_2; all three subspaces in Equation (2) transform covariantly under this sign change, so the energy of each component, the potential $\mathbf{s}$ up to a global constant, and β_1 are independent of the orientation choice. Only the entrywise signs of the component vectors depend on it: for example, $(\mathbf{b}_{\text{grad}})_{(i,j)} = s_j - s_i$ can be negative, and the sign encodes the relative order of s_i and s_j under the reference orientation.

F. Computational Cost

Star alignment requires $n-1$ weight-matching calls. MST alignment requires one initial star pass ($n-1$ calls), all $\binom{n}{2}$ pairwise barrier evaluations, one Hodge decomposition, and $n-1$ re-alignment calls—roughly twice the matching cost plus the barrier evaluations. The dominant term is barrier evaluation: $\binom{n}{2}$ pairs $\times$ 9 interior interpolation points, each requiring BN recalibration and a loss evaluation over 5,000 samples. The Hodge decomposition itself is an eigendecomposition of a $|E| \times |E|$ matrix with $|E| \leq \binom{24}{2} = 276$ and runs in seconds on CPU. For our 24 models, the full pipeline (alignment, barriers, Hodge analysis) takes about 2 hours on a single NVIDIA RTX PRO 6000 Blackwell GPU (96 GB), and the MST re-alignment pass adds about 1 hour. All experiments use PyTorch with CUDA 12.8.