Sheaf Cohomology for Neural Interpretability: Quantifying Cross-Context Semantic Consistency with Frozen Restriction Maps

Bryce Grant
Jiyawei Yang
bag100@case.edu
jxy1213@case.edu
Case Western Reserve University

ABSTRACT

We propose a sheaf-theoretic framework to quantify whether neural features preserve their semantics across contexts (augmentations or layers). We learn a *frozen sheaf* once on a FIT set: linear restrictions on edges, transports on faces, and keep these maps fixed. This makes the cohomology groups H^0/H^1 fixed properties of the ruler (cover+maps), while simple quadratic energies computed with the frozen maps become measurements on EVAL: a consistency energy $E_{\text{cons}} = \|\delta^0 a\|^2$ and a cycle energy $E_{\text{cycle}} = \|\delta^1 \delta^0 a\|^2$. On MNIST rotations and CIFAR-10 (ResNet layers), our ruler calibrates nulls (identity-sheaf ≈ 0; scrambled/random $\gg$ real) and locates layers where cross-context semantics drift. We explore sheaf-consistency regularization for sparse autoencoders, though current formulations require further optimization. We further introduce a discrete Hodge decomposition on edges to localize inconsistency into *exact* (node differences), *coexact* (cycle residuals), and *harmonic* (globally consistent) components, providing insight into the nature of semantic drift.

1 INTRODUCTION

The interpretability of neural network features remains a fundamental challenge. While sparse autoencoders (SAEs) [5, 6] have shown promise in extracting interpretable features through sparsity constraints, they provide no formal guarantees about semantic consistency across different contexts. A feature that appears to detect "edges" in one context might encode entirely different information under slight transformations.

We address this gap by introducing a topological framework based on sheaf cohomology [7, 12] that quantifies theconsistency of features across contexts. Our key insight is that monosemantic features, those with consistent causal effects, correspond to the kernel of a coboundary operator on a sheaf structure, while polysemantic features manifest as nontrivial cohomology classes in H^1.

2 RELATED WORK

2.1 Feature Interpretability

A central challenge in neural network interpretability is *polysemanticity* where individual neurons often activate for multiple, semantically unrelated concepts. [8] demonstrated that this phenomenon arises from superposition, where networks represent more features than they have neurons for by encoding sparse features in near-orthogonal directions. Their toy models revealed phase transitions governing when features are stored monosemantically versus in superposition, with connections to polytope geometry and adversarial examples.

Sparse autoencoders (SAEs) [6] offer a scalable, unsupervised approach to resolving superposition. By training autoencoders with L_1 penalties on activations, SAEs learn overcomplete dictionaries of sparsely activating features that are more interpretable than individual neurons. Bricken et al.[5] showed that SAE features can identify causally relevant directions invisible in the neuron basis, enabling finer-grained circuit analysis. However, SAEs optimize for sparsity and reconstruction, not semantic consistency across contexts, which leaves open questions about whether extracted features maintain stable meanings under distribution shift.

2.2 Concept-Based Explanations

Rather than attributing importance to low-level features like pixels, concept-based methods explain model behavior in terms of human-understandable abstractions. Testing with Concept Activation Vectors (TCAV) [15] learns linear classifiers to identify concept directions in activation space, then uses directional derivatives to quantify how sensitive predictions are to user-defined concepts. TCAV provides global explanations without requiring model retraining.

Automated Concept-based Explanations (ACE) [10] extends TCAV by automatically discovering concepts through multi-resolution segmentation and clustering in activation space, eliminating the need for manually curated concept datasets. Network Dissection [2] takes a complementary approach, quantifying interpretability by measuring alignment between individual hidden units and a broad vocabulary of semantic concepts (objects, parts, textures, materials). Their finding that interpretability is axis-aligned—destroyed by random rotations that preserve discriminative power—suggests networks learn meaningful decompositions, but provides no guarantees about cross-context consistency.

2.3 Causal Interpretability

Pearl's structural causal framework [19] provides foundations for understanding interventional effects, but connecting causality to neural network internals remains challenging. Invariant Risk Minimization (IRM) [1] learns representations that elicit invariant predictors across training environments, targeting features with stable causal relationships to the target. While IRM addresses distributional robustness, it operates at the representation level rather than providing tools for analyzing individual features.

2.4 Sheaves in Machine Learning

[12] established the spectral theory of cellular sheaves, introducing the sheaf Laplacian $L_{\mathcal{F}} = \delta^T \delta$ and connecting global sections to $\ker L_{\mathcal{F}}$ via the Hodge theorem. [11] applied this to graph neural networks, showing that sheaf Laplacians generalize standard graph diffusion and outperform GCNs on signed and heterogeneous graphs where edge relationships are non-constant or asymmetric.

[3] extended this with Neural Sheaf Diffusion, proving that nontrivial sheaf geometry addresses heterophily and oversmoothing in GNNs by enabling linear class separation in the infinite-time diffusion limit. For knowledge representation, [9] cast knowledge graph embedding as learning approximate global sections of a knowledge sheaf, unifying methods like TransE and RotatE as coboundary variations and enabling inference over composite relations. [13] introduced discourse sheaves, modeling private opinions and public discourse via restriction maps.

Most relevant to our approach, [18] address learning sheaf Laplacians from data by jointly inferring topology and restriction maps through total variation minimization with closed-form updates. This motivates our use of learned restriction maps as a frozen measurement instrument.

Rather than using sheaves to improve architectures or learn embeddings, we use sheaf cohomology to quantify interpretability of existing representations. The dimensions of H^0 and H^1 provide a principled metric for cross-context semantic consistency, features in $\ker \delta^0$ maintain consistent causal effects across contexts, while nontrivial H^1 classes signal obstructions to global consistency.

3 SHEAF-THEORETIC PRELIMINARIES

Definition 3.1 (Cellular Sheaf). *A cellular sheaf $\mathcal{F}$ on a regular CW-complex X assigns: (i) a vector space $\mathcal{F}(\sigma)$ to each cell $\sigma \in X$, and (ii) a linear map $\mathcal{F}_{\sigma \trianglelefteq \tau} : \mathcal{F}(\sigma) \to \mathcal{F}(\tau)$ for each incidence $\sigma \trianglelefteq \tau$ [4].*

For neural interpretability, we construct sheaves where:

- **0-cells (vertices)**: Contexts U_i with stalks $\mathcal{F}(U_i) = \mathbb{R}^K$ (feature activations)
- **1-cells (edges)**: Overlaps $U_i \cap U_j$ with stalks $\mathcal{F}(e_{ij}) = \mathbb{R}^{d_e}$
- **2-cells (faces)**: Triple overlaps with stalks $\mathcal{F}(f_{ijk}) = \mathbb{R}^{d_f}$

The coboundary operators form a cochain complex:

$$C^0(X;\mathcal{F}) \xrightarrow{\delta^0} C^1(X;\mathcal{F}) \xrightarrow{\delta^1} C^2(X;\mathcal{F}) \tag{1}$$

3.1 Frozen Sheaf Approach

Our innovation is the **frozen sheaf**:

(1) **Learn once (FIT):** On training data, learn linear restrictions $R_i^{ij} : V_i \to E_{ij}$ and transports $B_{ij} : E_{ij} \to E_{ij}$ that minimize reconstruction error on overlaps
(2) **Freeze maps:** Fix these learned maps to create a permanent "ruler"
(3) **Measure (EVAL):** Apply the frozen maps to new features to compute energies

This separation ensures that cohomology dimensions H^0/H^1 are structural invariants of the ruler, while energies become measurements of semantic consistency.

4 METHOD

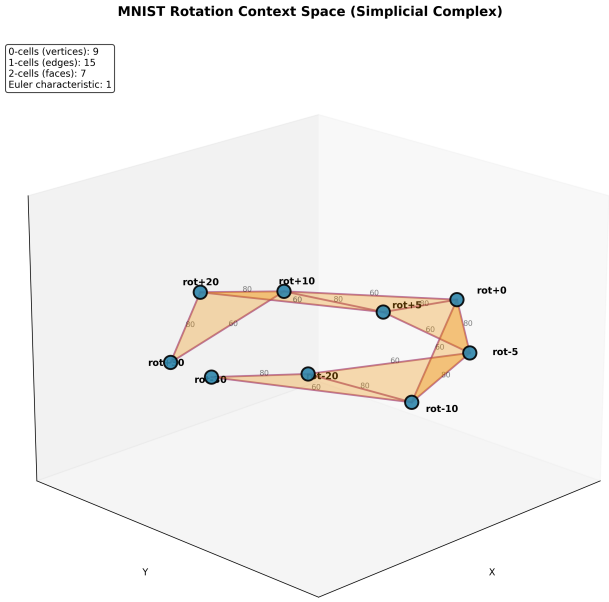

Figure 1: Simplicial complex of the MNIST context space

4.1 Complex Construction

Overlap Definition: For rotation contexts, samples appearing in multiple augmentations define Ω_{ij}. We use 70% partial overlaps to avoid trivial alignment. For layer contexts in CIFAR-10, all samples pass through all layers, so $\Omega_{ij} = \{1, \ldots, N\}$.

Edge Stalk Dimension: We set $d_e = \min(K, 16)$ where K is the feature dimension, balancing expressiveness with computational efficiency.

2-Skeleton: Faces are formed by all triangles (i, j, k) where contexts i, j, k are mutually adjacent in the cover topology. Face dimension is set to $d_f = d_e/2 = 8$ to ensure non-trivial δ^1.

4.2 Learning Restriction Maps

Given feature matrices $F_i \in \mathbb{R}^{n_i \times K}$ in contexts U_i and overlaps Ω_{ij} indexing shared samples, we learn restriction maps via:

Canonical Correlation Analysis (CCA): Find projections maximizing correlation between $F_i[\Omega_{ij}]$ and $F_j[\Omega_{ij}]$.

Procrustes Alignment: For edge $e = (i, j)$, solve:

$$\min_Q \|F_i[\Omega_{ij}]Q - F_j[\Omega_{ij}]\|_F^2 \text{ s.t. } Q^T Q = I \tag{2}$$

Transport Map: Learn optimal linear alignment:

$$B_e = (R_i^{e\top} R_i^e)^{-1} R_i^{e\top} R_j^e \tag{3}$$

4.3 Sheaf Construction

Given n contexts $\{U_1, \ldots, U_n\}$ with feature activations $a_i \in \mathbb{R}^K$, we construct a cellular sheaf on a simplicial complex X where vertices

are contexts and edges connect contexts with shared samples $\Omega_{ij} \neq \emptyset$.

For each edge $e = (i, j)$, we learn:

- Restriction maps $R_i^e : \mathbb{R}^K \to \mathbb{R}^{d_e}$ and $R_j^e : \mathbb{R}^K \to \mathbb{R}^{d_e}$
- Transport map $B_e : \mathbb{R}^{d_e} \to \mathbb{R}^{d_e}$ aligning the two restrictions

The coboundary operator $\delta^0 : C^0 \to C^1$ is:

$$(\delta^0 a)_e = R_j^e a_j - B_e R_i^e a_i \tag{4}$$

Learning Objective: On the FIT set, we minimize:

$$\min_{R,B} \sum_{e=(i,j)} \sum_{s \in \Omega_{ij}} \|R_j^e f_j(s) - B_e R_i^e f_i(s)\|^2 \tag{5}$$

where $f_i(s)$ is sample s's feature vector in context i.

4.4 Energy Metrics

For a section $a = \{a_i\}$ of node features, we define:

Consistency Energy:

$$E_{\text{cons}}(a) = \sum_{e=(i,j)} \|R_j^e a_j - B_e R_i^e a_i\|^2 \tag{6}$$

Cycle Energy:

$$E_{\text{cycle}}(a) = \|\delta^1 \delta^0(a)\|^2 \tag{7}$$

Remark: For a true cellular sheaf, $\delta^1 \circ \delta^0 = 0$. However, our learned transport maps may violate the cocycle condition $B_{jk} B_{ij} = B_{ik}$, making E_{cycle} a measure of how far the learned structure deviates from a valid sheaf.

4.5 Hodge Decomposition

Let $s = \delta^0 a$ be the edge mismatch. The discrete Hodge decomposition [18] with edge Laplacian $\Delta_1 = \delta^{1\top}\delta^1 + \delta^0\delta^{0\top}$ decomposes:

$$s = s_{\text{exact}} + s_{\text{harm}} + s_{\text{coexact}} \tag{8}$$

where $s_{\text{exact}} \in \text{im}(\delta^0)$, $s_{\text{harm}} \in \ker \Delta_1$, and $s_{\text{coexact}} \in \text{im}(\delta^{1\top})$.

4.6 Sheaf-Regularized SAE

We modify the SAE loss to penalize inconsistency across contexts:

$$\mathcal{L} = \|x - \hat{x}\|_2^2 + \lambda_1 \|a\|_1 + \lambda_{\text{cons}} \cdot E_{\text{cons}}(a) \tag{9}$$

where the consistency energy acts as a differentiable regularizer:

$$E_{\text{cons}}(a) = \sum_{e=(i,j)} \|R_j^e a_j - B_e R_i^e a_i\|_2^2 \tag{10}$$

5 EXPERIMENTS

5.1 setup

Datasets & Contexts:

- MNIST [17]: Rotations at $\{-30, -20, -10, -5, 0, 5, 10, 20, 30\}$
- CIFAR-10 [16]: ResNet-18 [14] layer features (layers 1-4)

Baselines:

- Standard SAE with L_1 regularization
- Identity sheaf (identical features + identity maps, should be ≈ 0)
- Random restriction maps (null model)
- Scrambled overlaps (permutation test)

Metrics.

- $\dim H^0$, $\dim H^1$, and ratio H^1/H^0
- Consistency and cycle energies
- Hodge decomposition percentages

5.2 Results

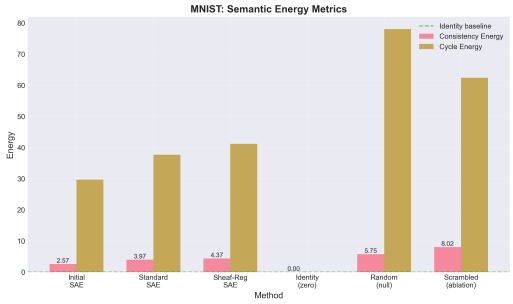

Figure 2: MNIST semantic energy metrics

Method	E_{cons}	E_{cycle}
Identity Sheaf	0.00	—
Initial SAE	2.57	29.71
Standard SAE	3.97	37.68
Sheaf-Reg SAE	4.37	41.18
Random	5.75	78.09
Scrambled	8.02	62.42

Table 1: MNIST energy metrics

Within Table 1 we observe the MNIST energy metrics. Sheaf-Reg SAE shows slightly higher energies, suggesting the regularization weight λ_{cons} may need tuning or that the regularizer optimizes for different objectives during training.

Layer	$\dim H^0$	Interpretation
conv1	192	Local features
layer1	896	Increasing abstraction
layer2	2304	More global
layer3	5120	Semantic invariants

Table 2: CIFAR-10 ResNet18 layer-wise H^0 dimensions

We observe that the frozen sheaf provides a calibrated ruler: identity sheaf achieves zero energy (0.00), while null models (random/scrambled) show 2-3× higher energies. Hodge decomposition reveals approximately equal exact (47%) and harmonic (48%) components, with minimal coexact (4%), suggesting features have both local and global inconsistencies. CIFAR-10 layer analysis in Table 2 shows increasing H^0 dimension with depth (192→5120), indicating more globally consistent features in deeper layers. Current regularization formulation requires further tuning—observed slight energy increase suggests competing objectives between reconstruction and consistency.

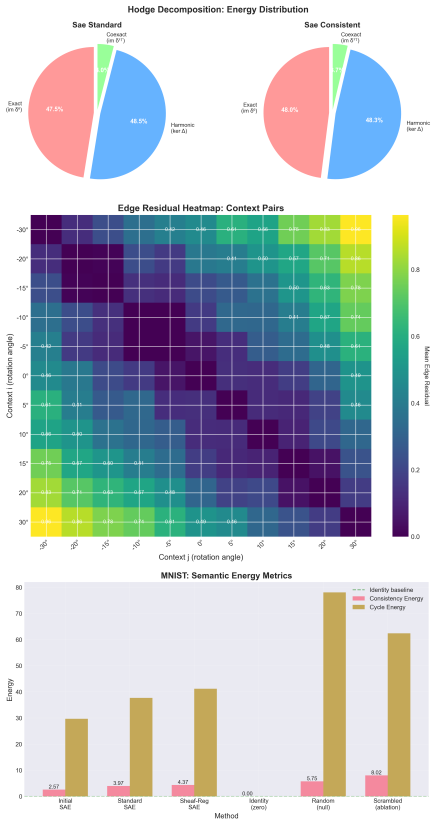

Figure 3: (Top) Hodge decomposition showing exact, harmonic, and coexact components. (Center) Edge residual heatmap showing per-edge inconsistencies. (Bottom) Energy comparison across different methods.

6 DISCUSSION

Our results demonstrate that sheaf cohomology provides meaningful quantification of feature interpretability. The frozen sheaf approach ensures clean separation between structural properties (the ruler) and behavioral measurements (energies). The Hodge decomposition adds interpretability by localizing sources of inconsistency. However, this is not without limitations. Linear restrictions may not capture all nonlinear relationships. Currently, the results depend on the context cover design and the current regularizer formulation needs careful tuning.

7 CONCLUSION

We introduced a topological framework for quantifying neural feature interpretability through sheaf cohomology. The frozen sheaf methodology provides: fixed structural invariants (H^0/H^1 dimensions), clean energy measurements of semantic consistency and-hodge decomposition to localize inconsistency types.

Our formulation connects abstract topological concepts to interpretability tools. Future work will explore nonlinear restriction maps, applications to large language models, and integration with existing interpretability methods. This framework extends to several directions:

Large Language Models: Applying frozen sheaves to transformer activations across context perturbations (paraphrases, translations, prompt variations) could identify which features maintain consistent semantics across linguistic contexts.

Vision-Language-Action Models: For robotics, vision-language-action models (VLAs), have allowed researchers to move closer towards generalist robotic policies [**?**]. By performing sheaf analysis over the action head, we could reveal which neurons encode control primitives. By constructing a sheaf over contexts that vary speed, direction, or manipulation type, we can identify H^0 features that represent these concepts consistently, enabling targeted steering for fine-grained robotic control and safety verification.

Scaling: While our experiments focus on SAE features and ResNet layers, the frozen sheaf methodology requires only linear projections and can scale to modern foundation models by operating on layer activations or learned sparse representations. This will require exploring nonlinear restriction maps, automated context cover design, and integration with causal intervention methods for more robust interpretability.

REFERENCES

[1] Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization, 2020.

[2] Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A. Network dissection: Quantifying interpretability of deep visual representations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017).

[3] Bodnar, C., and et al. Neural sheaf diffusion: A topological perspective on heterophily and oversmoothing in gnns. In *NeurIPS* (2022).

[4] Bredon, G. E. *Sheaf Theory*, 2 ed. Graduate Texts in Mathematics. Springer, New York, NY, Jan. 1997.

[5] Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread 2* (2023).

[6] Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models, 2023.

[7] Curry, J. Sheaves, cosheaves and applications, 2014.

[8] Elhage, N., Nanda, N., Olsson, C., et al. Toy models of superposition. Technical report, 2022.

[9] Gebhart, T., Hansen, J., and Schrater, P. Knowledge sheaves: A sheaf-theoretic framework for knowledge graph embedding. *AISTATS (arXiv:2110.03789)* (2023).

[10] Ghorbani, A., Wexler, J., Zou, J., and Kim, B. Towards automatic concept-based explanations. In *Advances in Neural Information Processing Systems (NeurIPS)* (2019).

[11] Hansen, J., and Gebhart, T. Sheaf neural networks. *arXiv:2012.06333* (2020).

[12] Hansen, J., and Ghrist, R. Toward a spectral theory of cellular sheaves. *Journal of Applied and Computational Topology 3*, 4 (Aug. 2019), 315–358.

[13] Hansen, J., and Ghrist, R. Opinion dynamics on discourse sheaves, 2020.

[14] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016).

[15] Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., and Sayres, R. Interpretability beyond feature attribution: Testing with concept activation vectors (tcav). In *International Conference on Machine Learning (ICML)* (2018).

[16] Krizhevsky, A. Learning multiple layers of features from tiny images. Tech. rep., University of Toronto, 2009.

[17] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* (1998).

[18] Nino, L. D., Barbarossa, S., and Lorenzo, P. D. Learning sheaf laplacian optimizing restriction maps, 2025.

[19] Pearl, J. *Causality: Models, Reasoning, and Inference*, 2 ed. Cambridge University Press, 2009.