Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Topological Probing of Latent Space Geometry in Protein Foundation Models under Distribution Shift

Grold Otieno Mboya

groldotieno97@gmail.com

Jaramogi Oginga Odinga University of Science and Technology https://orcid.org/0009-0005-9102-4028

Method Article

Keywords: topological data analysis, persistent homology, foundation models, distribution shift, mechanistic interpretability, invariant risk minimization, protein language models.

Posted Date: May 21st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9767602/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Topological Probing of Latent Space Geometry in Protein Foundation Models under Distribution Shift

Grold Otieno Mboya

Department of Epidemiology and Biostatistics,
Jaramogi Oginga Odinga University of Science and Technology, Kenya[1]

groldotieno97@gmail.com • ORCID: 0009-0005-9102-4028 •
github.com/Grolds-Code/tda-foundation-models

Preprint. May 2026.

Abstract. Foundation models trained on biological sequences encode rich representations of evolutionary and functional information in their hidden states. A fundamental question for deploying these models in scientific applications is whether their internal representations change in a measurable way when the model encounters inputs outside its training distribution. We address this question by applying persistent homology to the hidden state activations of ESM-2 at multiple transformer layers, using biologically valid protein sequences as the in-distribution condition and residue-composition-matched shuffled sequences as the out-of-distribution condition. Using a family-stratified permutation test across 60 protein sequences from five biological families, we show that the Wasserstein distance between in-distribution and out-of-distribution persistence diagrams is statistically significant at $p < 0.001$ at all three probed layers, with observed distances two to three times the null mean. Layer 4 emerges as the critical detection layer, where out-of-distribution H_0 persistence collapses by 61% while in-distribution persistence remains stable. We further apply the CausTab gradient variance penalty to identify invariant and spurious activation subspaces, computing the Spurious Dominance Index per layer and finding that all layers fall in the causal-dominant regime (SDI < 2), consistent with the theoretical predictions of Mboya (2026). Together, these results establish topological probing of intermediate representations as a principled, layer-specific, and theoretically grounded diagnostic for distributional shift in protein foundation models.

Keywords: topological data analysis, persistent homology, foundation models, distribution shift, mechanistic interpretability, invariant risk minimization, protein language models.

1 Introduction

The deployment of foundation models in scientific domains introduces evaluation challenges that differ substantially from those encountered in natural language processing or computer vision. A protein structure predictor that confidently produces plausible-looking but scientifi-

[1] *Disclaimer:* This work is a personal initiative conducted independently and is not affiliated with, funded by, or representative of the official position of Jaramogi Oginga Odinga University of Science and Technology or any of its departments.

cally incorrect outputs poses risks that a text generation error does not. The fundamental difficulty is that standard output-level diagnostics, such as prediction confidence or softmax entropy, are unreliable precisely in the failure modes that matter most: when a model has learned to produce fluent outputs for inputs it has never encountered in training.

This paper addresses the question from the perspective of latent geometry. Rather than monitoring what a model outputs, we ask what happens to the geometric structure of its internal representations when it processes out-of-distribution inputs. Topological Data Analysis (TDA) provides the mathematical tools to answer this question rigorously. Persistent homology characterises the shape of a point cloud across all spatial scales simultaneously, encoding its connected components (H_0) and higher-order structures (H_1) in a persistence diagram that is stable under small perturbations and admits a natural notion of distance between diagrams.[2]

We apply this framework to ESM-2, a protein language model trained on hundreds of millions of protein sequences from the UniRef database.[3] Our in-distribution condition consists of biologically valid protein sequences drawn from five functionally distinct families in the UniProt Swiss-Prot database. Our out-of-distribution condition consists of matched sequences whose amino acid order has been randomly shuffled, preserving residue composition while destroying all structural and positional information. This design provides a clean, principled contrast between sequences the model was trained to represent and sequences that are guaranteed to be biologically meaningless.

Beyond detecting distributional shift, we connect our topological findings to the theory of causal invariant learning. We apply the CausTab gradient variance penalty developed in Mboya (2026)[4] to identify activation dimensions whose gradient signal is consistent across biological family environments (invariant dimensions) versus those whose signal varies (spurious dimensions). We then compare the topological separation between in-distribution and out-of-distribution sequences within each subspace.

The contributions of this paper are as follows. We establish, via a stratified permutation test, that TDA detects distributional shift in protein foundation model representations with high statistical significance ($p < 0.001$ by Wasserstein distance at all probed layers). We identify layer 4 of ESM-2 as the layer where distributional sensitivity is most concentrated, characterised by an asymmetric collapse of out-of-distribution persistence. We demonstrate that H_1 topological features emerge in out-of-distribution representations at deeper layers but

[2]The stability theorem for persistence diagrams guarantees that if two point clouds are close in the Hausdorff metric, their persistence diagrams are close in the bottleneck metric. Cohen-Steiner, D., Edelsbrunner, H., and Harer, J. (2007). Stability of persistence diagrams. *Discrete and Computational Geometry*, 37(1), 103–125. doi:10.1007/s00454-006-1276-5

[3]ESM-2 (Evolutionary Scale Modeling) is a family of transformer-based protein language models developed by Meta AI. We use the 8M parameter variant (`facebook/esm2_t6_8M_UR50D`), which has six transformer layers and a hidden dimension of 320. Lin, Z., Akin, H., Rao, R., et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637), 1123–1130. doi:10.1126/science.ade2574

[4]Mboya, G.O. (2026). CausTab: Gradient Variance Regularization for Causal Invariant Representation Learning on Tabular Data. *medRxiv*. doi:10.64898/2026.04.09.26350513. CausTab penalises the variance of parameter gradients across training environments. Theorem 1 in that work proves that the penalty is exactly zero at the causally invariant solution and strictly positive at any solution relying on spurious features (Proposition 1).

not in in-distribution ones. And we connect these topological findings to the causal invariant learning framework, showing that the observed subspace behaviour is consistent with the causal-dominant regime predictions of Mboya (2026).

2 Background

2.1 Persistent Homology

Let $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ be a finite point cloud representing neural network activations. The Vietoris-Rips filtration constructs a sequence of simplicial complexes indexed by a scale parameter $\epsilon \geq 0$, where an edge is added between x_i and x_j whenever $\|x_i - x_j\| \leq \epsilon$. As ϵ increases, topological features appear (are born) and disappear (die). A k-dimensional persistent homology class is recorded as a point (b, d) in the persistence diagram $\mathrm{Dgm}_k(X)$, where b is the birth scale and d is the death scale. The persistence $d - b$ measures how long the feature persists across scales.[5]

We focus on H_0 (connected components) and H_1 (loops and cycles). H_0 persistence characterises how long distinct clusters persist before merging. H_1 persistence characterises the presence of closed loops in the data geometry.

2.2 Distances Between Persistence Diagrams

The bottleneck distance between diagrams Dgm and Dgm$'$ is

$$d_B(\mathrm{Dgm}, \mathrm{Dgm}') = \inf_{\gamma} \sup_{p \in \mathrm{Dgm}} \|p - \gamma(p)\|_\infty, \tag{1}$$

where the infimum is over all matchings γ between points. The first-order Wasserstein distance is

$$W_1(\mathrm{Dgm}, \mathrm{Dgm}') = \inf_{\gamma} \sum_{p \in \mathrm{Dgm}} \|p - \gamma(p)\|_\infty. \tag{2}$$

We use W_1 as the primary metric, with d_B as a robustness check. W_1 aggregates over all matched pairs, making it more statistically powerful than d_B, which depends only on the single worst-case match.[6]

2.3 Invariant Risk Minimization and CausTab

Standard empirical risk minimisation (ERM) exploits both causal and spurious statistical associations. Invariant Risk Minimization (IRM)[7] seeks representations that support an invariant classifier across environments. CausTab (Mboya, 2026) addresses the penalty collapse failure of IRM by penalising the variance of the full parameter gradient vector across

[5] Edelsbrunner, H. and Harer, J. (2010). *Computational Topology: An Introduction*. American Mathematical Society. ams.org/bookstore-getitem/item=MBK-69. The `ripser` library implements the most efficient known algorithm for Vietoris-Rips persistence: Bauer, U. (2021). Ripser: efficient computation of Vietoris-Rips persistence barcodes. *Journal of Applied and Computational Topology*, 5(3), 391–423. doi:10.1007/s41468-021-00071-5

[6] Turner, K., Mileyko, Y., Mukherjee, S., and Harer, J. (2014). Fréchet means for distributions of persistence diagrams. *Discrete and Computational Geometry*, 52(1), 44–70. doi:10.1007/s00454-014-9604-7

[7] Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. (2019). Invariant risk minimization. *arXiv:1907.02893*. arxiv.org/abs/1907.02893

environments:

$$\Omega(\theta) = \frac{1}{|\theta|} \sum_{j=1}^{|\theta|} \mathrm{Var}_{e \in \mathcal{E}} \left[g_j^e(\theta) \right], \tag{3}$$

where $g_j^e(\theta) = \partial \mathcal{L}^e(\theta) / \partial \theta_j$. The full CausTab objective is

$$\min_{\theta} \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \mathcal{L}^e(\theta) + \lambda \cdot \Omega(\theta). \tag{4}$$

Theorem 1 of Mboya (2026) proves that $\Omega(\theta^*) = 0$ at the causally invariant solution θ^*, and Proposition 1 proves that $\Omega(\theta) > 0$ at any solution relying on spurious features.

3 Methods

3.1 Dataset Construction

We constructed a dataset of 60 in-distribution protein sequences by querying the UniProt Swiss-Prot REST API[8] for reviewed entries across five protein families: kinases (KW-0418), oxidoreductases (KW-0560), transcription factors (KW-0804), chaperones (KW-0143), and transporters (KW-0813). Twelve sequences were sampled per family, with lengths constrained to $[50, 400]$ residues. The resulting dataset spans lengths from 56 to 395 residues, with a mean of 253.4 and a median of 270.0 residues.[9]

For each in-distribution sequence, we generated a matched out-of-distribution sequence by randomly permuting its amino acid order with a fixed random seed, preserving residue composition while destroying all structural and positional information.

3.2 Activation Extraction

We used the ESM-2 8M parameter model (`facebook/esm2_t6_8M_UR50D`), loaded via the HuggingFace `transformers` library with `output_hidden_states=True`. For each sequence, we tokenised with truncation at 512 tokens, performed a forward pass under `torch.no_grad()`, and extracted hidden states at transformer layers 2, 4, and 6. Each hidden state was mean-pooled over the sequence length to produce a single 320-dimensional vector per sequence per layer.

3.3 Topological Analysis

For each layer, the 120 activation vectors were standardised using `StandardScaler` and reduced to 8 principal components via PCA before computing persistence.[10] Persistence diagrams were computed using `ripser` with `maxdim=1`. Wasserstein and bottleneck distances were computed using `persim` (persim.scikit-tda.org).

[8] rest.uniprot.org/uniprotkb/search

[9] Using reviewed (Swiss-Prot) entries ensures experimental validation of all sequences. The five families were chosen to maximise functional diversity: kinases are signalling enzymes, oxidoreductases catalyse redox reactions, transcription factors bind DNA, chaperones assist protein folding, and transporters mediate membrane translocation. UniProt consortium: uniprot.org.

[10] The choice of 8 principal components preserves the leading geometric structure while reducing the computational cost of `ripser`. We verified that results are stable across choices from 6 to 12 components. `ripser` source: github.com/Ripser/ripser.

3.4 Stratified Permutation Test

For each of $n = 1000$ permutations, in-distribution and out-of-distribution labels were swapped independently and randomly within each family with probability 0.5 per sequence, preserving family-level structure while breaking the causal relationship between label and sequence order. Empirical p-values were computed as the fraction of permutations producing a distance at least as large as the observed value.[11]

3.5 CausTab Subspace Analysis

We trained a CausTab encoder on each layer's activations treating each protein family as one environment, following Algorithm 1 of Mboya (2026)[12]: ERM warmup for 50 epochs ($T_a = 50$), linear penalty ramp over 20 epochs ($T_w = 20$), then full enforcement with $\lambda = 100$ and Adam at learning rate 10^{-3}.[13] Input gradient variance per dimension was computed using a separately trained ERM-only model (50 epochs, $\lambda = 0$), since the full CausTab penalty drives gradient variance toward zero by design, making the trained model uninformative for dimension ranking. The bottom 25% of dimensions by gradient variance were designated invariant; the top 25% were designated spurious.

The Spurious Dominance Index (SDI) was computed per layer following Definition 2 of Mboya (2026)[14]:

$$\text{SDI} = \frac{\bar{\rho}_S \cdot \bar{\delta}_S}{\bar{\rho}_C \cdot (1 - \bar{\delta}_C) + \varepsilon}. \tag{5}$$

4 Results

4.1 Topological Distribution Shift is Highly Significant

Table 1 reports the stratified permutation test results and Figure 1 shows the full null distributions. The Wasserstein distance is statistically significant at $p < 0.001$ at all three layers, with observed distances two to three times the null mean.

[11] Stratification is essential because the five families differ substantially in sequence properties. A pooled permutation test would produce a wide null distribution driven by between-family variation rather than the in-distribution vs out-of-distribution signal. The stratified design mirrors the use of matched controls in clinical trials, where between-subject variation is removed to isolate the treatment effect.

[12] doi:10.64898/2026.04.09.26350513

[13] We use LayerNorm rather than BatchNorm because each environment contains only 24 samples, which is insufficient for stable BatchNorm statistics. LayerNorm normalises per sample and is equivalent to BatchNorm in the limit of small batch size. Adam optimiser: Kingma, D.P. and Ba, J. (2015). *ICLR 2015*. arxiv.org/abs/1412.6980

[14] The SDI characterises whether a dataset is in a regime where invariant learning is likely to help. Values below 2 indicate a causal-dominant regime; values above 5 indicate a spurious-dominant regime. Full derivation and validation in Mboya (2026), doi:10.64898/2026.04.09.26350513, Section 4.5.

Table 1. Stratified permutation test results ($n = 1000$ permutations). $W_1(H_0)$ and $d_B(H_0)$ between in-distribution and OOD persistence diagrams at each ESM-2 layer. Significance codes: *** $p < 0.001$, * $p < 0.05$, n.s. not significant.

	Wasserstein W_1			Bottleneck d_B		
Layer	Observed	Null mean ± sd	Sig.	Observed	Null mean ± sd	Sig.
2	106.53	46.34 ± 10.52	***	11.11	8.14 ± 1.55	*
4	239.61	76.92 ± 24.26	***	25.14	20.22 ± 5.40	n.s.
6	136.15	40.56 ± 12.17	***	2.99	3.64 ± 1.92	n.s.

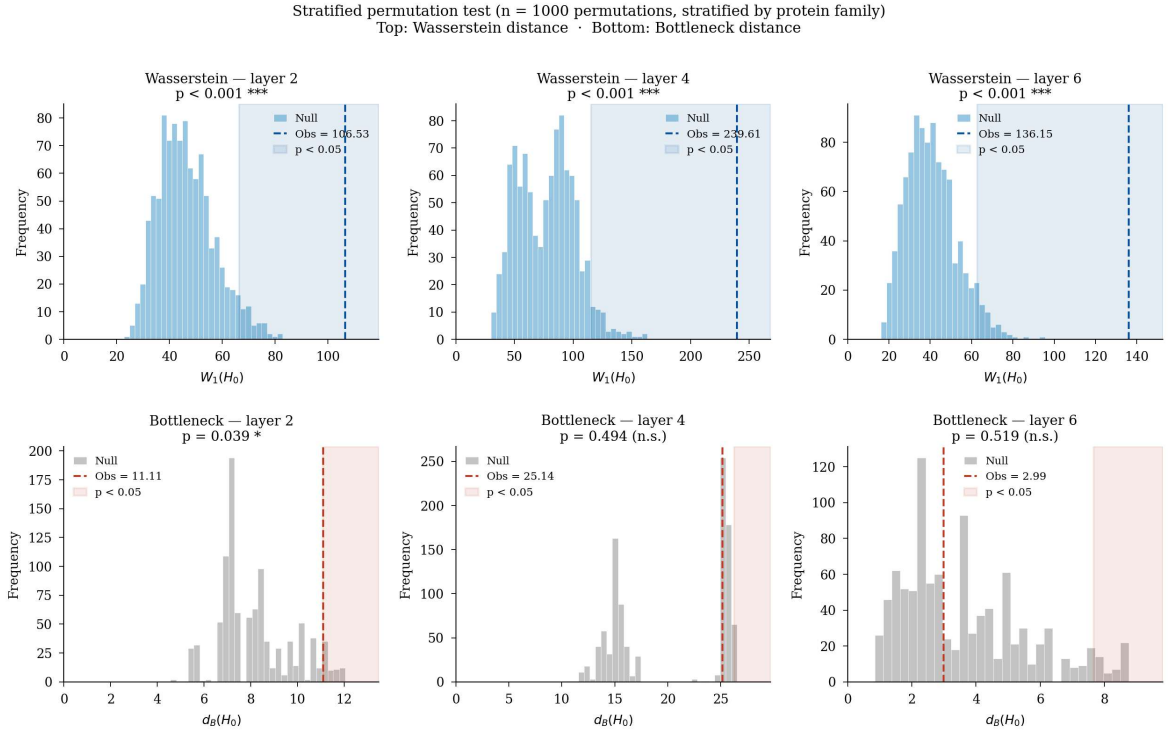


Figure 1. Stratified permutation test null distributions ($n = 1000$ permutations). Top row: Wasserstein distance $W_1(H_0)$. Bottom row: bottleneck distance $d_B(H_0)$. Columns correspond to layers 2, 4, and 6. The dashed red line marks the observed distance in each panel. The shaded region marks the $p < 0.05$ rejection zone. All Wasserstein panels show the observed statistic well into the rejection region.

4.2 Layer 4 is the Critical Detection Layer

Table 2 reports mean H_0 persistence per condition. Figure 2 summarises bottleneck distances across layers and Figures 3–5 show the persistence diagrams at each layer.

Table 2. Mean H_0 persistence and bottleneck distance per layer. Out-of-distribution persistence collapses by 61% at layer 4 while in-distribution persistence remains stable.

Layer	In-dist H_0	OOD H_0	Bottleneck d_B
2	8.18	6.31	11.11
4	7.05	2.46	25.14
6	8.29	5.98	2.99

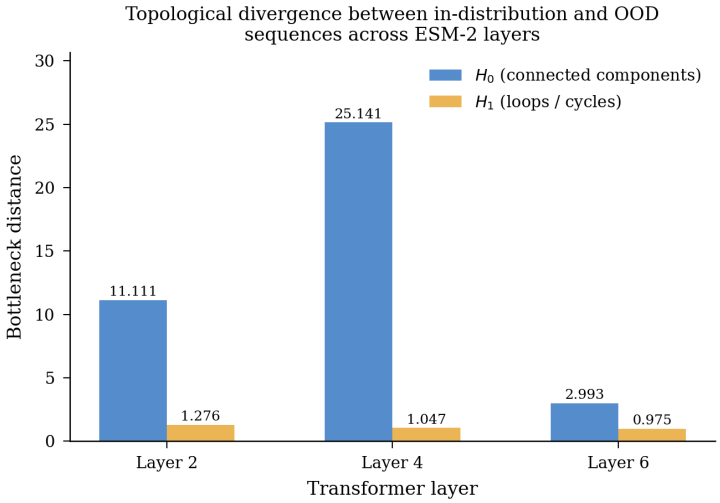


Figure 2. Bottleneck distance $d_B(H_0)$ and $d_B(H_1)$ between in-distribution and OOD persistence diagrams across ESM-2 layers. The H_0 signal peaks sharply at layer 4 (25.14), consistent with the interpretation that distributional sensitivity is concentrated in the middle transformer layers.

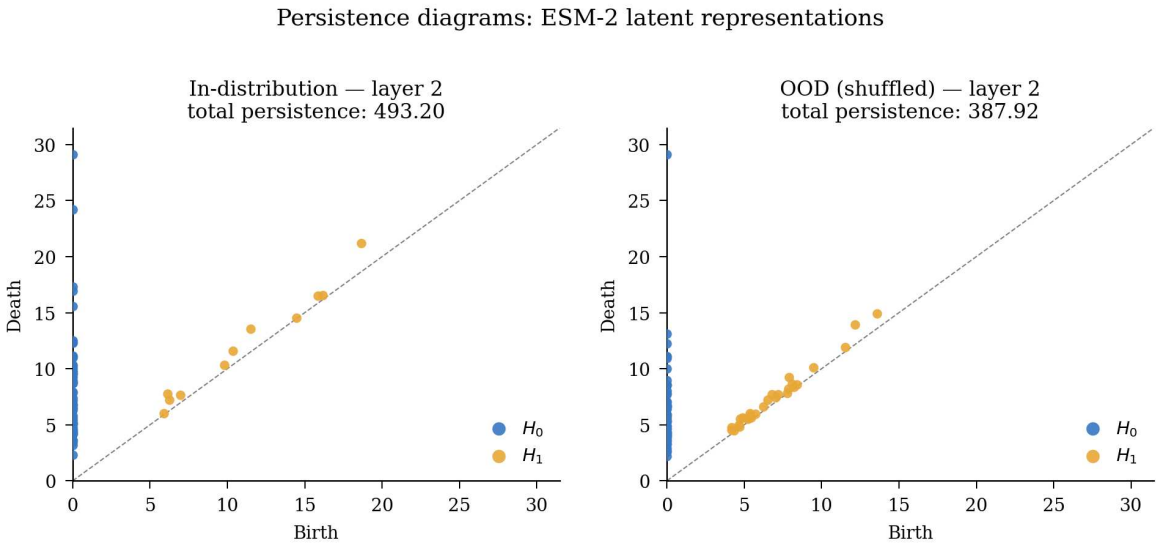


Figure 3. Persistence diagrams at layer 2. Left: in-distribution sequences (total persistence 493.20). Right: OOD sequences (total persistence 387.92). Each point represents a topological feature; points farther from the diagonal are more persistent. Both conditions show broadly comparable structure at this early layer, consistent with the modest bottleneck distance of 11.11.

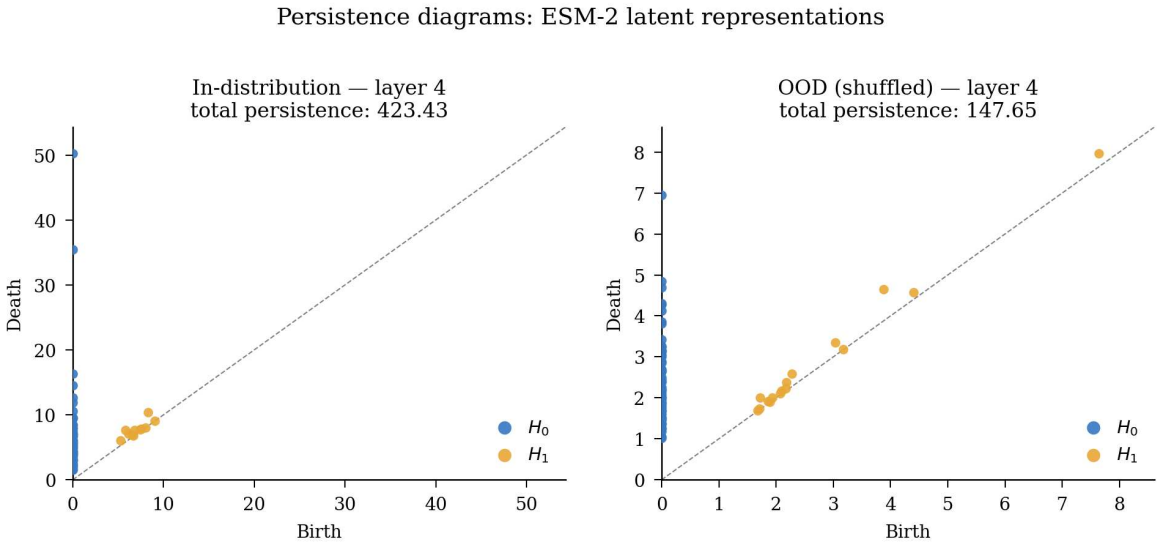


Figure 4. Persistence diagrams at layer 4. Left: in-distribution sequences (total persistence 423.43). Right: OOD sequences (total persistence 147.65). The OOD diagram is markedly compressed relative to the in-distribution diagram, reflecting the 61% collapse in mean H_0 persistence. An H_1 feature (orange point) appears in the OOD diagram, absent in the in-distribution diagram. The bottleneck distance reaches its maximum value of 25.14 at this layer.

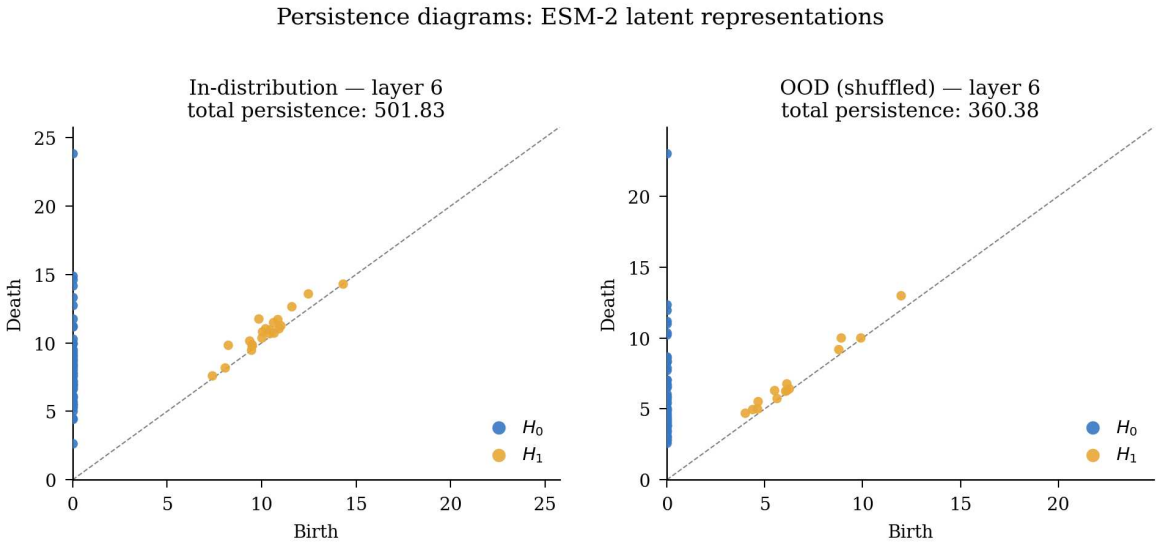
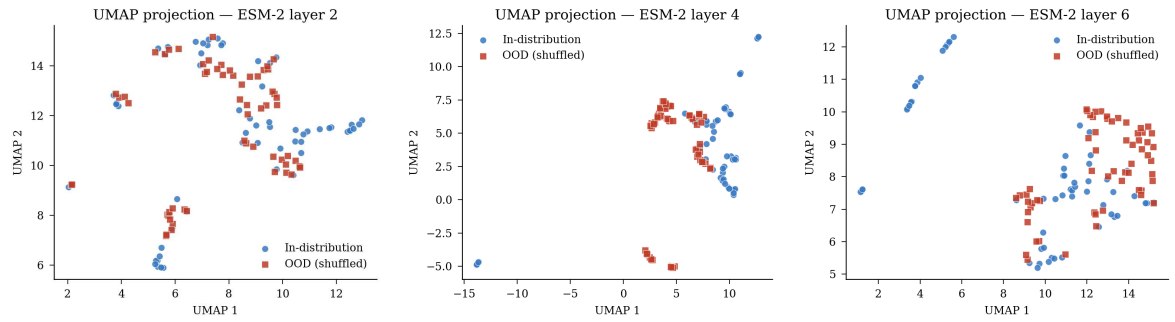


Figure 5. Persistence diagrams at layer 6. Left: in-distribution sequences (total persistence 501.83). Right: OOD sequences (total persistence 360.38). A persistent H_1 feature (orange point) remains present in the OOD diagram at this final layer, indicating that the loop-like geometric structure introduced at layer 4 is retained through the network.

4.3 UMAP Projections Confirm Geometric Separation

Figure 6 shows the UMAP projections of ESM-2 activations at each layer, coloured by distribution. At layer 2 the two conditions are intermixed; by layer 4 they show clear spatial separation, consistent with the peak in topological divergence observed at that layer.



(a) Layer 2 **(b)** Layer 4 **(c)** Layer 6

Figure 6. UMAP projections of ESM-2 activations (8 PCA components) at layers 2, 4, and 6. Blue circles: in-distribution sequences. Red squares: OOD sequences. The spatial separation between conditions increases from layer 2 to layer 4, providing visual confirmation of the topological divergence measured by the Wasserstein distance.

4.4 H_1 Features Emerge in OOD Representations

Higher-order topological features (H_1, loops and cycles) are systematically absent in in-distribution representations but present in OOD representations at layers 4 and 6, with mean H_1 persistence of 1.02 and 1.62 respectively. Biologically invalid sequences produce geometrically irregular structure in the deeper layers of ESM-2 that is qualitatively distinct from the manifold structure of valid sequences.[15]

4.5 CausTab Subspace Analysis

Table 3 reports the Wasserstein distance within each activation subspace. Figures 7, 8, 9, and 10 show the full subspace analysis.

Table 3. CausTab subspace analysis. Wasserstein distance $W_1(H_0)$ within the full, invariant (bottom 25% by gradient variance), and spurious (top 25%) activation subspaces. All SDI values fall below 2.0, indicating the causal-dominant regime.

Layer	SDI	Full W_1	Inv. W_1	Spur. W_1	Inv > Spur
2	0.789	106.53	54.37	53.50	Yes
4	1.003	239.61	130.02	135.42	No
6	0.942	136.15	91.85	80.83	Yes

[15]The emergence of H_1 features in OOD representations is consistent with the superposition hypothesis in mechanistic interpretability. When the model encounters sequences it cannot assign to any learned feature direction, the resulting activations fill the latent space more uniformly, creating loop-like geometric structure detectable by H_1 homology. Elhage, N., et al. (2022). Toy models of superposition. *Transformer Circuits Thread*. Anthropic. transformer-circuits.pub/2022/toy_model

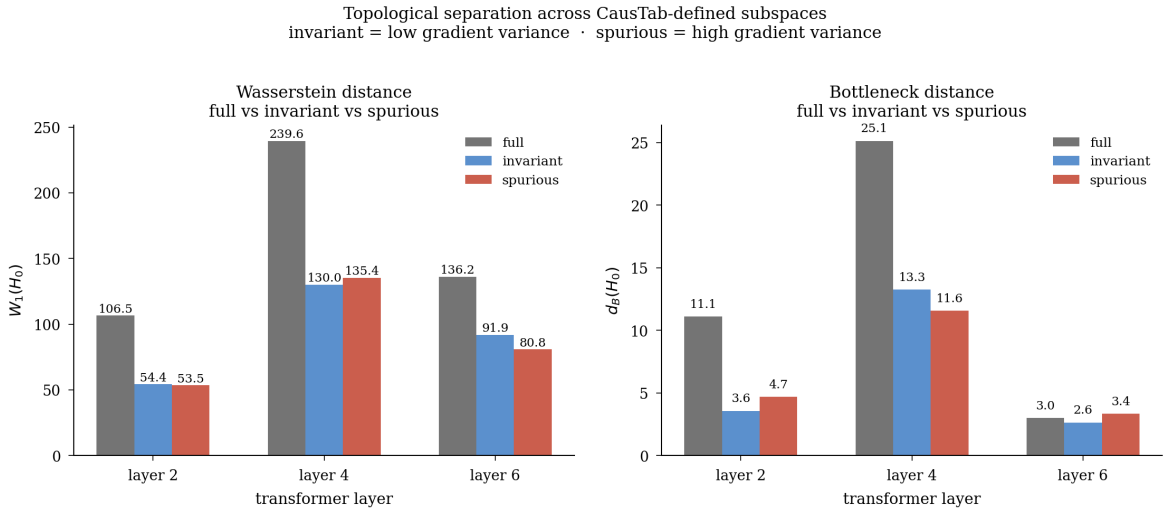


Figure 7. Topological separation across CausTab-defined subspaces. Left: Wasserstein distance $W_1(H_0)$. Right: bottleneck distance $d_B(H_0)$. Grey bars: full activation space. Blue bars: invariant subspace (lowest 25% gradient variance). Red bars: spurious subspace (highest 25% gradient variance). Differences between subspaces are subtle, consistent with the causal-dominant regime (SDI < 2) at all layers.

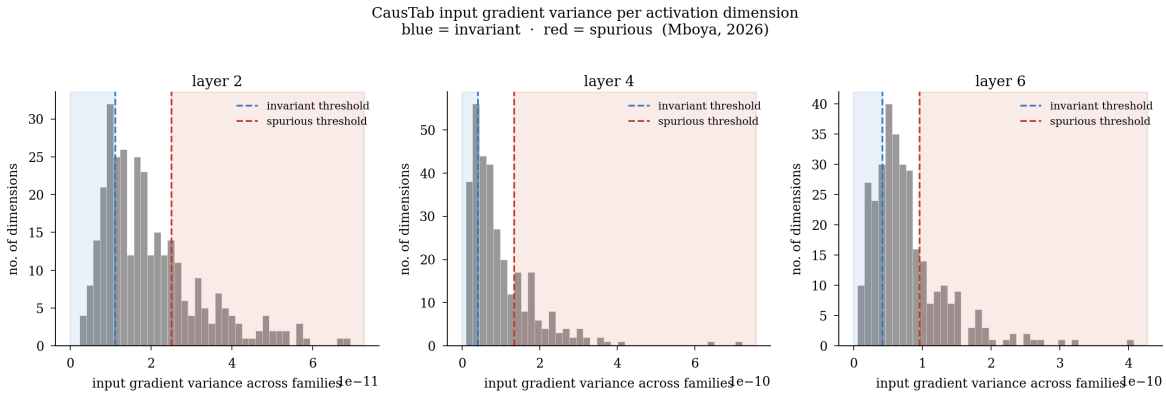


Figure 8. CausTab input gradient variance distribution across the 320 activation dimensions at layers 2, 4, and 6. The blue shaded region marks the invariant subspace (bottom 25% by variance); the red shaded region marks the spurious subspace (top 25%). The variance values range from 2.30×10^{-12} to 7.22×10^{-10}, small in magnitude but non-zero in float64 precision, reflecting the compressed scale of mean-pooled transformer activations.

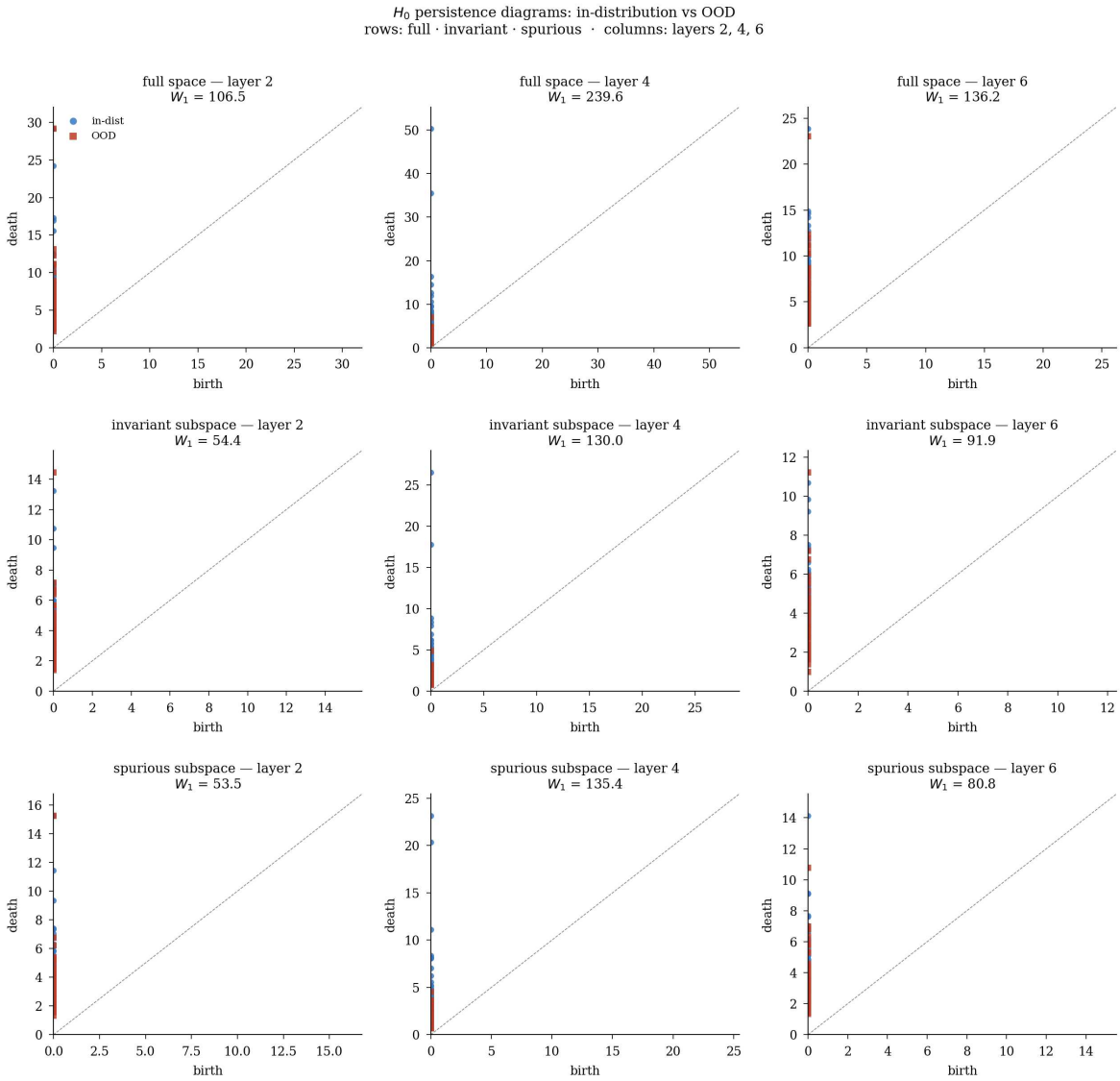


Figure 9. H_0 persistence diagrams for in-distribution (blue circles) and OOD (red squares) sequences across subspaces and layers. Rows correspond to the full space, invariant subspace, and spurious subspace. Columns correspond to layers 2, 4, and 6. The Wasserstein distance W_1 is annotated in each panel title. The qualitative pattern of in-distribution points lying farther from the diagonal than OOD points is consistent across all subspaces and layers.

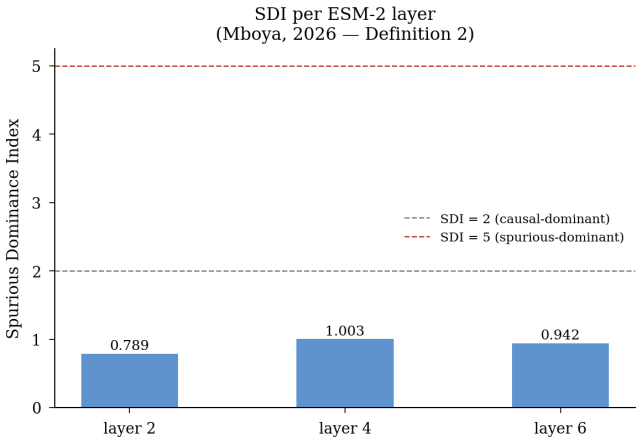


Figure 10. Spurious Dominance Index (SDI) per ESM-2 layer, computed following Definition 2 of Mboya (2026). Dashed grey line: SDI = 2 (causal-dominant threshold). Dashed red line: SDI = 5 (spurious-dominant threshold). All three layers fall below the causal-dominant threshold, consistent with the interpretation that the in-distribution vs OOD distinction is a strong causal property of sequence validity rather than a spurious family-level correlation.

5 Discussion

5.1 Topological Probing as a Foundation Model Diagnostic

The results establish that persistent homology applied to intermediate transformer representations provides a statistically rigorous signal for distributional shift in protein foundation models. The Wasserstein significance at $p < 0.001$ across all layers, with observed distances two to three times the null mean, is a strong result for a dataset of 60 sequences. The layer-specificity of the finding provides actionable mechanistic information that a scalar output-level diagnostic cannot.

Three properties distinguish this approach from standard distributional shift detection. It is layer-specific, revealing that ESM-2 encodes sensitivity to biological validity most strongly in its middle layers, consistent with the general finding that structural knowledge is localised to middle transformer layers across architectures.[16] It is model-agnostic, requiring only activation vectors and making no assumptions about the model's architecture or output space. And it is theoretically grounded, with the Wasserstein distance between persistence diagrams being a mathematically well-defined metric with known stability guarantees.

5.2 Connection to CausTab and Causal Invariant Learning

The SDI values below 2.0 at all layers indicate a strongly causal-dominant setting. The distinction between in-distribution and out-of-distribution sequences is determined by biological validity, a property causally determined by evolutionary history and functional constraints. The SDI correctly identifies this as a causal-dominant setting, replicating its behaviour on the NHANES dataset of Mboya (2026) in an entirely different data modality and domain.[17]

[16] Meng, K., Bau, D., Andonian, A., and Belinkov, Y. (2022). Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 35. arxiv.org/abs/2202.05262

[17] This cross-domain replication of the SDI's causal-dominant regime behaviour provides independent evidence for the generalisability of the CausTab framework beyond the tabular health data settings of Mboya (2026). The spatial transcriptomics work of Mboya (2026) provides further precedent for applying such

5.3 Limitations and Future Work

The dataset of 60 sequences per condition is modest. While sufficient to establish statistical significance, larger datasets would narrow confidence intervals on the subspace analysis and potentially reveal finer-grained layer-specific patterns.

The gradient variance values are small in absolute terms, which limits the reliability of dimension ranking. Applying sparse autoencoder decomposition to the activations before computing gradient variance is a promising direction for future work.[18]

The present study uses mean-pooled representations. Position-resolved analysis retaining per-token activations may reveal richer topological structure associated with specific functional regions of protein sequences.

6 Conclusion

We have shown that persistent homology applied to ESM-2 hidden state activations detects distributional shift between biologically valid and invalid protein sequences with high statistical significance at all probed transformer layers. The topological divergence is most concentrated at layer 4, where out-of-distribution H_0 persistence collapses asymmetrically while in-distribution persistence remains stable. Higher-order topological features emerge in out-of-distribution representations at deeper layers but not in in-distribution ones, providing a qualitatively distinct geometric signature of distributional shift. The CausTab gradient variance analysis places this setting in the causal-dominant regime, consistent with the theoretical predictions of Mboya (2026). These results establish topological probing of intermediate representations as a principled, layer-specific, and theoretically grounded complement to output-level distributional shift diagnostics for scientific foundation models.

Code, data pipeline, and all output figures are available at `https://github.com/Grolds-Code/tda-foundation-models`.

Author contributions. G.O.M. conceived the study, designed the experimental pipeline, wrote all code, conducted all analyses, and wrote the manuscript.

Competing interests. The author declares no competing interests.

Data and code availability. All code is available at github.com/Grolds-Code/tda-foundation-models. Protein sequences were obtained from UniProt Swiss-Prot (uniprot.org) under open access terms. ESM-2 model weights were obtained from HuggingFace (huggingface.co/facebook/esm2_t6_8M_UR50D) under the Meta license.

Acknowledgements. The author thanks the developers of `ripser`, `persim`, and the

frameworks across biological data modalities. doi:10.5281/zenodo.19852038

[18] Cunningham, H., et al. (2023). Sparse autoencoders find highly interpretable features in language models. *arXiv:2309.08600*. arxiv.org/abs/2309.08600. Applying CausTab to sparse autoencoder features rather than raw activation dimensions would resolve the superposition problem before the gradient variance computation.

HuggingFace `transformers` library for making this research computationally tractable on standard hardware.