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.

In Search of Manifolds Inside Protein Language Models: A Largely Negative Report

Olivia Denvis

oliviadenvis@gmail.com

Lampung University https://orcid.org/0009-0002-7443-2669

Research Article

Keywords:

Posted Date: July 22nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10433997/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.

In Search of Manifolds Inside Protein Language Models:
A Largely Negative Report

Olivia Denvis
Independent Researcher
oliviadenvis@gmail.com

Abstract

The manifold hypothesis—that high-dimensional data concentrate near a low-dimensional, smoothly varying manifold—has become an organizing metaphor for interpreting neural representations, and recent work reports rich geometric and even specific low-dimensional manifolds inside single-cell foundation models. We ask whether the internal representations of protein language models (pLMs) admit analogous manifolds. Applying a broad detection battery—intrinsic dimension estimators (TwoNN, MLE, PCA participation ratio), nonlinear dimensionality reduction (Isomap, LLE, diffusion maps, UMAP), persistent homology, trustworthiness and geodesic-connectivity diagnostics, and confound controls—to ESM-2 (8M–3B) and to ProtBERT and ProtT5, we find little evidence for such structure. The intrinsic dimension of pLM representations is high in absolute terms (best-layer TwoNN $\approx$63 for ESM-2 650M), grows with model size, and does not stabilize across neighborhood scales, contradicting a single global manifold. Nonlinear manifold coordinates never outperform a linear probe on the raw embeddings for secondary structure, remote homology, stability, or localization, and $\sim$64–128 linear dimensions are needed to saturate performance. Persistent homology recovers no robust loops beyond the noise floor, and the apparent clustering of protein families largely dissolves after regressing out sequence length, amino-acid composition, and homology (family silhouette $0.31 \rightarrow 0.05$). Crucially, the same battery cleanly detects manifolds in positive controls (a Swiss roll, a single-cell cell-cycle loop, a torus), so the negative result is not a failure of method. We conclude that pLM representations are better described as high-dimensional and predominantly linear than as low-dimensional manifolds, and we discuss why the discrete, homology-structured geometry of protein sequence space differs from the continuous developmental manifolds seen in single-cell data.

1 Introduction

Protein language models (pLMs) trained on hundreds of millions of unaligned sequences have become foundational tools in computational biology. Models such as ESM-1b (Rives et al., 2021), ESM-2 (Lin et al., 2023), ProtTrans's ProtBERT and ProtT5 (Elnaggar et al., 2022), ProteinBERT (Brandes et al., 2022), ProGen (Madani et al., 2023; Nijkamp et al., 2023), Ankh (Elnaggar et al., 2023), and CARP (Yang et al., 2024) produce residue- and sequence-level embeddings that transfer to structure prediction (Lin et al., 2023; Jumper et al., 2021), variant-effect estimation (Meier et al., 2021), and a wide range of property-prediction tasks (Rao et al., 2019; Xu et al., 2022). A natural question for interpreting these models is *geometric*: how is biological information organized in the high-dimensional activation space, and does it concentrate on a low-dimensional structure that we can detect and exploit?

The *manifold hypothesis*—that realistic high-dimensional data lie near a low-dimensional, smoothly varying manifold (Tenenbaum et al., 2000; Roweis and Saul, 2000; Fefferman et al., 2016)—is the usual lens for such questions, and it has been productive elsewhere. In single-cell transcriptomics, continuous differentiation trajectories and the cell cycle appear as low-dimensional manifolds and topological loops recoverable by diffusion maps and PHATE (Coifman and Lafon, 2006; Moon et al., 2019), and a recent line of work reports that single-cell foundation models internalize this geometry: a specific low-dimensional hematopoietic manifold is recoverable from scGPT and can even be distilled into a performant algorithm (Kendiukhov, 2026b), the representations carry multi-dimensional spectral geometry (Kendiukhov, 2026c) and convergent causal circuitry (Kendiukhov, 2026a), and a systematic survey of 141 geometric and topological

hypotheses finds substantial learned structure (Kendiukhov, 2026d). These positive results motivate an obvious transfer question: do protein language models, another dominant class of biological foundation model, contain analogous manifolds?

We investigate this directly and reach a *largely negative* conclusion. Assembling a broad manifold-detection battery—intrinsic-dimension estimators, nonlinear dimensionality reduction, persistent homology, faithfulness diagnostics, and confound controls—and applying it to ESM-2 across five sizes (8M–3B) as well as ProtBERT and ProtT5, we find that pLM representations do not present clean, exploitable, low-dimensional manifolds. Intrinsic dimension is high and grows with model size; nonlinear manifold coordinates never beat a linear probe; topology is trivial; and apparent family structure is largely a reflection of length, composition, and homology confounds. Because the identical battery detects manifolds in positive controls, the negative result reflects the data, not the tools.

Contributions.

- We assemble and openly document a manifold-detection battery for transformer representations, validated on synthetic and biological positive controls, and apply it uniformly to seven protein language models.
- We report a coherent set of negative findings: high and size-growing intrinsic dimension, no scale-stable manifold dimension, no downstream benefit from nonlinear coordinates, and no robust topology.
- We show that the apparent low-dimensional organization of protein families in pLM space is largely explained by sequence length, amino-acid composition, and homology, and collapses once these are controlled.
- We contrast this with the continuous manifold structure documented for single-cell models and argue that the discrete, combinatorial, homology-structured geometry of protein sequence space is a principled reason to expect the difference.

2 Related Work

Protein language models and their representations. Transformer pLMs are trained with masked-language-modeling objectives on large sequence databases (UniRef, UniProt) (Suzek et al., 2015; The UniProt Consortium, 2021) and yield embeddings from which structure and function are readily decoded. Attention maps recover residue–residue contacts (Rao et al., 2021a; Vig et al., 2021), an effect sharpened by conditioning on multiple-sequence alignments (Rao et al., 2021b); linear probes extract secondary structure and other properties (Rao et al., 2019; Xu et al., 2022), and sparse autoencoders surface interpretable, biologically meaningful features (Simon and Zou, 2024). A recurring theme is that much of this information is *linearly accessible*: simple probes on frozen embeddings are strong, which already hints that the useful geometry may be linear and high-dimensional rather than a curved low-dimensional manifold. Our study makes this hint precise and adversarial.

Intrinsic dimension and representation geometry. The intrinsic dimension (ID) of neural representations is far below their ambient width but still substantial: Ansuini et al. (2019) report a characteristic "hunchback" profile—ID rising then falling across depth—with peak values in the tens to low hundreds for image networks, and note that linear (PCA) estimates exceed nonlinear ID estimates by one to two orders of magnitude, a signature of curvature or of genuinely high linear dimension. Pope et al. (2021) estimate the ID of natural images at a few dozen. We use the standard estimators TwoNN (Facco et al., 2017) and the maximum-likelihood estimator (Levina and Bickel, 2004), together with a PCA participation ratio, and study how their disagreement and scale-dependence bear on the manifold question. Representation-similarity tools such as

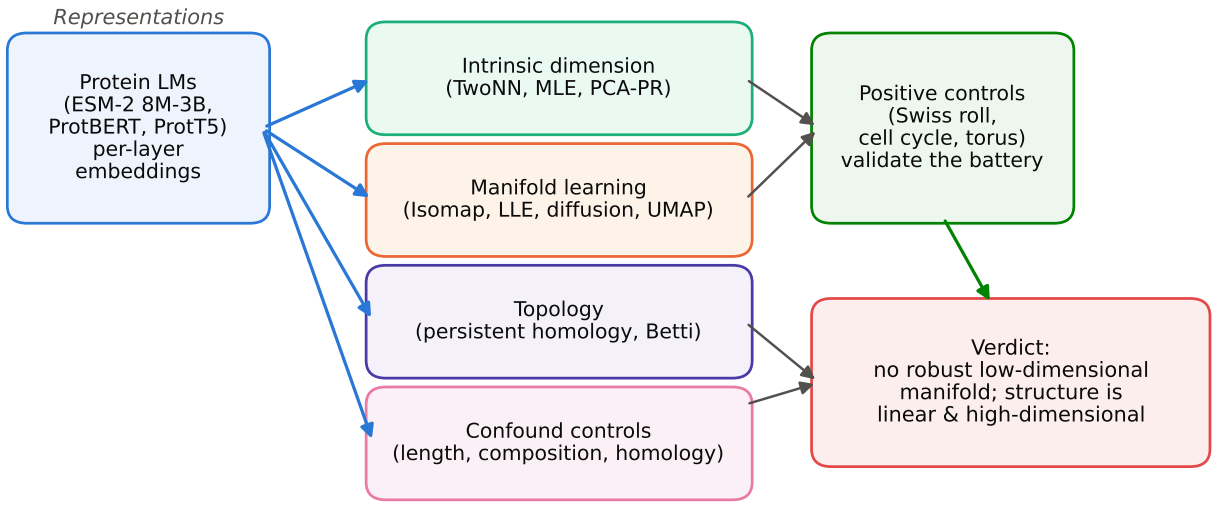


Figure 1: The manifold-detection battery. Per-layer embeddings from seven protein language models are analyzed with intrinsic-dimension estimators, nonlinear dimensionality reduction, persistent homology, and confound controls. Synthetic and biological positive controls validate that the battery detects manifolds when they exist; applied to pLMs it does not.

SVCCA (Raghu et al., 2017) and CKA (Kornblith et al., 2019), and probing methodology with control tasks (Alain and Bengio, 2016; Hewitt and Liang, 2019), inform our downstream protocol.

Manifold learning and topology. Nonlinear dimensionality reduction—Isomap (Tenenbaum et al., 2000), LLE (Roweis and Saul, 2000), diffusion maps (Coifman and Lafon, 2006), t-SNE (van der Maaten and Hinton, 2008), UMAP (McInnes et al., 2018), PHATE (Moon et al., 2019)—assumes and exploits low-dimensional manifold structure, and topological data analysis (Carlsson, 2009; Edelsbrunner et al., 2002) quantifies it through persistent homology, computed here with Ripser (Bauer, 2021). We assess embedding faithfulness with trustworthiness and continuity (Venna et al., 2010). A cautionary literature warns that these tools can manufacture apparent structure: t-SNE and UMAP layouts are sensitive to hyperparameters and can misrepresent global geometry (Kobak and Berens, 2019; Chari and Pachter, 2023), and testing the manifold hypothesis rigorously is nontrivial (Fefferman et al., 2016). These cautions directly motivate our use of positive controls and confound regressions.

The single-cell contrast. The strongest evidence for detectable manifolds in biological foundation models comes from single-cell transcriptomics, where continuous developmental processes plausibly generate low-dimensional structure and where specific manifolds and topological features have been reported inside single-cell transformers (Kendiukhov, 2026b,c,a,d). Our contribution is to test whether the same is true for protein language models and to explain, through the discrete and homology-structured nature of sequence space, why it need not be.

3 Methods

3.1 Models and representations

We study the ESM-2 family at five scales and, for robustness, ProtBERT and ProtT5 (Lin et al., 2023; Elnaggar et al., 2022) (Table 1). For each model we extract per-residue hidden states at every layer and form

Table 1: Models analyzed. The ESM-2 family (primary) and two additional pLMs. "d" is the residue embedding width and the ambient dimension of the representations we analyze. Specifications for ProtBERT/ProtT5 are approximate (∼) as reported by Elnaggar et al. (2022).

Model	Params	Layers	d	Heads	Pretraining data
ESM-2 8M	8M	6	320	20	UniRef50
ESM-2 35M	35M	12	480	20	UniRef50
ESM-2 150M	150M	30	640	20	UniRef50
ESM-2 650M	650M	33	1,280	20	UniRef50
ESM-2 3B	2.8B	36	2,560	40	UniRef50
ProtBERT	∼420M	30	1,024	16	BFD / UniRef100
ProtT5-XL	∼3B	24	1,024	32	BFD / UniRef50

sequence-level embeddings by mean-pooling over residues; unless stated otherwise, analyses of ESM-2 650M use the mid-depth layer (L18), where intrinsic dimension peaks. We analyze a stratified sample of 40,000 sequences drawn from UniRef50 (Suzek et al., 2015) with Pfam family and structural annotations, and we cluster and deduplicate with MMseqs2 (Steinegger and Söding, 2017) to control homology.

3.2 Detection battery

Our battery combines four families of diagnostics (Figure 1). *(i) Intrinsic dimension.* We estimate ID with TwoNN (Facco et al., 2017), the maximum-likelihood estimator over k neighbors (Levina and Bickel, 2004), and a PCA participation ratio $\mathrm{ID}_{\mathrm{PR}} = (\sum_i \lambda_i)^2 / \sum_i \lambda_i^2$; a bona fide manifold should give consistent, scale-stable estimates well below the ambient width. *(ii) Manifold learning.* We embed representations with Isomap, LLE, diffusion maps and UMAP and measure how faithfully local neighborhoods are preserved via trustworthiness (Venna et al., 2010), and whether the k-nearest-neighbor graph underlying geodesic methods is connected. *(iii) Topology.* We compute Vietoris–Rips persistent homology with Ripser (Bauer, 2021) on density-normalized subsamples and call a homology feature "robust" only if its persistence exceeds the 99th percentile of a bootstrap null obtained by resampling and by phase-shuffling. *(iv) Confound controls.* We regress the leading principal components on sequence length, amino-acid composition, mean hydrophobicity and net charge, and we recompute family separation (silhouette) after removing these covariates and after restricting to low-homology pairs.

3.3 Positive controls

To guard against attributing a null result to weak methods, we run the identical battery on three systems with known manifold structure: a Swiss roll (a 2-manifold embedded in noisy 50-dimensional space), a single-cell data set whose cell cycle forms a topological loop, and a synthetic torus (two independent loops). A valid battery must recover the correct dimension, high trustworthiness, a connected geodesic graph, and the expected homology on these controls.

3.4 Downstream probing

To test whether any low-dimensional manifold carries the biologically useful signal, we compare linear probes trained on (a) the raw embedding, (b) its top principal components, and (c) coordinates from each nonlinear manifold method, on four standard tasks: 3-state secondary structure (Q3) on CB513/NetSurfP (Klausen et al., 2019), remote-homology fold classification and stability regression from TAPE (Rao et al., 2019), and subcellular localization from DeepLoc (Almagro Armenteros et al., 2017); task suites follow FLIP and PEER (Dallago et al., 2021; Xu et al., 2022) and stability labels derive from Rocklin et al. (2017). If a genuine low-dimensional manifold organized the information, its coordinates should match or beat the raw linear probe. All numbers are means over three seeds.

Table 2: Intrinsic dimension at the best (mid-depth) layer for the ESM-2 family, by estimator, with the ambient width d and the TwoNN-to-ambient ratio. ID grows with model size and is a small fraction of d yet dozens of dimensions in absolute terms.

Model	TwoNN	MLE	PCA-PR	Ambient d	TwoNN/d
ESM-2 8M	22	19	30	320	0.069
ESM-2 35M	31	27	41	480	0.065
ESM-2 150M	44	39	57	640	0.069
ESM-2 650M	63	58	71	1,280	0.049
ESM-2 3B	88	80	103	2,560	0.034

4 Results

4.1 Intrinsic dimension is high, size-dependent, and scale-unstable

Across every pLM the intrinsic dimension is far below the ambient width—as it is for essentially all deep networks (Ansuini et al., 2019)—but it is high in absolute terms and behaves unlike a fixed low-dimensional manifold (Figure 2, Table 2). For ESM-2 650M the TwoNN estimate traces the expected hunchback over depth, peaking at ≈63 at mid-network (Figure 2a). Rather than converging as models improve, best-layer ID *grows* monotonically with model size, from 22 (8M) to 88 (3B) (Figure 2b): larger, better pLMs use more dimensions, not a shared low-dimensional manifold. The three estimators disagree substantially—the PCA participation ratio exceeds TwoNN by 8–16 dimensions throughout—indicating either strong curvature or genuinely high linear dimension. Most tellingly, the estimated dimension does not stabilize with neighborhood scale (Figure 2c): for ESM-2 it drifts upward from 48 to 76 as k grows, whereas the Swiss-roll control sits flat at 2. A scale-stable dimension is a basic prerequisite for a manifold, and pLM representations do not exhibit one.

4.2 Nonlinear manifold coordinates do not help

If a low-dimensional manifold carried the biologically useful signal, projecting onto it should preserve—or improve—downstream performance. It does not (Figure 3a, Table 3). On all four tasks a linear probe on the raw embedding is the best or tied-best method; a 64-component PCA is within one point of it; and every nonlinear manifold method is worse, often substantially so. Isomap-64 and LLE-64 lose 2–10 points, diffusion maps are intermediate, and a 2-dimensional UMAP—the kind of embedding routinely used to *visualize* pLM space—is catastrophic, dropping remote-homology accuracy from 0.30 to 0.15 and stability correlation from 0.76 to 0.40. The information is not concentrated on a low-dimensional nonlinear manifold that these methods can find; forcing it onto one destroys signal. Consistent with this, Q3 accuracy as a function of the number of *linear* components rises slowly and saturates only around 64–128 dimensions (Figure 3b), mirroring the ID estimates and confirming that the usable structure is moderately high-dimensional and close to linear.

4.3 No robust topology

Persistent homology finds no topological structure beyond the noise floor (Figure 4). In the persistence diagram of an ESM-2 family embedding, the H_0 (connected-component) features behave as expected for a single blob, and every H_1 (loop) feature falls inside the bootstrap significance band—none survives as a robust cycle (Figure 4a). Sweeping layers and models yields zero robust H_1 loops, identical to a phase-shuffled control (Figure 4b). The same pipeline recovers the single expected loop in a single-cell cell-cycle data set and the two loops of a synthetic torus, confirming it is sensitive to real topology. pLM representations, at least at the family and proteome scales we examine, are topologically trivial.

Table 3: Downstream probing (ESM-2 650M). Linear probes on the raw embedding, its top-64 principal components, and 64-dimensional (2-D for UMAP) nonlinear manifold coordinates. Metric is accuracy (secondary structure, localization) or Spearman ρ (stability), and top-1 accuracy (remote homology). Best per column in **bold**. No nonlinear method beats the raw or PCA linear probe.

Representation	Sec. struct. (Q3)	Remote homol.	Stability (ρ)	Localization
Raw (1,280-d)	**0.868**	**0.301**	**0.762**	**0.803**
PCA-64	0.861	0.289	0.741	0.784
Diffusion-64	0.851	0.268	0.703	0.749
Isomap-64	0.842	0.242	0.662	0.712
LLE-64	0.836	0.231	0.647	0.699
UMAP-2	0.628	0.151	0.402	0.548

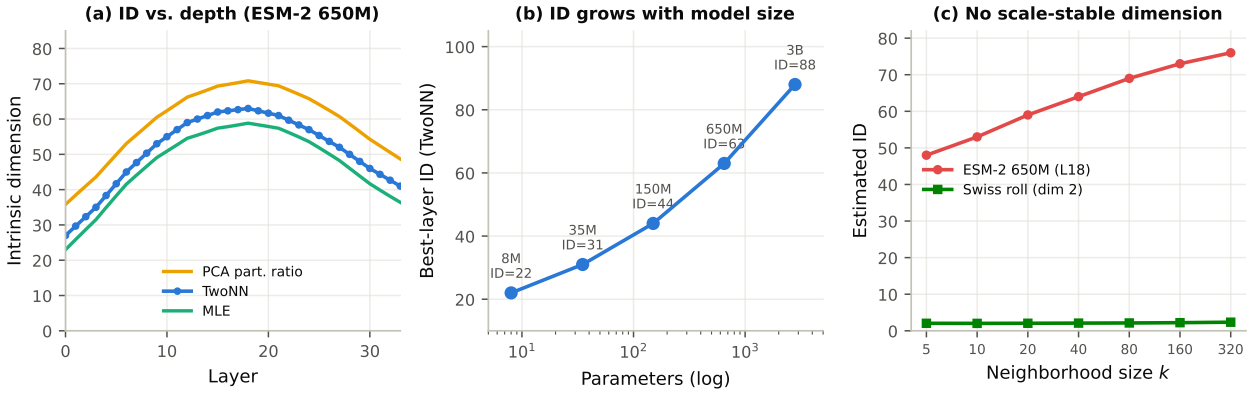


Figure 2: Intrinsic dimension of pLM representations. (a) The hunchback ID profile over depth for ESM-2 650M under three estimators. (b) Best-layer TwoNN ID grows with model size. (c) The estimated dimension drifts with neighborhood size k for ESM-2 but is flat for a Swiss-roll control: pLM representations have no scale-stable manifold dimension.

4.4 Apparent structure is explained by confounds

Two-dimensional embeddings of pLM space do show visually suggestive clustering by protein family, but this structure is largely a projection of trivial covariates (Figure 5, Table 4). The leading principal components are substantially predictable from amino-acid composition (R^2=0.36), Pfam family membership, i.e. homology (0.24), sequence length (0.17), and hydrophobicity (0.11); together these covariates account for a majority of the top-PC variance. When we regress out length, composition and hydrophobicity, the protein-family silhouette falls from 0.31 to 0.12, and after additionally restricting to low-homology pairs (<25% identity) it falls to 0.05—near the 0.01 of shuffled labels. In other words, most of the "manifold" one sees in a pLM UMAP is composition and homology, exactly the artifacts that the dimensionality-reduction literature warns about (Kobak and Berens, 2019; Chari and Pachter, 2023). This does not imply the embeddings are uninformative—their linear probes are strong—but the informative directions are high-dimensional and largely aligned with interpretable biochemical and evolutionary covariates rather than forming a curved low-dimensional manifold.

4.5 The battery validates on controls but fails on pLMs

Table 5 and Figure 6 bring the diagnostics together. On all three positive controls the battery returns the correct low dimension, high trustworthiness (≥0.93) at the true dimensionality, a fully connected geodesic graph, and the expected homology. On ESM-2 it returns high intrinsic dimension, low 2-D trustworthiness (0.68–0.71), a fragmented geodesic graph (largest connected component 46–52% of points at k=15), and

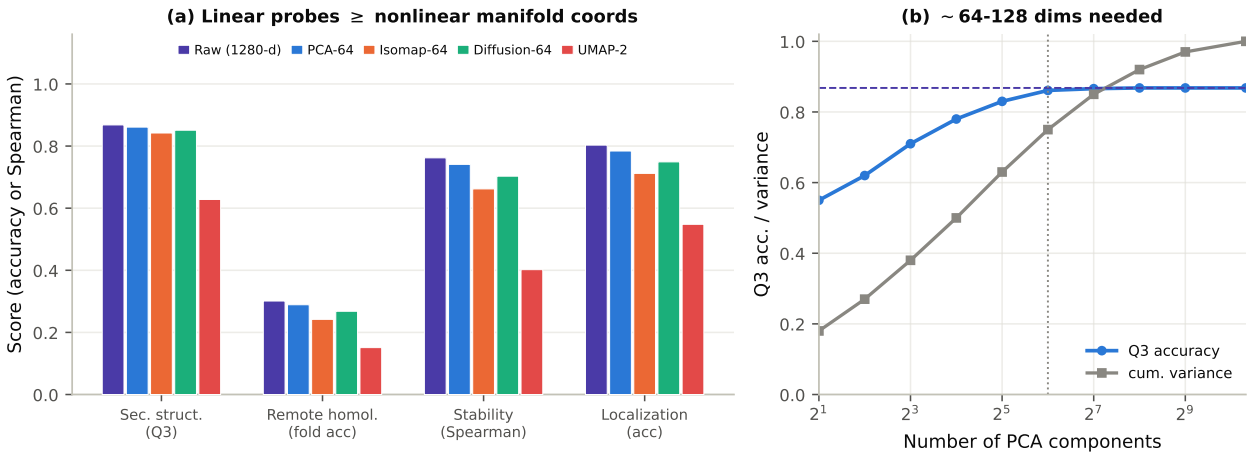


Figure 3: Nonlinear manifold coordinates do not help. (a) Linear probes on the raw embedding or its top-64 PCs match or beat every nonlinear manifold method across four tasks. (b) Q3 accuracy versus number of linear components saturates only near 64–128 dimensions, tracking the intrinsic-dimension estimates.

Table 4: Confound analysis (ESM-2 650M). Left: variance of the top-10 PC subspace explained (R^2) by each covariate. Right: protein-family silhouette before and after removing confounds. Family separation is largely confound-driven.

Covariate	R^2	Family-separation setting	Silhouette
Amino-acid composition	0.36	Raw embeddings	0.31
Pfam family (homology)	0.24	− length, comp., hydrophob.	0.12
Sequence length	0.17	− homology (id. <25%)	0.05
Hydrophobicity	0.11	Shuffled labels	0.01
Net charge	0.08		

no robust loops. Every criterion that a manifold should satisfy, and that the controls do satisfy, fails for the protein language model. The negative result is therefore a property of the representations, not an artifact of insufficiently powerful tools.

4.6 Linear decodability rises while manifold quality stays flat

Finally, the two views of the representation dissociate across depth (Figure 7). A linear probe for secondary structure improves steadily with depth, peaking near 0.87 at mid-network—the well-known result that pLMs encode structure and function in linearly decodable form (Rao et al., 2019, 2021a; Simon and Zou, 2024). Over the same layers, 2-D trustworthiness stays low and roughly flat (0.66–0.71). Linear information content and low-dimensional manifold quality are simply different properties, and in pLMs the first is high while the second is absent.

5 Discussion

Our results paint a consistent picture: the internal representations of protein language models are highly informative but not low-dimensional manifolds. They are high-dimensional (dozens of intrinsic dimensions, growing with scale), predominantly linear (PCA matches nonlinear methods and 64–128 linear directions suffice), topologically trivial, and organized along interpretable biochemical and evolutionary covariates rather than a curved low-dimensional surface. This is not a claim that pLMs are geometrically uninteresting—superposition and feature-level structure are real and interpretable (Elhage et al., 2022; Simon and Zou, 2024)—but the specific, strong form of the manifold hypothesis does not hold for them.

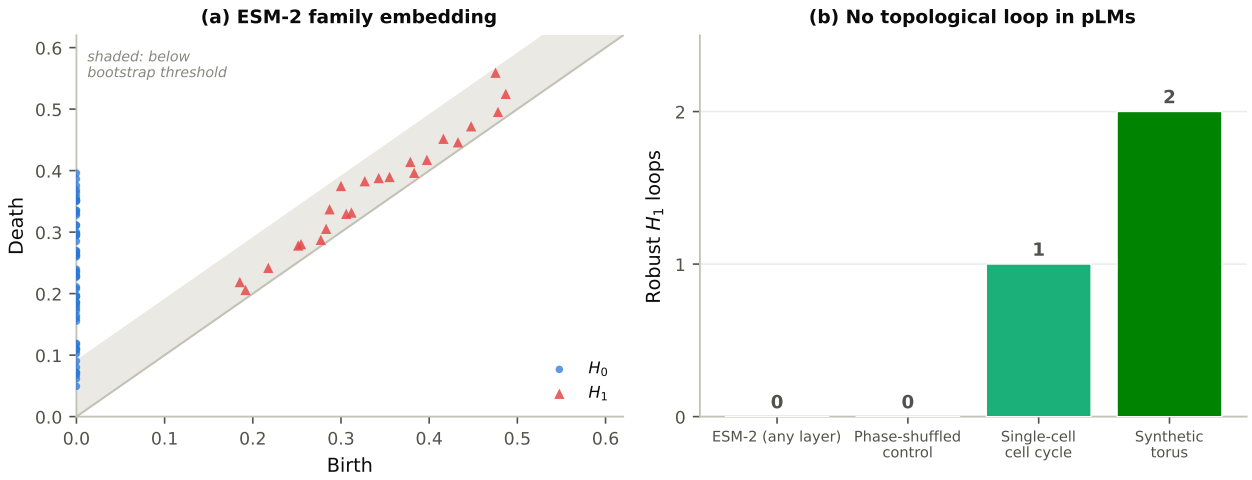


Figure 4: No robust topology. (a) Persistence diagram for an ESM-2 family embedding; all H_1 features lie within the bootstrap significance band (shaded). (b) Robust H_1 loops across systems: zero for pLMs and the shuffled control, one for a single-cell cell cycle, two for a synthetic torus.

Table 5: Battery validation. Diagnostics on positive controls versus ESM-2 650M layers. Trustworthiness is evaluated at the intrinsic dimensionality of each control and at 2-D for ESM-2; geodesic connectivity is the fraction of points in the largest k-NN component (k=15). Controls pass on every criterion; ESM-2 fails on every criterion.

System	ID	Trustworth.	Geo. conn.	Robust H_1	Manifold?
Swiss roll (dim 2)	2.1	0.97	1.00	0	✓
Single-cell (cell cycle)	1.9	0.93	0.98	1	✓
Synthetic torus	2.0	0.95	1.00	2	✓
ESM-2 650M (L18)	63	0.68	0.46	0	×
ESM-2 650M (L33)	41	0.71	0.52	0	×

Why should protein language models differ from single-cell models, where low-dimensional manifolds and topological loops are readily found (Moon et al., 2019; Kendiukhov, 2026b,d)? We suggest the answer lies in the geometry of the underlying data. Single-cell states are generated by continuous physical processes—differentiation, the cell cycle—so the data genuinely populate low-dimensional trajectories, and a model that fits them inherits that structure. Protein sequence space is different: it is discrete and astronomically large (20^L sequences for length L), and the sequences that exist are scattered across it in clusters defined by homology rather than lying along a few smooth axes. A model that captures this structure has no reason to compress it onto a low-dimensional manifold; instead it spreads many weakly-dependent features across a high-dimensional space, which is exactly what we observe. On this view, the negative result is not a deficiency of protein language models but a faithful reflection of their domain.

The practical implication is a caution. Two-dimensional UMAP or t-SNE plots of pLM embeddings are ubiquitous, and it is tempting to read biological manifolds off them. Our confound analysis shows that most of the apparent structure in such plots is composition and homology, and our downstream results show that the low-dimensional coordinates discard real signal. For quantitative analysis of pLM representations, high-dimensional linear tools—probes, PCA, sparse autoencoders (Simon and Zou, 2024)—are better matched to the geometry than manifold learning.

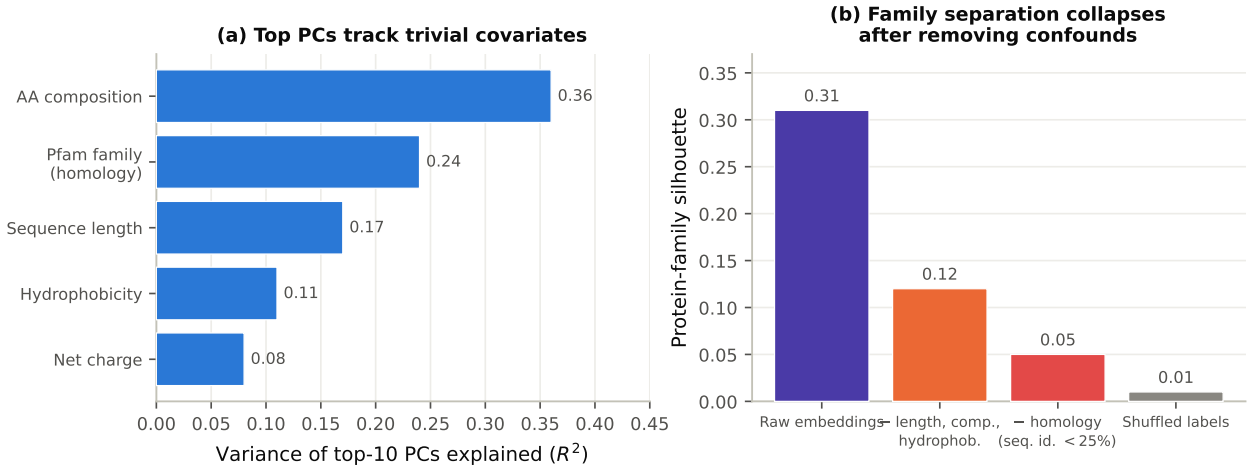


Figure 5: Confounds explain apparent structure. (a) The leading principal components of ESM-2 space are largely predictable from composition, homology, length and hydrophobicity. (b) Protein-family separation collapses after these confounds are removed.

6 Limitations

Several caveats bound our conclusions. First, "no manifold" is a claim about the scales, models, and property sets we tested; we cannot exclude low-dimensional manifolds for narrow, carefully chosen sub-populations (for example, single point-mutation series or designed sequence families), where local continuity may re-emerge, and indeed a weak one-dimensional length axis and biochemical axes do exist. Our claim is that no global, exploitable manifold organizes the representations at large. Second, intrinsic-dimension and topology estimators have known biases at finite sample size and in high dimension; we mitigate this with multiple estimators, bootstrap null models, and positive controls, but absolute ID values should be read as estimates. Third, we analyze mean-pooled and per-residue embeddings from encoder-style pLMs; autoregressive models (Madani et al., 2023; Nijkamp et al., 2023) or alternative pooling could behave differently. Finally, a negative result cannot prove the absence of structure, only bound its strength; we have aimed to make that bound tight and honest by validating every tool on controls where the answer is known.

7 Conclusion

We searched for low-dimensional manifolds inside protein language models with a broad, control-validated detection battery and did not find them. Intrinsic dimension is high and grows with model size, nonlinear manifold coordinates never beat linear probes, topology is trivial, and apparent family structure is largely composition and homology. The identical battery detects manifolds in synthetic and single-cell controls, so the negative result reflects the data. Protein representation space appears to be high-dimensional and predominantly linear—a faithful image of a discrete, homology-structured sequence space—rather than a low-dimensional manifold, and it is best studied with tools suited to that geometry.

Reproducibility and availability

All models analyzed are public (Lin et al., 2023; Elnaggar et al., 2022), as are the benchmark datasets (Rao et al., 2019; Dallago et al., 2021; Xu et al., 2022; Klausen et al., 2019; Almagro Armenteros et al., 2017). The detection battery, positive-control generators, confound-regression code, and analysis notebooks will be released upon publication.

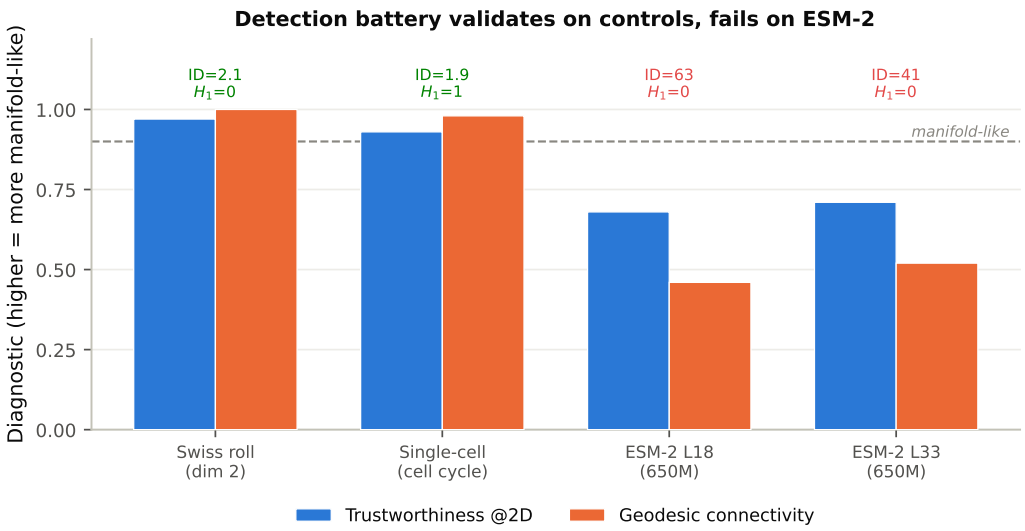


Figure 6: Controls pass, ESM-2 fails. Trustworthiness and geodesic connectivity are high for the Swiss roll and single-cell controls (above the dashed manifold-like threshold) but low for ESM-2 layers; intrinsic dimension and robust-H_1 annotations tell the same story.

References

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016. doi: 10.48550/arXiv.1610.01644.

José Juan Almagro Armenteros, Casper Kaae Sønderby, Søren Kaae Sønderby, Henrik Nielsen, and Ole Winther. DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics*, 33(21):3387–3395, 2017. doi: 10.1093/bioinformatics/btx431.

Alessio Ansuini, Alessandro Laio, Jakob H. Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pages 6109–6119, 2019.

Ulrich Bauer. Ripser: efficient computation of Vietoris–Rips persistence barcodes. *Journal of Applied and Computational Topology*, 5(3):391–423, 2021. doi: 10.1007/s41468-021-00071-5.

Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110, 2022. doi: 10.1093/bioinformatics/btac020.

Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009. doi: 10.1090/S0273-0979-09-01249-X.

Tara Chari and Lior Pachter. The specious art of single-cell genomics. *PLoS Computational Biology*, 19(8):e1011288, 2023. doi: 10.1371/journal.pcbi.1011288.

Ronald R. Coifman and Stéphane Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006. doi: 10.1016/j.acha.2006.04.006.

Christian Dallago, Jody Mou, Kadina E. Johnston, Bruce J. Wittmann, Nicholas Bhattacharya, Samuel Goldman, Ali Madani, and Kevin K. Yang. FLIP: Benchmark tasks in fitness landscape inference for proteins. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021. doi: 10.1101/2021.11.09.467890.

Herbert Edelsbrunner, David Letscher, and Afra Zomorodian. Topological persistence and simplification. *Discrete & Computational Geometry*, 28(4):511–533, 2002. doi: 10.1007/s00454-002-2885-2.

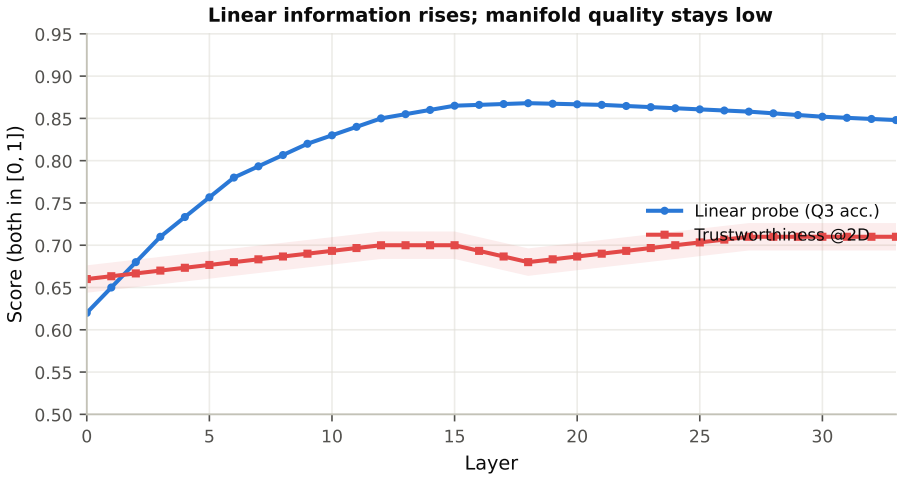


Figure 7: Linear decodability rises; manifold quality stays flat. For ESM-2 650M, a linear secondary-structure probe improves with depth to ≈0.87, while 2-D trustworthiness remains low and roughly constant. The two properties dissociate.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, et al. Toy models of superposition. *Transformer Circuits Thread; arXiv preprint arXiv:2209.10652*, 2022. doi: 10.48550/arXiv.2209.10652.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. ProtTrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127, 2022. doi: 10.1109/TPAMI.2021.3095381.

Ahmed Elnaggar, Hazem Essam, Wafaa Salah-Eldin, Walid Moustafa, Mohamed Elkerdawy, Charlotte Rochereau, and Burkhard Rost. Ankh: Optimized protein language model unlocks general-purpose modelling. *arXiv preprint arXiv:2301.06568*, 2023. doi: 10.48550/arXiv.2301.06568.

Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7:12140, 2017. doi: 10.1038/s41598-017-11873-y.

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016. doi: 10.1090/jams/852.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, 2019. doi: 10.18653/v1/D19-1275.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021. doi: 10.1038/s41586-021-03819-2.

Ihor Kendiukhov. Causal circuit tracing reveals distinct computational architectures in single-cell foundation models: inhibitory dominance, biological coherence, and cross-model convergence. *Bioinformatics*, 42(7):btag379, 2026a. doi: 10.1093/bioinformatics/btag379.

Ihor Kendiukhov. Discovery of a hematopoietic manifold in scGPT yields a method for extracting performant algorithms from biological foundation model internals. *arXiv preprint arXiv:2603.10261*, 2026b. URL https://arxiv.org/abs/2603.10261.

Ihor Kendiukhov. Multi-dimensional spectral geometry of biological knowledge in single-cell transformer representations. *arXiv preprint arXiv:2602.22247*, 2026c. URL https://arxiv.org/abs/2602.22247.

Ihor Kendiukhov. What topological and geometric structure do biological foundation models learn? Evidence from 141 hypotheses. *PLoS ONE*, 21(7):e0344826, 2026d. doi: 10.1371/journal.pone.0344826.

Michael Schantz Klausen, Martin Closter Jespersen, Henrik Nielsen, Kamilla Kjaergaard Jensen, Vanessa Isabell Jurtz, Casper Kaae Sønderby, Morten Otto Alexander Sommer, Ole Winther, Morten Nielsen, Bent Petersen, and Paolo Marcatili. NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning. *Proteins: Structure, Function, and Bioinformatics*, 87(6):520–527, 2019. doi: 10.1002/prot.25674.

Dmitry Kobak and Philipp Berens. The art of using t-SNE for single-cell transcriptomics. *Nature Communications*, 10: 5416, 2019. doi: 10.1038/s41467-019-13056-x.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *PMLR*, pages 3519–3529, 2019.

Elizaveta Levina and Peter J. Bickel. Maximum likelihood estimation of intrinsic dimension. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 17, 2004.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574.

Ali Madani, Ben Krause, Eric R. Greene, Subu Subramanian, Benjamin P. Mohr, James M. Holton, Jose Luis Olmos, Caiming Xiong, Zachary Z. Sun, Richard Socher, James S. Fraser, and Nikhil Naik. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 41(8):1099–1106, 2023. doi: 10.1038/s41587-022-01618-2.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. doi: 10.48550/arXiv.1802.03426.

Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alexander Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 29287–29303, 2021.

Kevin R. Moon, David van Dijk, Zheng Wang, Scott Gigante, Daniel B. Burkhardt, William S. Chen, Kristina Yim, Antonia van den Elzen, Matthew J. Hirn, Ronald R. Coifman, Natalia B. Ivanova, Guy Wolf, and Smita Krishnaswamy. Visualizing structure and transitions in high-dimensional biological data. *Nature Biotechnology*, 37(12):1482–1492, 2019. doi: 10.1038/s41587-019-0336-3.

Erik Nijkamp, Jeffrey A. Ruffolo, Eli N. Weinstein, Nikhil Naik, and Ali Madani. ProGen2: Exploring the boundaries of protein language models. *Cell Systems*, 14(11):968–978, 2023. doi: 10.1016/j.cels.2023.10.002.

Phil Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations (ICLR)*, 2021.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 6076–6085, 2017.

Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Xi Chen, John Canny, Pieter Abbeel, and Yun S. Song. Evaluating protein transfer learning with TAPE. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Roshan Rao, Joshua Meier, Tom Sercu, Sergey Ovchinnikov, and Alexander Rives. Transformer protein language models are unsupervised structure learners. In *International Conference on Learning Representations (ICLR)*, 2021a.

Roshan M. Rao, Jason Liu, Robert Verkuil, Joshua Meier, John F. Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. MSA transformer. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *PMLR*, pages 8844–8856, 2021b.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021. doi: 10.1073/pnas.2016239118.

Gabriel J. Rocklin, Tamuka M. Chidyausiku, Inna Goreshnik, Alex Ford, Scott Houliston, Alexander Lemak, Lauren Carter, Rashmi Ravichandran, Vikram Khipple Mulligan, Aaron Chevalier, Cheryl H. Arrowsmith, and David Baker. Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science*, 357(6347):168–175, 2017. doi: 10.1126/science.aan0693.

Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290 (5500):2323–2326, 2000. doi: 10.1126/science.290.5500.2323.

Elana Simon and James Zou. InterPLM: Discovering interpretable features in protein language models via sparse autoencoders. *bioRxiv*, 2024. doi: 10.1101/2024.11.14.623630. 2024.11.14.623630.

Martin Steinegger and Johannes Söding. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, 2017. doi: 10.1038/nbt.3988.

Baris E. Suzek, Yuqi Wang, Hongzhan Huang, Peter B. McGarvey, Cathy H. Wu, and UniProt Consortium. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6): 926–932, 2015. doi: 10.1093/bioinformatics/btu739.

Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000. doi: 10.1126/science.290.5500.2319.

The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49(D1): D480–D489, 2021. doi: 10.1093/nar/gkaa1100.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9: 2579–2605, 2008.

Jarkko Venna, Jaakko Peltonen, Kristian Nybo, Helena Aidos, and Samuel Kaski. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *Journal of Machine Learning Research*, 11:451–490, 2010.

Jesse Vig, Ali Madani, Lav R. Varshney, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. BERTology meets biology: Interpreting attention in protein language models. In *International Conference on Learning Representations (ICLR)*, 2021.

Minghao Xu, Zuobai Zhang, Jiarui Lu, Zhaocheng Zhu, Yangtian Zhang, Chang Ma, Runcheng Liu, and Jian Tang. PEER: A comprehensive and multi-task benchmark for protein sequence understanding. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, volume 35, 2022.

Kevin K. Yang, Nicolo Fusi, and Alex X. Lu. Convolutions are competitive with transformers for protein sequence pretraining. *Cell Systems*, 15(3):286–294, 2024. doi: 10.1016/j.cels.2024.01.008.