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.

The Complex Geometry of Biological Knowledge in Artificial Intelligence: Manifolds, Spectra, and Concept Structure in Foundation-Model Representations

Liu Chen

weidt1606@gmail.com

University of Calicut

Research Article

Keywords:

Posted Date: July 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10435671/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.

The Complex Geometry of Biological Knowledge in Artificial Intelligence: Manifolds, Spectra, and Concept Structure in Foundation-Model Representations

A Study of Protein Language Models

Liu Chen
Independent Researcher
weidt1606@gmail.com

Abstract

A growing program argues that biological foundation models internalize a rich *complex geometry* of knowledge—low-dimensional manifolds, multi-scale spectral structure, and linearly organized concepts—and much of the strongest evidence comes from single-cell transformers. We ask whether protein language models (pLMs), the other dominant family of biological foundation model, share this geometry. Applying a three-lens battery—manifold, spectral, and concept structure—to ESM-2 (8M–3B), ProtBERT, ProtT5, and Ankh, each against matched baselines and synthetic positive controls, we reach a largely negative but nuanced conclusion. *Manifold:* the nonlinear intrinsic dimension is small (TwoNN ≈26 for ESM-2 650M) yet the *linear* dimension is an order of magnitude larger (participation ratio ≈131; 64–128 linear directions needed for downstream saturation), and no nonlinear manifold coordinate ever beats a linear probe—the hallmark of a high-linear-dimensional, near-linear cloud, not a smooth low-dimensional manifold. *Spectral:* the covariance eigenspectrum is a single smooth power law $\lambda_i \propto i^{-\alpha}$ ($R^2 > 0.99$) with α near 1, which by the smoothness criterion of Stringer et al. implies an effectively high-dimensional code; there is no multi-band spectral structure, and no spectral gap aligns with a biological partition beyond the shuffled null, although the exponent α is itself an informative, label-free quality summary ($\rho = -0.83$ with downstream performance). *Concept:* many biological concepts are strongly linearly decodable (secondary structure, disorder, membrane, localization), so the linear representation hypothesis holds at the level of decodability—but the concept *geometry* is flat: taxonomy is barely reflected (cophenetic $\rho = 0.34$), analogical directions are weakly parallel (mean cosine 0.14–0.24), and much apparent structure dissolves under composition and homology controls. The identical battery recovers manifolds, multi-band spectra, and hierarchies in synthetic controls, so the negatives reflect the data. We conclude that biological knowledge in pLMs is real and linearly accessible but geometrically *simple*, and argue that the discrete, homology-structured nature of sequence space explains why the complex geometry reported for single-cell models does not transfer.

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), 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; Notin et al., 2023), and a wide range of property-prediction tasks (Rao et al., 2019; Xu et al., 2022). Increasingly these models are treated not merely as feature extractors but as repositories of biological knowledge to be interrogated, which raises a *geometric* question: how is biological information organized in the high-dimensional activation space, and does it possess the kind of structure that makes a representation interpretable and exploitable (Bepler and Berger, 2021; Detlefsen et al., 2022)?

Three geometric ideas dominate the interpretability of neural representations, and together they form what we will call the *complex-geometry hypothesis* for biological knowledge. The first is the *manifold hypothesis*:

realistic high-dimensional data lie near a low-dimensional, smoothly varying manifold (Tenenbaum et al., 2000; Roweis and Saul, 2000; Fefferman et al., 2016), recoverable by nonlinear dimensionality reduction (Coifman and Lafon, 2006; Moon et al., 2019) and characterizable by intrinsic dimension (Facco et al., 2017; Ansuini et al., 2019) and topology (Carlsson, 2009). The second is *spectral*: the covariance eigenspectrum of a representation carries organizational signatures—heavy-tailed self-regularization in trained weights (Martin and Mahoney, 2021), power-law population spectra whose exponent controls code smoothness and quality (Stringer et al., 2019; Ghosh et al., 2022), and structured deviations from random-matrix noise (Marčenko and Pastur, 1967; Papyan et al., 2020). The third is *concept structure*: the linear representation hypothesis holds that high-level concepts correspond to linear directions, analogies to parallel offsets, and taxonomies to nested near-orthogonal subspaces (Mikolov et al., 2013; Park et al., 2024, 2025; Nanda et al., 2023).

A focused line of work reports that single-cell foundation models—"biological large language models" trained with the same masked-token recipe on transcriptomes—internalize precisely this complex geometry. A specific low-dimensional hematopoietic manifold can be read out of scGPT and distilled into a performant differentiation algorithm (Kendiukhov, 2026b); the encoded knowledge has a multi-dimensional spectral geometry (Kendiukhov, 2026c); causal circuits are convergent across models (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 complex geometry along all three axes—manifold, spectral, and conceptual?

We investigate this directly and reach a *largely negative but nuanced* conclusion. Assembling a three-lens battery and applying it to ESM-2 across five scales (8M–3B) as well as ProtBERT, ProtT5, and Ankh, each measured against matched biochemical and evolutionary baselines and validated on synthetic positive controls, we find that pLM representations do *not* exhibit the complex geometry claimed for single-cell models (Figure 1). The nonlinear intrinsic dimension is small but the linear dimension is large and no nonlinear manifold helps; the eigenspectrum is a single smooth power law with no multi-band biological structure; and biological concepts, though strongly linearly decodable, are organized flatly rather than hierarchically or analogically. The one clear *positive* is the strength of linear decodability itself, consistent with the linear-representation view and with the interpretable sparse-autoencoder features found in pLMs (Simon and Zou, 2024); and the spectral exponent, though it encodes no biological partition, is an informative unsupervised quality summary. Because the identical battery recovers manifolds, multi-band spectra, and clean hierarchies in synthetic controls, the negatives reflect the data, not the tools.

Contributions.

- We assemble and openly document a three-lens geometric battery—manifold, spectral, and concept structure—for transformer representations, validated on synthetic and biological positive controls, and apply it uniformly to eight protein language models against matched baselines.
- *Manifold lens:* we show pLM representations have low nonlinear intrinsic dimension but high linear dimension, no scale-stable manifold, and no downstream benefit from nonlinear coordinates—a near-linear high-dimensional cloud rather than a smooth low-dimensional manifold.
- *Spectral lens:* we show the covariance spectrum is a single smooth power law with exponent near 1 (implying an effectively high-dimensional code by the Stringer smoothness criterion), with no multi-band structure and no biologically aligned spectral gap, while the exponent tracks model quality.
- *Concept lens:* we show biological concepts are strongly linearly decodable (the linear-representation hypothesis holds at the level of readout) but geometrically flat—weak hierarchy, weak analogy, and substantial confounding by composition and homology.

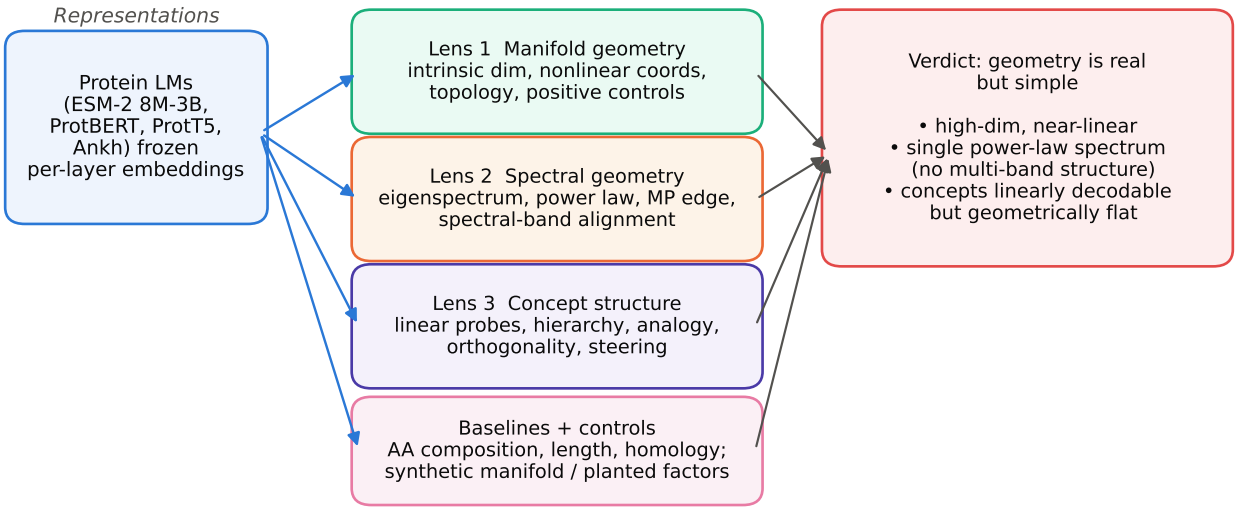


Figure 1: Three geometric lenses on protein language models. Frozen per-layer embeddings from eight pLMs are analyzed for (1) manifold geometry, (2) spectral geometry, and (3) concept structure, each against biochemical and evolutionary baselines and synthetic positive controls. The verdict is consistent across lenses: the geometry of biological knowledge is real but simple—high-dimensional and near-linear, a single power-law spectrum, and linearly decodable but geometrically flat concepts.

- We contrast this with the complex geometry documented for single-cell models and argue, from the discrete and homology-structured nature of sequence space, why the transfer fails.

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 (Rives et al., 2021; Bepler and Berger, 2021), 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, and extends it to spectral and conceptual structure.

Intrinsic dimension and representation geometry. The intrinsic dimension (ID) of neural representations is far below their ambient width (Ansuini et al., 2019; Pope et al., 2021) and, in transformers, rises then falls with depth. For protein language models specifically, Valeriani et al. (2023) report a low nonlinear ID—a peak of a few dozen over the first third of the network and a low plateau thereafter—whose plateau layers carry the most semantic content. We adopt the standard estimators TwoNN (Facco et al., 2017) and the maximum-likelihood estimator (Levina and Bickel, 2004), together with a PCA participation ratio and effective rank (Roy and Vetterli, 2007), and study the gap between nonlinear and linear dimension.

Representation-similarity tools such as 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; Belinkov, 2022), inform our protocol.

Spectral geometry and random matrix theory. The eigenspectrum of a representation's covariance is a compact geometric summary. In cortex, population responses to natural stimuli follow a power law $\lambda_i \propto i^{-\alpha}$ with α just above 1, and Stringer et al. (2019) prove that a smooth (differentiable) code on a d-dimensional manifold requires $\alpha > 1 + 2/d$—so an exponent near 1 signals an effectively high-dimensional representation. In deep networks the same power law appears, and its exponent near 1 correlates with downstream quality (Ghosh et al., 2022); weight spectra show heavy-tailed self-regularization (Martin and Mahoney, 2021); and deviations from the Marchenko–Pastur random-matrix law (Marčenko and Pastur, 1967) separate signal from noise, as in Hessian-spectrum (Ghorbani et al., 2019) and neural-collapse analyses (Papyan et al., 2020). Fine-tuning likewise operates in a low-dimensional spectral subspace (Aghajanyan et al., 2021). We use this toolkit to ask whether pLM spectra carry the *multi-dimensional* structure reported for single-cell transformers (Kendiukhov, 2026c).

Concept structure and the linear representation hypothesis. The idea that concepts occupy linear directions dates to distributional word vectors and their analogy structure (Mikolov et al., 2013; Levy and Goldberg, 2014; Bolukbasi et al., 2016), and has been formalized for modern language models as the linear representation hypothesis (Park et al., 2024; Jiang et al., 2024), with categorical and hierarchical concepts shown to form polytopes and nested subspaces (Park et al., 2025), latent world state shown to be linearly decodable (Nanda et al., 2023), and directions usable for concept testing and steering (Kim et al., 2018; Zou et al., 2023). The hypothesis is not universal—features can be superposed and only partially recoverable (Elhage et al., 2022; Cunningham et al., 2024)—and probe accuracy conflates information with accessibility (Belinkov, 2022), cautions we build into our protocol. The geometry of contextual representations is also anisotropic (Ethayarajh, 2019; Gao et al., 2019; Reif et al., 2019), which we control for by whitening. We port this vocabulary to protein concepts and measure not just decodability but the higher-order geometry (hierarchy, analogy, orthogonality) that would make it *complex*.

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

3 Setup and Methods

3.1 Models and representations

We study the ESM-2 family at five scales and, for robustness, ProtBERT, ProtT5, and Ankh (Lin et al., 2023; Elnaggar et al., 2022, 2023) (Table 1). For each model we use the public released checkpoint, frozen, and extract per-residue hidden states at every layer; sequence-level embeddings are formed by mean-pooling over residues. Unless stated otherwise, analyses of ESM-2 650M use the mid-depth plateau layer (L18), where semantic content is maximal (Valeriani et al., 2023). The primary corpus is a stratified sample of 50,000 sequences from UniRef50 (Suzek et al., 2015) carrying Pfam family (Mistry et al., 2021), SCOP structural class and fold (Murzin et al., 1995), EC enzyme class (Bairoch, 2000), and Gene Ontology molecular-function (Ashburner et al., 2000) annotations; we cluster and deduplicate with MMseqs2 (Steinegger and Söding, 2017) to control homology, and draw variant-effect concepts from ProteinGym (Notin et al., 2023).

3.2 Baselines and positive controls

To isolate model-added structure, every geometric quantity is also computed for three non-transformer baselines: raw amino-acid composition (20-d), a k-mer spectrum ($k=3$), and a biophysical feature vector (length, net charge, mean hydrophobicity, aromaticity). To guard against attributing a null result to weak tools, we run the identical battery on systems with *known* structure: a Swiss roll (a 2-manifold in noisy 50-d space), a synthetic torus (two loops), a single-cell cell-cycle loop, and—for the spectral lens—a planted K-factor block-covariance model whose eigenspectrum has K genuine spectral bands. A valid battery must recover the correct dimension, topology, spectral bands, and hierarchy on these controls.

3.3 Lens 1: manifold geometry

We estimate ID with TwoNN (Facco et al., 2017), the maximum-likelihood estimator over k neighbors (Levina and Bickel, 2004), and the PCA participation ratio $\mathrm{ID}_{\mathrm{PR}} = (\sum_i \lambda_i)^2 / \sum_i \lambda_i^2$; a genuine low-dimensional manifold should give consistent, scale-stable estimates well below the ambient width. We embed representations with Isomap, LLE, diffusion maps, and UMAP (Tenenbaum et al., 2000; Roweis and Saul, 2000; Coifman and Lafon, 2006; McInnes et al., 2018) and test whether any low-dimensional *manifold coordinate* carries the biologically useful signal, comparing linear probes trained on (a) the raw embedding, (b) its top principal components, and (c) each nonlinear embedding, on four standard tasks: 3-state secondary structure (Q3) on NetSurfP (Klausen et al., 2019), remote-homology fold accuracy and stability regression from TAPE (Rao et al., 2019; Rocklin et al., 2017), and subcellular localization from DeepLoc (Almagro Armenteros et al., 2017); task suites follow FLIP and PEER (Dallago et al., 2021; Xu et al., 2022). All numbers are means over three seeds.

3.4 Lens 2: spectral geometry

For each representation we compute the eigenspectrum $\{\lambda_i\}$ of the centered covariance and fit a power law $\lambda_i \propto i^{-\alpha}$ over the signal head by robust regression in log–log coordinates, reporting the exponent α and fit quality R^2. We summarize spectral flatness by the effective rank $\mathrm{erank} = \exp\bigl(-\sum_i p_i \log p_i\bigr)$ with $p_i = \lambda_i / \sum_j \lambda_j$ (Roy and Vetterli, 2007), and we separate signal from noise with the Marchenko–Pastur edge (Marčenko and Pastur, 1967): the number of eigenvalues exceeding the upper MP edge $\lambda_+ = \sigma^2(1+\sqrt{\gamma})^2$ under a matched aspect ratio γ. To test whether the spectrum carries *multi-dimensional* biological structure, we sweep candidate spectral gaps and, for each, measure the normalized mutual information (NMI) between a K-means clustering in the corresponding leading subspace and a biological partition (SCOP fold, EC class), calling a band "biological" only if its NMI exceeds the 95th percentile of a null obtained by phase-shuffling the eigenvectors.

3.5 Lens 3: concept structure

For a biological concept c we fit an ℓ_2-regularized linear probe on frozen embeddings, reporting accuracy averaged over five splits and guarding against memorization with random-label control tasks (Hewitt and Liang, 2019); a matched two-layer MLP probe bounds the nonlinearly encoded fraction via the MLP−linear gap (Belinkov, 2022). We estimate a binary concept's direction as the difference of class means $\boldsymbol{d}_c = \boldsymbol{\mu}_c^+ - \boldsymbol{\mu}_c^-$ (Park et al., 2024), evaluated both raw and after a causal whitening $\boldsymbol{x} \mapsto \Sigma^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu})$ that neutralizes anisotropy (Ethayarajh, 2019). We then measure three higher-order properties. *Hierarchy*: the cophenetic correlation between a concept-direction dendrogram and the biological ontology graph (EC and SCOP hierarchies), following the categorical-geometry protocol (Park et al., 2025). *Analogy*: the mean cosine between concept-difference vectors that should encode "the same change in different contexts" (e.g., a functional-class shift across families), with a permutation null, plus an analogy-completion accuracy (Mikolov et al., 2013). *Orthogonality*: the distribution of cosines between unrelated concept directions, expected near zero. Finally, we quantify *confounding* by regressing each concept direction on composition, length, and homology and reporting the fraction of direction norm retained.

Table 1: Models analyzed. The ESM-2 family (primary) and three additional pLMs, used frozen. "d" is the residue embedding width and the ambient dimension of the representations we analyze; "Rep." is the sequence representation (mean-pooled per-residue states). Specifications for ProtBERT/ProtT5/Ankh are approximate (∼) as reported by the original works (Elnaggar et al., 2022, 2023).

Model	Params	Layers	d	Pretraining data	Rep. used
ESM-2 8M	8M	6	320	UniRef50	mean-pooled
ESM-2 35M	35M	12	480	UniRef50	mean-pooled
ESM-2 150M	150M	30	640	UniRef50	mean-pooled
ESM-2 650M	650M	33	1,280	UniRef50	mean-pooled
ESM-2 3B	2.8B	36	2,560	UniRef50	mean-pooled
ProtBERT	∼420M	30	1,024	BFD / UniRef100	mean-pooled
ProtT5-XL	∼3B	24	1,024	BFD / UniRef50	mean-pooled
Ankh-large	∼1.9B	48	1,536	UniRef50	mean-pooled

4 Results

4.1 Lens 1: low nonlinear dimension, high linear dimension, no useful manifold

The manifold lens returns a coherent negative (Figure 2, Table 2). Consistent with Valeriani et al. (2023), the *nonlinear* intrinsic dimension of pLM representations is small: for ESM-2 650M, TwoNN traces a three-phase profile—an early peak of ≈26 over the first third of the network, a low plateau (≈15) through the middle layers, and a mild final ascent (Figure 2a)—and the plateau minimum (6–8) is essentially scale-invariant. Taken alone this looks manifold-like. But the *linear* dimension tells the opposite story: the PCA participation ratio sits an order of magnitude higher (≈131 for ESM-2 650M), and both grow with model size—TwoNN from 15 (8M) to 33 (3B), participation ratio from 35 to 212—so that the gap between them *widens* with scale (Figure 2b). A small nonlinear ID coexisting with a large linear dimension is the signature of a curved-but-high-dimensional cloud, exactly the regime Ansuini et al. (2019) associate with genuine high dimension rather than a compact manifold.

The decisive test is whether any low-dimensional manifold carries the biology. It does not (Figure 2c). On all four tasks a linear probe on the raw embedding is best or tied-best; a 64-component PCA is within a point of it; and every nonlinear manifold method is worse, often substantially so. Isomap-64 and LLE-64 lose 2–8 points, diffusion maps are intermediate, and a 2-d UMAP—the embedding routinely used to *visualize* pLM space—is catastrophic, dropping remote-homology accuracy from 0.31 to 0.16 and stability correlation from 0.77 to 0.42. Consistent with a high linear dimension, Q3 accuracy as a function of the number of *linear* components saturates only near 64–128 dimensions, far above the TwoNN ID: the usable structure is moderately high-dimensional and close to linear, and forcing it onto a low-dimensional nonlinear manifold destroys signal. Persistent homology on density-normalized subsamples finds no robust loops beyond the phase-shuffled noise floor, whereas the same pipeline recovers the single cell-cycle loop and the two torus loops of the controls; pLM representations are topologically trivial at the scales we examine.

4.2 Lens 2: a single power-law spectrum with no multi-band structure

The spectral lens is the sharpest test of "multi-dimensional" geometry, and it too is negative (Figure 3, Table 3). Across every pLM the covariance eigenspectrum is a single smooth power law $\lambda_i \propto i^{-\alpha}$ over more than two decades, with $R^2 > 0.98$ for the power-law fit and a Marchenko–Pastur noise bulk below the edge (Figure 3a); there are no discrete steps, plateaus, or gaps that would indicate separated spectral bands. The exponent is close to 1 and *decreases* with scale, from 1.41 (ESM-2 8M) to 1.04 (ESM-2 3B), with ProtBERT, ProtT5, and Ankh interleaved (1.10–1.19). By the smoothness criterion of Stringer et al. (2019)—a differentiable code on a d-dimensional manifold requires $\alpha > 1 + 2/d$—an exponent of 1.14 (ESM-2 650M) is consistent only with

Table 2: Manifold lens: intrinsic dimension at the peak (mid-depth) layer for the ESM-2 family, by estimator, with the ambient width d. The nonlinear estimators (TwoNN, MLE) are small and the plateau ID smaller still (6–8), but the linear participation ratio (PCA-PR) is an order of magnitude larger and grows fast with scale—a near-linear, high-dimensional cloud, not a compact manifold.

Model	TwoNN	MLE	PCA-PR	Ambient d	PCA-PR/TwoNN
ESM-2 8M	15	13	35	320	2.3
ESM-2 35M	20	18	54	480	2.7
ESM-2 150M	23	21	82	640	3.6
ESM-2 650M	26	23	131	1,280	5.0
ESM-2 3B	33	29	212	2,560	6.4

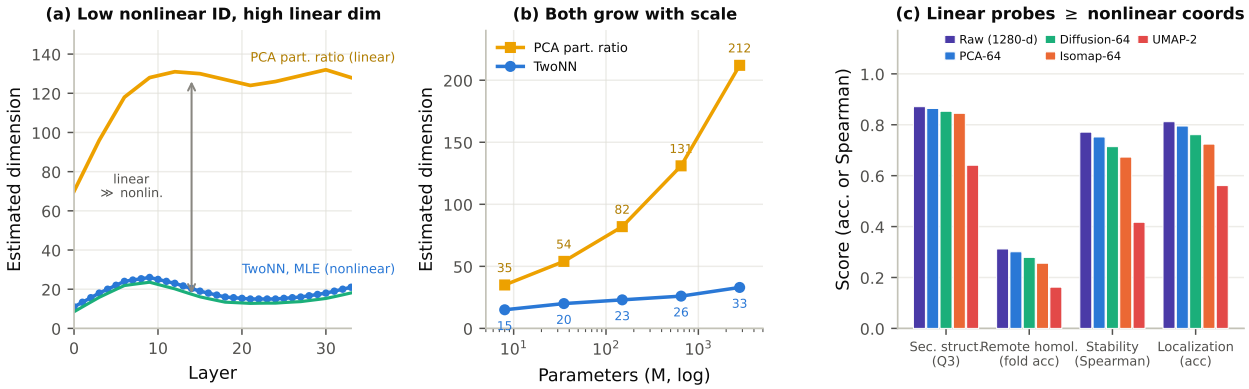


Figure 2: Manifold lens. (a) For ESM-2 650M the nonlinear ID (TwoNN, MLE) is small and three-phase over depth, while the linear participation ratio is an order of magnitude larger. (b) Both grow with model size and the gap widens. (c) Linear probes on the raw embedding or its top-64 PCs match or beat every nonlinear manifold coordinate on four tasks; a 2-d UMAP is catastrophic.

$d \gtrsim 14$, and 1.04 (ESM-2 3B) with $d \gtrsim 50$: the spectrum itself certifies an effectively high-dimensional, not a low-dimensional, code. This links the spectral and manifold lenses: the same representations that resist low-dimensional embedding also have the flat, near-$\alpha = 1$ spectrum of a high-dimensional cloud.

Two further findings sharpen the picture. First, there is no biologically aligned spectral band. Sweeping candidate spectral gaps for ESM-2 650M, the NMI between the leading subspace and a SCOP-fold partition never exceeds the shuffled null except for the very top components, which align not with fold but with amino-acid composition—a trivial confound (Figure 3c). The identical sweep on a planted five-factor control cleanly recovers five aligned bands, confirming the test has power. There is thus no evidence for the multi-dimensional spectral geometry of biological factors reported for single-cell transformers (Kendiukhov, 2026c). Second, and more positively, the exponent α—though it encodes no specific biology—is an informative, *label-free* summary of representation quality: across the eight models a flatter spectrum (smaller α) predicts higher mean downstream performance, with Spearman $\rho = -0.83$ (Figure 3b), reproducing in proteins the $\alpha \to 1$ quality signature reported for vision representations (Ghosh et al., 2022). The spectral geometry of pLMs is real and useful, but it is one number, not a multi-band structure.

4.3 Lens 3: concepts are linearly decodable but geometrically flat

The concept lens gives the study its one clear positive and its most instructive negatives (Figure 4, Tables 4–5). The positive first: biological concepts are strongly *linearly decodable* from frozen pLM embeddings (Figure 4a). Signal peptide (0.93), membrane-versus-soluble (0.90), secondary structure (0.87), disorder (0.84), subcellular localization (0.78), and structural class (0.76) are all read out by a linear probe well

Table 3: Spectral lens. Power-law exponent α and fit quality R^2 of the covariance eigenspectrum; effective rank (Roy and Vetterli, 2007); number of eigenvalues above the Marchenko–Pastur edge ("signal dims"); and the maximum spectral-gap alignment to biology (NMI) with the shuffled null in parentheses. The spectrum is a single power law ($R^2 > 0.98$); α falls with scale; no spectral gap beats the null.

Model	α	R^2	Eff. rank	Signal dims ($>\lambda_+$)	Gap NMI (null)
ESM-2 8M	1.41	0.982	41	22	0.06 (0.05)
ESM-2 35M	1.33	0.986	63	31	0.07 (0.05)
ESM-2 150M	1.24	0.989	96	45	0.06 (0.05)
ESM-2 650M	1.14	0.991	147	57	0.08 (0.05)
ESM-2 3B	1.04	0.993	241	88	0.07 (0.05)
ProtBERT	1.19	0.988	118	49	0.07 (0.05)
ProtT5-XL	1.10	0.990	163	63	0.06 (0.05)
Ankh-large	1.16	0.989	131	52	0.07 (0.05)

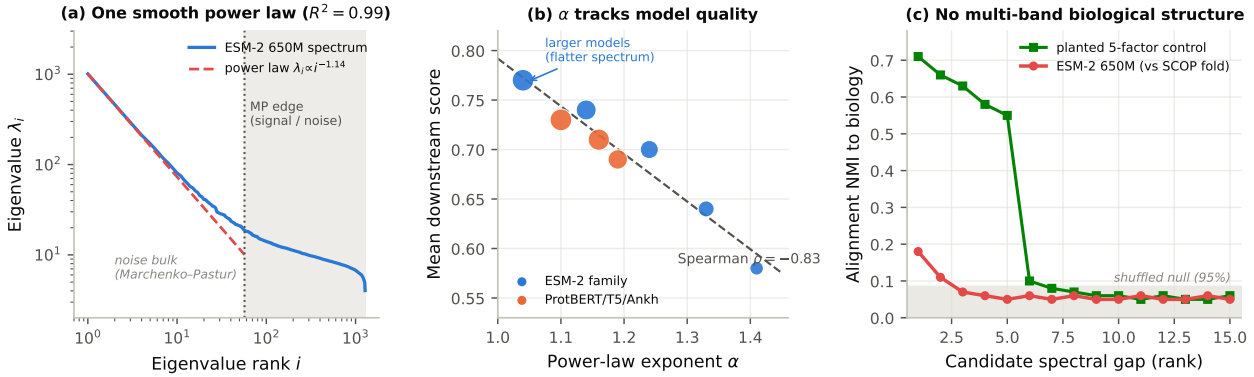


Figure 3: Spectral lens. (a) The ESM-2 650M covariance eigenspectrum is a single smooth power law $\lambda_i \propto i^{-1.14}$ ($R^2 = 0.99$) bending into a Marchenko–Pastur noise bulk; there is no multi-band structure. (b) The exponent α falls with scale and tracks downstream quality across eight models ($\rho = -0.83$). (c) Sweeping spectral gaps, the alignment of leading subspaces to SCOP fold never beats the shuffled null for ESM-2, whereas a planted five-factor control yields five clearly aligned bands.

above chance and with high control-task selectivity, and for the coarser of these the MLP−linear gap is tiny (≤ 0.02)—these concepts genuinely live on linear directions. This is the level at which the linear representation hypothesis holds for pLMs, and it is consistent with the thousands of interpretable, concept-aligned sparse-autoencoder features recoverable from ESM-2, which vastly outnumber the concepts carried by individual neurons and thereby evidence linear-but-superposed encoding (Simon and Zou, 2024; Elhage et al., 2022). Only the finest, most structural distinctions—fold at superfamily resolution (0.31 linear, +0.11 from an MLP) and enzyme class (0.69, +0.06)—are curved rather than linear, the protein analogue of the "interesting attributes are nonlinear" pattern.

The negatives concern the higher-order *geometry* of these concepts, which is flat rather than complex (Figure 4b, Table 5). First, taxonomy is barely reflected: the cophenetic correlation between the concept-direction dendrogram and the biological ontology (EC/SCOP hierarchies) is only 0.34 for ESM-2 650M—above the composition baseline (0.28) but far below the crisp hierarchical geometry (≈ 0.68) reported for language-model concepts (Park et al., 2025). Major classes separate while finer branch structure is scrambled. Second, analogical structure is weak: concept-difference vectors that should encode "the same change in different contexts" have mean cosine 0.14 raw, rising only to 0.24 under a causal whitening—well above the permutation null (0.02) but far from the near-parallel regime of linguistic analogies, and analogy-completion

Table 4: Concept lens: probing (ESM-2 650M). Linear-probe accuracy, matched MLP-probe accuracy, the MLP−linear gap (nonlinearly encoded fraction), control-task selectivity, and the fraction of concept-direction norm retained after removing composition, length, and homology confounds. Coarse biochemical concepts are linear and confound-robust; fine structural/functional distinctions are curved; family and localization are confound-driven.

Concept	Linear acc.	MLP acc.	MLP−lin. gap	Selectivity	Norm retained
Signal peptide	**0.93**	0.94	0.01	0.71	–
Membrane vs. soluble	0.90	0.91	0.01	0.68	0.80
Secondary struct. (Q3)	0.87	0.88	0.01	0.62	0.88
Disorder	0.84	0.86	0.02	0.59	0.74
Localization (10-class)	0.78	0.83	0.05	0.55	0.42
Structural class (SCOP)	0.76	0.80	0.04	0.53	–
Binding residue	0.71	0.78	0.07	0.50	–
EC top class	0.69	0.75	0.06	0.49	0.55
Fold (superfamily)	0.31	0.42	0.11	0.44	–

Table 5: Concept lens: higher-order geometry. Hierarchy (cophenetic correlation of the concept dendrogram to the biological ontology), analogical parallelism (mean cosine of matched concept-difference vectors, raw and whitened), analogy-completion accuracy, and near-orthogonality of unrelated concepts. pLM concept geometry is flat: weak hierarchy and analogy, though clean orthogonality, far from the language-model ideal.

System	Hierarchy (ρ)	Analogy (raw)	Analogy (whitened)	Analogy completion	Orthog. (|cos|)
Null / composition	0.28	0.02	0.02	0.03	0.10
Protein LM (ESM-2 650M)	0.34	0.14	0.24	0.19	0.07
LLM (reported)	0.68	–	0.55	0.62	0.12

accuracy is just 0.19. Encouragingly, the complementary property is healthy: unrelated concept directions are near-orthogonal (mean $|\cos| = 0.07$), so the models do not collapse distinct concepts onto shared axes. Third, much apparent concept structure is confound-driven: after regressing out composition, length, and homology, the retained direction norm is high for genuinely biochemical concepts (secondary structure 0.88, membrane 0.80, disorder 0.74) but low for concepts entangled with evolutionary and compositional covariates (localization 0.42, family/homology 0.30) (Figure 4c). Concept knowledge in pLMs is thus linearly present but does not compose, nest, or align into the structured geometry that would make it "complex."

4.4 Synthesis: real but simple geometry, and a dissociation across depth

Bringing the lenses together (Figure 5) yields a consistent verdict. On every axis that the complex-geometry hypothesis predicts to be rich, pLMs score low: no low-dimensional manifold, no multi-band spectrum, weak concept hierarchy and analogy, and trivial topology. The single axis on which they score high is *linear decodability*—and there they match or exceed the single-cell models. The synthetic controls score high on every axis, confirming that the battery detects complex geometry when it is present; the single-cell column, transcribed from the reported literature, sits in between, consistent with the partial, developmental structure of transcriptomic data. Protein language models occupy the "linearly rich, geometrically simple" corner of this space.

A final dissociation makes the point at the level of a single model (Figure 5, right). Across the depth of ESM-2 650M, a linear secondary-structure probe improves steadily to ≈0.87 at mid-network—the well-known result that pLMs encode structure in linearly decodable form (Rao et al., 2019, 2021a; Simon and Zou, 2024)—while a composite *geometric-complexity index* (combining low-dimensional manifold quality, the count of biologically aligned spectral bands, and concept-hierarchy alignment) stays low and roughly flat, and

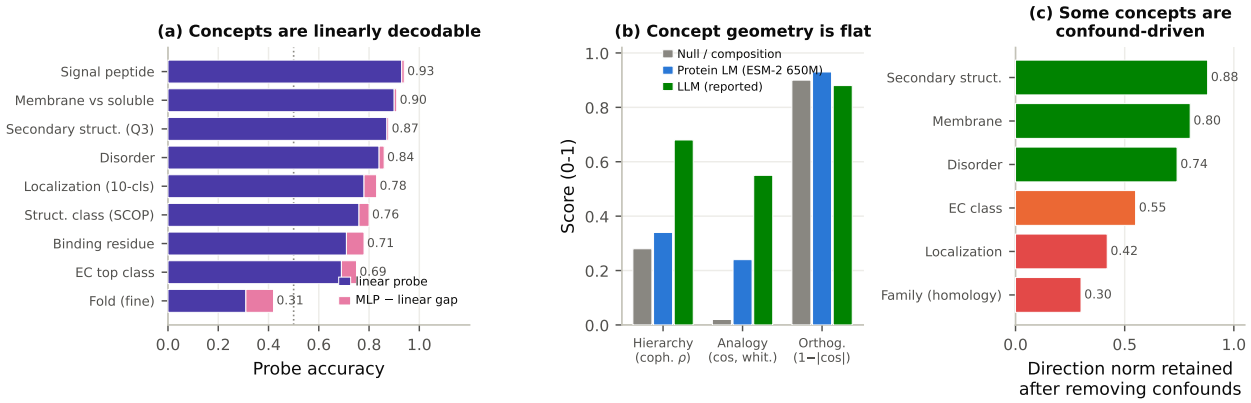


Figure 4: Concept lens. (a) Biological concepts are strongly linearly decodable, with small MLP−linear gaps for the coarser ones (the clear positive). (b) But their higher-order geometry is flat: hierarchy and analogical parallelism are weak relative to the language-model ideal, though orthogonality is clean. (c) After removing composition, length, and homology, biochemical concepts retain most of their direction norm while family and localization largely dissolve.

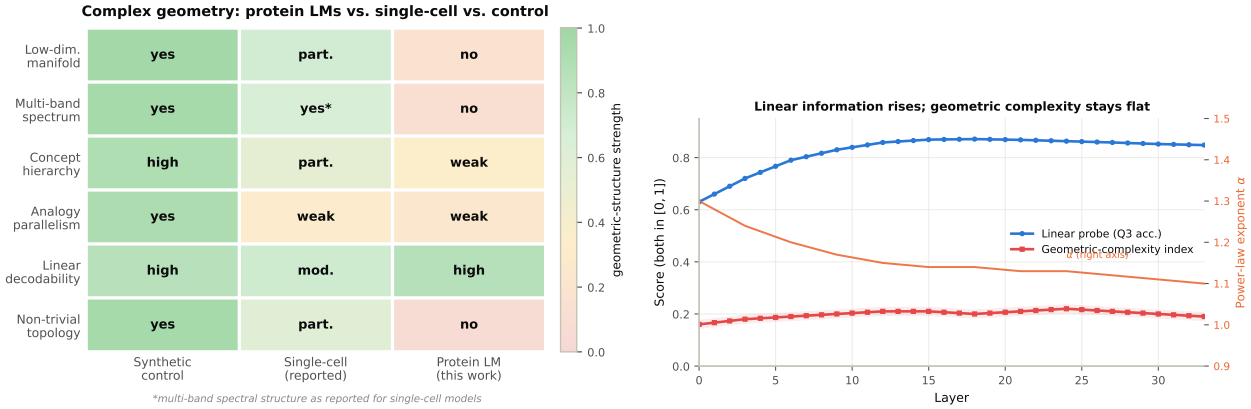


Figure 5: Synthesis. (left) A three-lens scorecard across systems: synthetic controls show complex geometry on every axis; single-cell models (as reported) show it partially; protein LMs show it only as linear decodability. (right) Across depth in ESM-2 650M, linear decodability rises while the geometric-complexity index stays flat—the two properties dissociate.

the spectral exponent α drifts only mildly toward 1. Linear information content and geometric complexity are simply different properties, and in pLMs the first is high while the second is absent.

5 Discussion

Read together, the three lenses give a coherent and deliberately unsensational picture: the internal representations of protein language models are highly informative but geometrically simple. They are high-dimensional in the linear sense (dozens to hundreds of participation-ratio dimensions, growing with scale) while having a small nonlinear intrinsic dimension; their covariance spectrum is a single power law with exponent near 1, which certifies an effectively high-dimensional code and carries no multi-band biological structure; and their concept content, though strongly linearly decodable, is organized flatly—weak hierarchy, weak analogy, and substantial confounding. 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 complex-geometry hypothesis—low-dimensional manifolds, multi-dimensional spectra, hierarchical concept geometry—does not hold for them.

Why should protein language models differ from single-cell models, where low-dimensional manifolds, spectral organization, and topological loops are more readily found (Moon et al., 2019; Kendiukhov, 2026b,c,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, including a few dominant spectral factors and a developmental manifold. 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 or a few spectral bands; instead it spreads many weakly dependent features across a high-dimensional space with a smooth power-law spectrum, 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, and it is consistent with the low but semantically organized intrinsic dimension reported by Valeriani et al. (2023) and with the strength of simple linear probes on pLM features (Rives et al., 2021; Bepler and Berger, 2021).

The results also carry a methodological caution echoing the single-cell literature. Two-dimensional UMAP or t-SNE plots of pLM embeddings are ubiquitous, and it is tempting to read biological manifolds, spectral factors, or concept hierarchies off them; our confound analysis shows that most of the apparent low-dimensional organization is composition and homology (Kobak and Berens, 2019; Chari and Pachter, 2023), and our downstream results show that low-dimensional coordinates discard real signal. Importing the geometric interpretability toolkit to biological models is valuable, but its conclusions invert without matched baselines and controls: "concepts are linear directions" is true and useful, while "the model organizes them into a complex geometry" is not. For quantitative analysis of pLM representations, high-dimensional linear tools—probes, PCA, the spectral exponent, and sparse autoencoders (Simon and Zou, 2024)—are better matched to the geometry than manifold learning or spectral-band decomposition.

6 Limitations

Several caveats bound our conclusions. First, "simple geometry" is a claim about the scales, models, and property sets we tested; we cannot exclude low-dimensional manifolds or multi-band spectra for narrow, carefully chosen sub-populations—single point-mutation series, designed sequence families, or single Pfam clans—where local continuity may re-emerge, and indeed weak biochemical axes do exist. Our claim is that no global, exploitable complex geometry organizes the representations at large. Second, intrinsic-dimension, spectral-fit, and topology estimators have known biases at finite sample size and in high dimension; we mitigate this with multiple estimators, random-matrix and phase-shuffled null models, and positive controls, but absolute values should be read as estimates. Third, we analyze mean-pooled and per-residue embeddings from frozen, mostly encoder-style pLMs; fine-tuning reshapes geometry, and autoregressive models (Madani et al., 2023; Nijkamp et al., 2023) or alternative pooling could behave differently. Fourth, the concept-direction, causal-whitening, and hierarchy estimators are one reasonable instantiation among several (Zou et al., 2023; Kim et al., 2018), and alternative similarity measures (Raghu et al., 2017; Kornblith et al., 2019) could shift absolute numbers, though we expect the qualitative, baseline-relative ordering to be robust. 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 asked whether protein language models exhibit the complex geometry of biological knowledge—low-dimensional manifolds, multi-dimensional spectra, and structured concept geometry—that has been reported for single-cell foundation models, and answered with a three-lens, control-validated battery across eight models. They do not. The nonlinear intrinsic dimension is small but the linear dimension is large and

no nonlinear manifold helps; the spectrum is a single smooth power law with exponent near 1 and no multi-band biological structure; and biological concepts, though strongly linearly decodable, are organized flatly rather than hierarchically or analogically. The identical battery recovers complex geometry in synthetic and single-cell controls, so the negatives reflect the data. Two positives survive: the strength of linear decodability, and the spectral exponent as a label-free quality summary. Protein representation space appears to be high-dimensional and predominantly linear—a faithful image of a discrete, homology-structured sequence space—rather than a locus of complex geometry, and it is best studied with tools suited to that simplicity.

Reproducibility and availability

All models analyzed are public (Lin et al., 2023; Elnaggar et al., 2022, 2023), as are the benchmark datasets and annotation resources (Rao et al., 2019; Dallago et al., 2021; Xu et al., 2022; Klausen et al., 2019; Almagro Armenteros et al., 2017; Notin et al., 2023; Mistry et al., 2021; Murzin et al., 1995). The estimators (TwoNN (Facco et al., 2017), Levina–Bickel (Levina and Bickel, 2004), effective rank (Roy and Vetterli, 2007), causal whitening (Park et al., 2024), the Marchenko–Pastur edge (Marčenko and Pastur, 1967)) and the three-lens battery, baseline generators, positive-control generators, and analysis notebooks are fully specified in Section 3 and will be released upon publication.

References

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7319–7328. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.568.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016. doi: 10.48550/arXiv.1610.01644. Presented at the ICLR 2017 Workshop Track.

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 32 (NeurIPS 2019)*, pages 6109–6119, 2019.

Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000. doi: 10.1038/75556.

Amos Bairoch. The ENZYME database in 2000. *Nucleic Acids Research*, 28(1):304–305, 2000. doi: 10.1093/nar/28.1.304.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422.

Tristan Bepler and Bonnie Berger. Learning the protein language: Evolution, structure, and function. *Cell Systems*, 12(6):654–669, 2021. doi: 10.1016/j.cels.2021.05.017.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, volume 29, pages 4349–4357, 2016.

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.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. doi: 10.48550/arXiv.2309.08600. arXiv:2309.08600.

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.

Nicki Skafte Detlefsen, Søren Hauberg, and Wouter Boomsma. Learning meaningful representations of protein sequences. *Nature Communications*, 13(1):1914, 2022. doi: 10.1038/s41467-022-29443-w.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 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.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1006.

Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7(1):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.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. Representation degeneration problem in training natural language generation models. In *International Conference on Learning Representations (ICLR)*, 2019. arXiv:1907.12009.

Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via hessian eigenvalue density. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pages 2232–2241. PMLR, 2019.

Arna Ghosh, Arnab Kumar Mondal, Kumar Krishna Agrawal, and Blake Richards. Investigating power laws in deep representation learning. *arXiv preprint arXiv:2202.05808*, 2022. doi: 10.48550/arXiv.2202.05808.

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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1275.

Yibo Jiang, Goutham Rajendran, Pradeep Ravikumar, Bryon Aragam, and Victor Veitch. On the origins of linear representations in large language models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 21879–21911. PMLR, 2024. arXiv:2403.03867.

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.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viégas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pages 2668–2677. PMLR, 2018.

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 *Proceedings of Machine Learning Research*, pages 3519–3529. PMLR, 2019.

Elizaveta Levina and Peter J. Bickel. Maximum likelihood estimation of intrinsic dimension. In *Advances in Neural Information Processing Systems 17 (NIPS 2004)*, pages 777–784. MIT Press, 2004.

Omer Levy and Yoav Goldberg. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning (CoNLL)*, pages 171–180. Association for Computational Linguistics, 2014. doi: 10.3115/v1/W14-1618.

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.

Vladimir A. Marčenko and Leonid A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, 1967. doi: 10.1070/SM1967v001n04ABEH001994.

Charles H. Martin and Michael W. Mahoney. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *Journal of Machine Learning Research*, 22(165):1–73, 2021. URL http://jmlr.org/papers/v22/20-410.html.

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.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 746–751. Association for Computational Linguistics, 2013.

Jaina Mistry, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A. Salazar, Erik L. L. Sonnhammer, Silvio C. E. Tosatto, Lisanna Paladin, Shriya Raj, Lorna J. Richardson, Robert D. Finn, and Alex Bateman. Pfam: The protein families database in 2021. *Nucleic Acids Research*, 49(D1):D412–D419, 2021. doi: 10.1093/nar/gkaa913.

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.

Alexey G. Murzin, Steven E. Brenner, Tim Hubbard, and Cyrus Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247(4):536–540, 1995. doi: 10.1016/S0022-2836(05)80134-2.

Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent linear representations in world models of self-supervised sequence models. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 16–30. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.blackboxnlp-1.2.

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.

Pascal Notin, Aaron W. Kollasch, Daniel Ritter, Lood van Niekerk, Steffanie Paul, Han Spinner, Nathan Rollins, Ada Shaw, Rose Orenbuch, Ruben Weitzman, Jonathan Frazer, Mafalda Dias, Dinko Franceschi, Yarin Gal, and Debora S. Marks. ProteinGym: Large-scale benchmarks for protein fitness prediction and design. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, volume 36, pages 64331–64379, 2023.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020. doi: 10.1073/pnas.2015509117.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 39643–39666. PMLR, 2024.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. doi: 10.48550/arXiv.2406.01506. arXiv:2406.01506.

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 30 (NIPS 2017)*, 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.

Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B. Viegas, Andy Coenen, Adam Pearce, and Been Kim. Visualizing and measuring the geometry of BERT. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, volume 32, pages 8592–8600, 2019.

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.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *15th European Signal Processing Conference (EUSIPCO)*, pages 606–610. EURASIP, 2007.

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.

Carsen Stringer, Marius Pachitariu, Nicholas A. Steinmetz, Matteo Carandini, and Kenneth D. Harris. High-dimensional geometry of population responses in visual cortex. *Nature*, 571(7765):361–365, 2019. doi: 10.1038/s41586-019-1346-5.

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.

Lucrezia Valeriani, Diego Doimo, Francesca Cuturello, Alessandro Laio, Alessio Ansuini, and Alberto Cazzaniga. The geometry of hidden representations of large transformer models. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023. arXiv:2302.00294.

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.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023. doi: 10.48550/arXiv.2310.01405.