A Look at the Isotropy of Pretrained Protein Language Models

Sheikh Azizul Hakim [*1] **Kowshic Roy** [*1] **M Saifur Rahman** [1]

Abstract

Large pretrained language models have transformed natural language processing, and their adaptation to protein sequences—viewed as strings of amino acid characters—has advanced protein analysis. However, the distinct properties of proteins, such as variable sequence lengths and lack of word-sentence analogs, necessitate a deeper understanding of protein language models (LMs). We investigate the isotropy of protein LM embedding spaces using average pairwise cosine similarity and the IsoScore method, revealing that models like ProtBERT and ProtXLNet are highly anisotropic, utilizing only 2–14 dimensions for global and local representations. In contrast, multi-modal training in ProteinBERT, which integrates sequence and gene ontology data, enhances isotropy, suggesting that diverse biological inputs improve representational efficiency. We also find that embedding distances weakly correlate with alignment-based similarity scores, particularly at low similarity.

1. Introduction

The most sophisticated machines in nature are proteins. Across the tree of life, proteins play a crucial role in catalyzing biochemical reactions, providing structural support for other cell organelles, facilitating cell signaling, contributing to immune defense, and even synthesizing other proteins (Kessel & Ben-Tal, 2018). Proteins can be regarded as sequences of amino acid characters. Consequently, machine-learning techniques tailored for natural language and other sequences are ideally suited for forecasting protein-related tasks (Ofer et al., 2021).

In recent years, natural language processing has been significantly advanced with the advent of large pretrained attention-based language models such as BERT (Devlin et al., 2019), XLNet (Yang et al., 2019), Albert (Lan et al., 2020), GPT (Radford et al., 2019) etc. (for a detailed survey, check (Min et al., 2023)). The same concepts, and often the same architectures, have been applied to proteins (Rao et al., 2019; Elnaggar et al., 2022; Brandes et al., 2022). The attention mechanism, in particular, has been shown to correlate with many known biological and biochemical properties (Vig et al., 2020). However, prior works have also noted that protein sequences behave differently from natural languages; for example, protein sequences can vary significantly in length, from under fifty amino acids to over thousands, unlike words and sentences, and we cannot break down proteins into analogs of words and sentences in the first place (Brandes et al., 2022).

Language models (LM) use several types of embeddings that map a linguistic concept into a geometric space. Traditionally, static embeddings have been utilized (Pennington et al., 2014), and such approaches have been theoretically explained as the factorization of a word-context matrix containing a co-occurrence statistic (Levy & Goldberg, 2014b;a). Theoretical and empirical evidence suggests that many of these models are *isotropic*, i.e., angularly uniform (Arora et al., 2016). However, context-sensitive word representations can be found from pretrained language models, such as BERT (Devlin et al., 2019), GPT (Radford et al., 2019), and are useful for several downstream tasks. Ethayarajh (Ethayarajh, 2019) investigated the isotropic properties of the contextualized embedding spaces of such pretrained models using average pairwise cosine similarity. The cosine similarity of two vectors $\mathbf{x}$ and $\mathbf{y}$ is defined as the normalized dot product between them $\left(\frac{\mathbf{x}.\mathbf{y}}{|\mathbf{x}||\mathbf{y}|}\right)$. The contextual embedding spaces of the pretrained LMs came out to be, somewhat surprisingly, *highly anisotropic*. Increasing isotropy has been suggested as a way to improve the performance of BERT (Rogers et al., 2020), but (Rajaee & Pilehvar, 2021) showed that increasing isotropy using existing methods of post-processing pretrained LMs may hurt performance. (Cai et al., 2021) argued that a different notion of isotropy might indeed exist for the contextualized embedding spaces and identified some other geometric properties.

(Rudman et al., 2022) argued that all existing measures of

[1]Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology, Dhaka, Bangladesh. Correspondence to: M Saifur Rahman <mrahman@cse.buet.ac.bd>.

Proceedings of the ICML 2025 Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences, Vancouver, Canada. 2025. Copyright 2025 by the author(s).

measuring isotropy have fundamental shortcomings. They identified some key properties of isotropy, such as mean agnosticity, global stability, rotational invariance, etc., and proposed a new scoring method, named *IsoScore*, based on the covariance matrix of the principal components. They also showed that this score can be used to approximate the number of dimensions effectively used by the point cloud in consideration.

Although several works have been done to analyze the isotropy and geometry of the embedding spaces for natural languages, the attempt to do so is scarce (if any) for protein sequences. We analyze both the cosine similarity-based and IsoScore-based approaches to analyze the isotropy of protein embedding spaces. We find that protein LMs are highly anisotropic, and a much lower dimensional embedding space might come equally handy for downstream tasks. We also find that protein embedding distances (cosine and Euclidean) exhibit weak overall correlations with traditional alignment-based similarity scores, reliably capturing biological relationships only at high similarity; at low similarity, their high variance highlights limitations in representing distant relationships, underscoring the need for multi-modal models to integrate diverse biological signals.

To extend our analysis, we investigate the isotropy and geometry of local (per-residue) embeddings in protein language models, finding them to be highly anisotropic, utilizing only approximately 14 dimensions on average across models (Table 3). By visualizing these embeddings in a 3D space defined by the first three principal components, we observe distinct clustering patterns for each amino acid, suggesting that local embeddings capture residue-specific biochemical properties. These findings indicate significant redundancy in local representations, similar to global embeddings, and highlight opportunities for dimensionality reduction in multi-modal protein models.

The contribution of this study is to explore various properties of protein embedding spaces. At first, We find that protein LMs are highly anisotropic, and a much lower dimensional embedding space might come equally handy for downstream tasks. Then, we explore the relationship between distances in embedding space and the alignment distances between the protein sequences. We extend the same result of anisotropy in the case of local (per-residue) representations. We also explore the geometry of the local embeddings for each amino acid. (#TODO: rewrite the paragraph)

2. Materials and Methods

2.1. Dataset

We use the SwissProt subset of the UniProt database (Consortium, 2023), consisting of approximately 570,000 protein sequences with experimentally validated annotations. SwissProt is manually curated and includes high-quality functional and structural information, making it a reliable benchmark for evaluating protein language models. Its focus on experimentally verified proteins ensures that downstream tasks—such as similarity analysis or embedding evaluation—are grounded in biologically meaningful data.

2.2. Protein Language Models in Consideration

We evaluated three pretrained protein language models from (Elnaggar et al., 2022): ProtXLNet, ProtBERT, and ProtBERT-BFD, the latter trained on a distinct dataset. Pretrained weights were obtained from Hugging Face[1]. In these models, protein sequences are treated as sentence-like sequences, with each amino acid residue represented as a word-like token. The underlying architectures, adapted from their natural language counterparts (XLNet and BERT), remain unmodified. These models generate per-residue (local) embeddings for input proteins, with per-protein (global) embeddings derived through average pooling of local embeddings. (Elnaggar et al., 2022) explored alternative pooling strategies, including minimum, maximum, and concatenation pooling, but found average pooling to be the most effective for generating robust global representations.

We also evaluated ProteinBERT from (Brandes et al., 2022), which employs a distinct architecture tailored for protein modeling. Unlike sequence-only models, ProteinBERT is trained on both protein sequences and gene ontology (GO) annotations, enabling a multi-modal approach that captures functional and structural insights. Its architecture directly generates both per-residue (local) and per-protein (global) embeddings, eliminating the need for pooling local embeddings to derive global representations. Pretrained weights were obtained from the model's GitHub repository[2].

3. Results and Discussion

3.1. Anisotropy of Global Embeddings

We computed IsoScores for embeddings generated by protein language models (LMs). This metric quantifies isotropy through mean agnosticity, global stability, and rotational invariance (Rudman et al., 2022). If the IsoScore of a pointcloud ($\mathbf{X} \in \mathcal{R}^n$) is $i(\mathbf{X})$, then according to (Rudman et al., 2022), effectively $\dim(\mathbf{X}) = (i(\mathbf{X}) \times (n-1) + 1)$ dimensions are utilized. If $i(\mathbf{X}) \approx 0$, then the pointcloud is highly anisotropic, and no more than one dimension is being effectively utilized. If $i(\mathbf{X}) \approx 1$, then the pointcloud is highly isotropic and all the dimensions are being effectively utilized. We tabulate the IsoScores and the number of effectively utilized dimensions (fractions are rounded up to the next integer) in Table 1. We find that the protein LMs

[1] https://huggingface.co/Rostlab
[2] https://github.com/nadavbra/protein_bert

Model Name	Embedding Dimension	IsoScore	Effectively Used Dimensions
ProtBERT	1024	0.001658146	3
ProtBERT-BFD	1024	0.003967522	6
ProtXLNet	1024	0.001502474	3
ProteinBERT	512	**0.231227934**	**120**

Table 1. Embedding dimension, IsoScore, and effectively used dimensions for different protein language models.

Model Name	**cosine**	**sq_euclidean**	**alignment_score**	**similarity_score**
cosine	1.000000	0.791068	0.013804	-0.011159
sq_euclidean		1.000000	-0.102698	-0.145814
alignment_score			1.000000	0.847258
similarity_score				1.000000

Table 2. Correlation matrix between different distance metrics for ProtBERT.

developed (Elnaggar et al., 2022), which is trained using the eponymous model architectures built for natural language, under consideration are *highly anisotropic* and use very few (2-5) dimensions. On the other hand, ProteinBERT has a relatively high isotropy score (0.23) and uses 120 dimensions effectively. For comparison purposes, the reported IsoScores in (Rudman et al., 2022) for BERT and GPT are 0.11 and 0.18, respectively. Thus, while protein LMs using traditional architectures are, in general, more anisotropic than natural LMs, ProteinBERT is more isotropic. We think this is because ProteinBERT uses a different architecture and is trained not only from protein sequences, but also from gene ontology (GO) annotations. ProteinBERT's architecture enables it to output global and local representations separately, instead of the other models, where local embeddings are pooled to generate global embeddings.

3.2. Comparison between Alignment Distances and Embedding Distances

We investigated the relationship between traditional similarity measures (alignment score and similarity score) and embedding-based measures (squared Euclidean distance and cosine similarity). We used BioPython (Cock et al., 2009) to calculate the alignments using the PAM-250 scoring matrix (Dayhoff et al., 1978). We define similarity score as the fraction of identical residues in optimal alignments. For this experiment, we randomly sampled 1% of the Swissprot proteins which resulted in 6.4×10^6 protein pairs. Our results show that the two traditional measures are strongly correlated with each other, as are the two embedding-based measures. However, correlations between traditional and embedding-based metrics are weaker—often low or even negative—suggesting that these approaches do not capture the same aspects of protein similarity. We report the pairwise correlation coefficients of the four similarity measures, as calculated for ProtBERT, in Table 2. Other models exhibit similar trends.

We further investigated the relationship between embedding-based distances and traditional alignment-based similarity scores and observed consistent non-linear patterns. As shown in Figures 1a and 1b, while the overall correlations are weakly negative, a clear structural trend emerges: for low similarity scores, both squared Euclidean and cosine distances exhibit high variance and span a wide range of values, indicating poor predictive power. In contrast, at high similarity scores, both metrics converge—Euclidean distances become consistently low, and cosine similarities cluster near 1.0. This asymmetric behavior suggests that embedding distances, while informative at the high end of biological similarity, are unreliable indicators of similarity in the low-alignment regime, revealing a key limitation in how current embeddings capture biologically meaningful relationships. While Figures 1a and 1b are generated for ProteinBERT, this trend generalizes across all models in consideration.

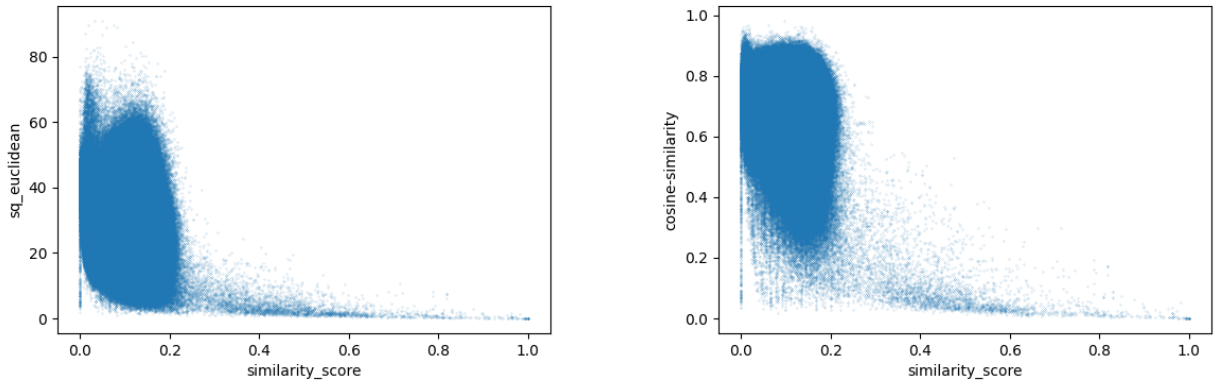
3.3. Anisotropy of Local Embeddings

To investigate the geometric structure of protein embedding spaces, we measured the IsoScore of individual amino acid token embeddings. Table 3 reveals a clear anisotropy in the embedding space for each amino acid. For 1024 embedding dimensions for each of the three models, only about 14 dimensions are effectively used on average.

4. Conclusion

Our analysis reveals a critical mismatch between biological richness and embedding geometry in pretrained protein language models. Sequence-only models like ProtBERT and ProtXLNet produce highly anisotropic embeddings that utilize minimal representational capacity, while multi-modal ProteinBERT demonstrates improved isotropy through biological priors.

These results have direct implications for generative biology,



(a) Squared Euclidean distance of ProteinBERT embeddings vs alignment similarity score.

(b) Cosine similarity of ProteinBERT embeddings vs alignment similarity score.

Figure 1. Relationship between embedding-based distances and traditional alignment-based similarity scores.

where diverse and informative latent spaces are essential for tasks such as protein design, variant prediction, and molecular optimization. The strong anisotropy in current embeddings suggests that models may fail to explore biologically meaningful subspaces during generation, leading to reduced diversity or biological invalidity. Moreover, we observe that learned distances in embedding space—based on cosine and Euclidean metrics—correlate poorly with biologically grounded similarity measures like sequence alignment, particularly at low similarity. This divergence implies that embedding spaces lack robustness when modeling distant or novel proteins, a serious limitation for generative models that seek to extrapolate beyond known data.

In a concurrent work, (Tule et al., 2025) analyzed the phylogenetic properties captured by protein LMs. Future studies might look for if improving isotropy leads to better phylogenetic relationships.

Looking ahead, we advocate for the development of next-generation protein LMs that explicitly optimize for geometric richness, isotropy, and biological alignment. Promising directions include biologically supervised contrastive pre-training, isotropy-promoting regularization, and functional embedding constraints grounded in ontologies or structural data. Such efforts could produce embeddings that are simultaneously compact, generative, and biologically meaningful—making them ideal backbones for AI-driven discovery in protein science. By better understanding and shaping the geometry of protein embedding spaces, we lay the groundwork for interpretable, multi-modal, and experimentally actionable generative models in biology.

Code Availability:

Our implementation and the generated plots can be found in https://github.com/vodro/geometry_of_proteins.

Amino Acid	BERT	BERT-BFD	XLNet
Alanine (A)	0.013340	0.017366	0.011098
Cysteine (C)	0.012388	0.013517	0.010462
Aspartic Acid (D)	0.012656	0.013981	0.011726
Glutamic Acid (E)	0.013102	0.017743	0.012512
Phenylalanine (F)	0.013049	0.012437	0.010798
Glycine (G)	0.011422	0.011228	0.009837
Histidine (H)	0.011963	0.011971	0.011251
Isoleucine (I)	0.013796	0.012815	0.011363
Lysine (K)	0.012053	0.021384	0.012672
Leucine (L)	0.013934	0.013625	0.011241
Methionine (M)	0.015887	0.017329	0.010793
Asparagine (N)	0.010028	0.017633	0.010436
Proline (P)	0.011258	0.011756	0.011503
Glutamine (Q)	0.012816	0.020504	0.011812
Arginine (R)	0.012033	0.012910	0.012437
Serine (S)	0.010018	0.017425	0.009935
Threonine (T)	0.011633	0.014803	0.010239
Valine (V)	0.014402	0.013638	0.010828
Tryptophan (W)	0.012764	0.011212	0.011547
Tyrosine (Y)	0.013153	0.012430	0.011601
Unknown (X)	0.010520	0.006424	0.007258

Table 3. Per-amino acid IsoScore values for three models: BERT, BERT-BFD, and XLNet, rounded to six decimal places.

References

Arora, S., Li, Y., Liang, Y., Ma, T., and Risteski, A. A Latent Variable Model Approach to PMI-based Word. *Transactions of the Association for Computational Linguistics*, 4: 385–399, December 2016. doi: 10.1162/tacl_a_00106.

Brandes, N., Ofer, D., Peleg, Y., Rappoport, N., and Linial, M. ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8): 2102–2110, April 2022. ISSN 1367-4803. doi: 10.1093/bioinformatics/btac020.

Cai, X., Huang, J., Bian, Y., and Church, K. Isotropy in the contextual embedding space: Clusters and manifolds. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=xYGNO86OWDH.

Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422, 2009.

Consortium, T. U. Uniprot: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531, 2023.

Dayhoff, M. O., Schwartz, R. M., and Orcutt, B. C. A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure*, volume 5, pp. 345–352. National Biomedical Research Foundation, 1978.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *North American Chapter of the Association for Computational Linguistics*, 2019.

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., and Rost, B. ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127, October 2022. ISSN 1939-3539. doi: 10.1109/TPAMI.2021.3095381.

Ethayarajh, K. How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. *ACL Anthology*, pp. 55–65, November 2019. doi: 10.18653/v1/D19-1006.

Kessel, A. and Ben-Tal, N. *Introduction to Proteins: Structure, Function, and Motion, SECOND EDITION (Chapman & Hall/CRC Mathematical and Computational Biology)*. March 2018. ISBN 978-1-49874717-2. doi: 10.1201/9781315113876.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations, April 2020. URL https://iclr.cc/virtual_2020/poster_H1eA7AEtvS.html. [Online; accessed 5. Nov. 2023].

Levy, O. and Goldberg, Y. Neural word embedding as implicit matrix factorization. In *NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, pp. 2177–2185. MIT Press, Cambridge, MA, USA, December 2014a. doi: 10.5555/2969033.2969070.

Levy, O. and Goldberg, Y. Linguistic Regularities in Sparse and Explicit Word Representations. *ResearchGate*, pp. 171–180, January 2014b. doi: 10.3115/v1/W14-1618.

Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., Agirre, E., Heintz, I., and Roth, D. Recent Advances in Natural Language Processing via Large Pre-trained Language Models: A Survey. *ACM Computing Surveys*, 56(2):1–40, September 2023. ISSN 0360-0300. doi: 10.1145/3605943.

Ofer, D., Brandes, N., and Linial, M. The language of proteins: NLP, machine learning & protein sequences. *Computational and Structural Biotechnology Journal*, 19: 1750–1758, January 2021. ISSN 2001-0370. doi: 10.1016/j.csbj.2021.03.022.

Pennington, J., Socher, R., and Manning, C. Glove: Global Vectors for Word Representation. *EMNLP*, 14:1532–1543, January 2014. doi: 10.3115/v1/D14-1162.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language Models are Unsupervised Multitask Learners, 2019. [Online; accessed 6. Nov. 2023].

Rajaee, S. and Pilehvar, M. T. How Does Fine-tuning Affect the Geometry of Embedding Space: A Case Study on Isotropy. *Conference on Empirical Methods in Natural Language Processing*, 2021.

Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, P., Canny, J., Abbeel, P., and Song, Y. Evaluating Protein Transfer Learning with TAPE. *Advances in Neural Information Processing Systems*, 32, 2019.

Rogers, A., Kovaleva, O., and Rumshisky, A. A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics*, 8:842–866, January 2020. doi: 10.1162/tacl_a_00349.

Rudman, W., Gillman, N., Rayne, T., and Eickhoff, C. IsoScore: Measuring the uniformity of embedding space

utilization. In Muresan, S., Nakov, P., and Villavicencio, A. (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 3325–3339, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.262. URL https://aclanthology.org/2022.findings-acl.262.

Tule, S., Foley, G., and Bodén, M. Do protein language models learn phylogeny? *Briefings in Bioinformatics*, 26 (1):bbaf047, 02 2025. ISSN 1477-4054. doi: 10.1093/bib/bbaf047. URL https://doi.org/10.1093/bib/bbaf047.

Vig, J., Madani, A., Varshney, L. R., Xiong, C., and Rajani, N. F. BERTology Meets Biology: Interpreting Attention in Protein Language Models. *ResearchGate*, June 2020. doi: 10.1101/2020.06.26.174417.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le, Q. V. XLNet: generalized autoregressive pretraining for language understanding. In *NIPS'19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pp. 5753–5763. Curran Associates Inc., December 2019. doi: 10.5555/3454287.3454804.