Probing the Embedding Space of Protein Foundation Models through Intrinsic Dimension Analysis

Soojung Yang, Juno Nam, Tynan Perez, Jinyeop Song, Xiaochen Du, Rafael Gómez-Bombarelli *
Massachusetts Institute of Technology

Abstract

Protein foundation models produce embeddings that are valuable for various downstream tasks, yet the structure and information content of these embeddings remain poorly understood, particularly in relation to diverse pre-training tasks and input modalities. We apply intrinsic dimension (I_d) analysis to quantify the complexity of protein embeddings from several widely used models, including ESM-2, ESM-IF, ProstT5, and ProteinMPNN. We also employ I_d correlation (I_dCor) to measure the shared information between different embeddings. Our results reveal a universality in protein embeddings, with similar I_d scales across models and strong correlations between protein and residue embeddings. We observe significant redundancy, with I_d values much smaller than the original embedding dimensions. We also show that models capture both spatial and sequential long-range correlation, with correlation decay rate differing based on the input modalities and pre-training tasks. Lastly, we analyze mutant embeddings, revealing that mutations cluster effectively by site, and fine-tuning further reduces the I_d to capture task-specific representations.

1 Introduction

Protein foundation models, trained on large scale protein sequence and structure datasets (e.g., PDB [5] and UniProt [29]), have emerged as powerful tools for encoding protein sequences and structures into embeddings that capture rich, biologically meaningful information [1, 24]. These embeddings have been successfully employed in a wide range of downstream tasks, including protein function prediction, protein design, and structural analysis [1, 12, 23, 24, 25]. The landscape of protein foundation models is highly diverse, with models being trained on different pretraining tasks (e.g., masked language modeling, inverse folding), accepting various input types (e.g., amino acid sequences, structural data), and come in different sizes and embedding dimensions.

Despite the success of protein foundation models, the information content of their embeddings, and how the pretraining objectives and input modalities shape these embeddings, remains unclear. While assumptions can be made based on a model's training task and inputs, a systematic and quantitative understanding of the embeddings' characteristics is still lacking. Previous efforts to analyze and compare protein embeddings have primarily focused on visualizing dimensionally-reduced projections and benchmarking performance on downstream tasks [1, 24, 25]. While these methods provide valuable insights, they may not fully capture the intrinsic properties of the embeddings or the underlying task-independent relationships between different models [30].

In this work, we analyze protein embeddings in a data-intrinsic manner using a concept of intrinsic dimension I_d. The intrinsic dimension is defined as the minimum number of variables required to describe a dataset effectively, providing a measure of the complexity and variability of the data. By estimating the I_d of protein embeddings, we aim to gain insights into their underlying structure and

*rafagb@mit.edu

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

information content. Furthermore, we employ a recent method [4] to estimate the mutual information that are shared across different protein embeddings. This approach has been shown to outperform traditional Euclidean distance-based methods in high-dimensional settings.

I_d analysis has been performed in other domains such as NLP and CV and provided insights, but its application in protein domain was limited, and I_dCor analysis has yet to be done so far to our knowledge.

Our key findings are as follows:

- **Universality of protein embeddings**. We observe that the I_ds of proteins and residues are consistent across foundation models. The fact that the sets of amino acids with small I_ds are nearly identical across different models, along with the high correlation between protein and residue embeddings, suggests that these embeddings capture a shared, universal structure across both the structural and sequence modalities of proteins.
- **Redundancy in protein embeddings.** Estimated I_d values are much smaller than the original embedding dimensions, indicating a high level of redundancy. High I_dCor values also suggest that a considerable amount of information is shared between the residues within the same protein and the residues and their corresponding protein embeddings.
- **Local and long-range awareness**. Residue embeddings are more correlated with sequentially or spatially proximal residues than with distal ones. The degree of decay in correlation varies depending on the pretraining task of the model, indicating that different models capture local and global geometries to varying degrees.
- **Understanding mutant embeddings**. ESM-2 mutant embeddings cluster by mutation sites and amino acid types, showing that the embeddings capture both local mutations and global protein context. Fine-tuning on the mutant stability score prediction task further reduces the I_d to 2.5, reflecting the compactness of task-specific representation.

2 Related Works

Protein foundation models We analyze embeddings from four popular protein foundation models: ESM-2 [17], ESM Inverse Folding (ESM-IF) [11], ProstT5 [10], and ProteinMPNN [7]. The following reviews key similarities and differences in architecture and training objectives.

- **ESM-2 [17]:** A Transformer encoder pretrained using masked language modeling to predict masked amino acids from sequence context, capturing statistical dependencies without structural information. We utilize embeddings from models of varying sizes (8M, 150M, and 650M parameters).
- **ESM Inverse Folding (ESM-IF) [11]**: Employs a GVP-Transformer encoder based on Geometric Vector Perceptron (GVP) [14] to embed structural information and a Transformer decoder to autoregressively generate amino acid sequences, effectively reversing the ESMFold process. It is trained on both experimental structures and AlphaFold2 predictions.
- **ProstT5 [10]:** A bidirectional encoder-decoder Transformer finetuned from pretrained ProtT5 model [8] to embed both sequence and structure. It uses 1) masked language modeling to predict amino acid (AA) or geometric (3Di) tokens given the partial information of the other, and 2) a bidirectional translation objective to obtain an entire sequence from complete structural information or vice versa, producing separate embeddings for amino acid identities and structure.
- **ProteinMPNN (MPNN) [7]:** A purely GNN-based encoder-decoder model that generates amino acid identities from structural information. Its encoder provides geometric embeddings for nodes (V), and also edges (E), an aggregation of edge features of spatial k-nearest neighbors. We also utilize the decoder's concatenated embeddings: ESV (nodes, edges, amino acid identity from a lookup table) and EXV (nodes, edges, mask token). We use the MPNN model trained with the highest Gaussian noise level since it showed the best benchmark performance in [7].

Intrinsic dimension-based analysis of embeddings The intrinsic dimension (I_d) provides insights into the structure of latent manifolds in deep neural models [13, 18, 21, 26] . I_d is also known to

relate to neural scaling laws in computer vision tasks by predicting loss trends [3, 21]. Tracking how I_d changes through neural network layers helps understand information flow: it typically increases in early layers and decreases in later ones, with the final layer's I_d known to predict test set classification accuracy [2, 15, 30]. For protein models, [30] analyzed I_d across layers in transformer-based architectures.

Recently, [4] introduced the I_dCor metric to estimate mutual information between datasets, assessing multimodal correlations between image and text models. However, no prior work has applied I_d analysis or I_dCor to protein or chemical models to understand their relationships. While methods like Singular Value Canonical Correlation Analysis (SVCCA) [22], Centered Kernel Alignment (CKA) [16], and Distance Correlation [27, 31] can be used to compare embeddings in a similar manner, they are most effective within the same data modality [4]. Since we aim to compare models with different input types (e.g., amino acid sequences vs. 3D coordinates), we employ I_dCor analysis.

3 Methods

Dataset preparation and caching of embeddings From the 16,380 cluster representatives of 30% sequence identity clusters of protein chains in the PDB, we selected 4,591 structures that do not contain any residues with missing structural information. For each protein foundation model, we extracted and cached residue embeddings from either the sequences or their corresponding PDB structures, and generated protein-level embeddings through mean pooling. During this process, no sequences were truncated for embedding generation or mean pooling; however, for residue-level embedding analysis, sequences were truncated at the 500th residue. To make the residue embeddings to fit in GPU memory, we downsampled them by a factor of 20, resulting in 25,170 residue embeddings. For amino acid-specific embedding analysis, no downsampling was applied.

Intrinsic dimension estimation Traditionally, I_d can be estimated with Principal Component Analysis or non-linear methods such as Multidimensional Scaling [6]. These methods project the data into a lower-dimensional space. Here, we employ the TwoNN method [9] that does not require projecting the data into lower dimensional space. I_d is estimated from nearest neighbor distances with good speed and performance on non-uniform embedding distributions.

Intrinsic dimension correlation Intrinsic Dimension Correlation (I_dCor) was introduced by Basile et al. [4] to measure the correlation between different data representations. It is defined as:

$$I_d\text{Cor} = \frac{id_1 + id_2 - id_C}{\max\{id_1, id_2\}},$$

where id_1 and id_2 are the I_d of the dataset described by feature vectors from first and second representation methods, respectively. id_C is the I_d of the combined representation obtaining by concatenating the two feature vectors. id_1 and id_2 quantify the information content of the individual representations, while id_C captures the information content of the joint representation. Hence, the numerator reflects the mutual information between the two representations, while the denominator normalizes I_dCor between 0 and 1. Note that I_dCor depends on the method used to compute I_d, and when id_C is significantly underestimated in high-dimensional spaces, I_dCor may not be strictly confined to the range of $[0, 1]$.

To assess the statistical significance of I_dCor estimation, a p-value associated with the null hypothesis that the two data representations are uncorrelated is also calculated, following the method described by Basile et al. [4]. The pairings between the two data representations are randomly shuffled S times, and I_d is calculated for the concatenated features of each shuffle (denoted by $id_S \in \mathbb{R}^S$). The p-value is then defined as $p = \frac{S'+1}{S+1}$, where $S' = |\{id \leq id_C \,|\, id \in id_S\}|$. Intuitively, a high value of S' indicates that the features are highly uncorrelated, leading to a higher id_C. The full algorithm is adapted from Basile et al. [4] with $S = 100$, and we apply a significance threshold of $p < 0.05$ in our analysis.

Table 1: The I_d values for residue and protein embeddings, and the I_dCor between embeddings with p-values. Res. (H) refers to the I_d for histidine residues.

		I_d			I_dCor (p-value)		
Model	Dim.	Residue	Res. (H)	Protein	Residue–Protein	Same Protein	All Protein
ESM-2 (8M)	320	22.6	14.3	12.4	0.65 (0.01)	0.68 (0.01)	0.61 (0.23)
ESM-2 (150M)	640	29.5	16.1	12.5	0.62 (0.01)	0.66 (0.01)	0.60 (0.55)
ESM-2 (650M)	1280	37.0	17.4	12.7	0.59 (0.01)	0.59 (0.01)	0.47 (0.84)
ESM-IF	512	29.6	25.2	17.0	0.62 (0.01)	0.28 (0.13)	0.25 (0.36)
ProstT5 (AA)	1024	23.1	15.7	11.5	0.40 (0.01)	0.00 (0.42)	0.00 (0.45)
ProstT5 (3Di)	1024	19.8	18.2	11.8	0.46 (0.01)	0.00 (0.77)	0.00 (0.13)
MPNN_{V}	128	15.9	13.7	13.3	0.74 (0.01)	0.35 (0.08)	0.33 (0.90)
MPNN_{E}	128	16.7	14.5	9.0	0.63 (0.01)	0.64 (0.01)	0.59 (0.85)
MPNN_{ESV}	384	19.3	14.1	14.4	0.72 (0.01)	0.55 (0.01)	0.00 (0.33)
MPNN_{EXV}	384	19.2	14.8	11.9	0.64 (0.01)	0.58 (0.01)	0.32 (0.86)

4 Results

4.1 Universality and redundancy in protein embeddings

Table 1 presents the I_d and I_dCor values for both residue and protein embeddings. First, despite the large variations in original feature dimensions, the I_d values for residue and protein embeddings remain relatively consistent, with 15–30 for residues and ≈ 10 for proteins. This suggests a universal behavior in how protein modalities are encoded in across the foundation models examined, and the significant dimensionality reduction indicates a high degree of redundancy in the information across different dimensions. Additionally, several models exhibit high I_dCor values, with statistically significant p-values, for residue embeddings randomly selected from the same protein sequence (**Same Protein**). This indicates that a considerable amount of information is shared among residues within the same protein. In contrast, the I_dCor values between two sets of randomly shuffled residues across different proteins (**All Protein**) yield high p-values. This supports our conclusion that the previously observed I_dCor results represent meaningful correlations specifically in the context of the same protein.

Overall, the I_d of residue embeddings is higher than that of the pooled protein embeddings, indicating a potential loss of residue-specific contextual information during the mean pooling process. However, we consistently observe reasonable I_dCor values between residue embeddings and their corresponding protein embeddings (**Residue–Protein**) across all models. This implies that while some information may be lost, a significant portion of the information in the residue embeddings is still retained through mean pooling operations.

When embeddings are clearly clustered into specific contexts, computing the I_ds for each context separately would provide more meaningful results. The I_d for individual amino acid types is lower than the I_d for all residues combined, indicating that computing I_d for specific contexts captures more focused information. The clustering of the residue embeddings, shown in Figures 5 and 6, reveals the types of contexts the models have learned. In the ESM-2 models, residue embeddings are distinctly clustered by amino acid type, while in other models, although the clustering is less pronounced, the embeddings can still be grouped by both amino acid types and 3Di geometric tokens.

Figure 1 ranks the amino acids by their I_d values for each model. Interestingly, the amino acid types with the lowest I_d values are consistent across different models. The amino acids with the lowest I_d values—cysteine (C), tryptophan (W), and histidine (H)—are highly conserved residues. For example, cysteines participate in disulfide bridges, which are crucial for structural stability and conserved across evolutionarily relevant protein families. Thus, their low I_d values align with biological expectations, as conserved positions correspond to less variability in features, given their critical roles in protein structure.

Figure 2 highlights the strong correlations between protein embeddings from different models. In particular, embeddings from the ESM-2 and ProstT5-AA models show very high correlations, which can be attributed to their reliance on similar sequential input types and pretraining tasks that focus on sequence-based objectives. Among the ESM-2 models (8M, 150M, and 650M), which share the same

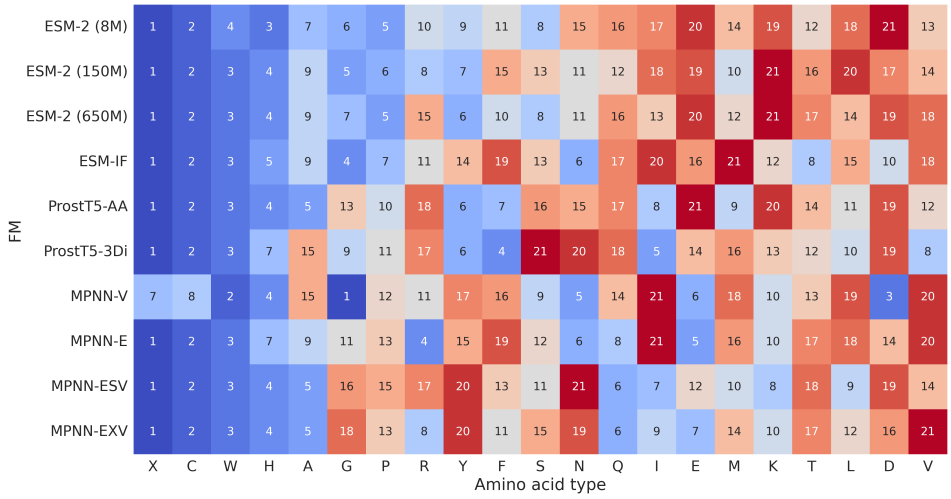
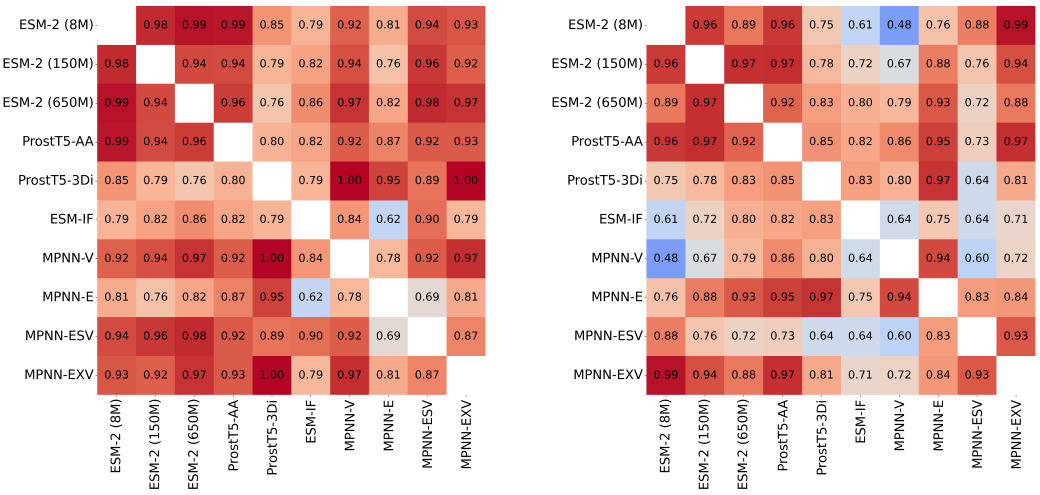


Figure 1: I_d ranking of amino acids for each embedding, listed in ascending order.



(a) Protein I_{d}Cor

(b) Residue (H) I_{d}Cor

Figure 2: Comparison of I_{d}Cor between protein and residue embeddings. (a) Protein I_{d}Cor, (b) Residue (H; histidine) I_{d}Cor. All I_{d}Cor estimates had p-value of 0.01, indicating the significance.

data and task but differ in the number of parameters, the correlations are nearly perfect (>0.94). This reflects the consistency of the embeddings across models that vary in size but are trained in a similar manner. In contrast, the different variants of MPNN (V, E, ESV, EXV) show weaker correlations with each other, suggesting that these models disentangle information differently depending on the type of embeddings they generate.

The correlation patterns for residue embeddings (amino acid: histidine (H)) are similar but generally weaker I_dCor values than those at the protein level. This weaker correlation may be due to residue embeddings capturing local context and fine-grained details that vary more between models, whereas protein embeddings summarize this information globally.

4.2 Local and long-range awareness

We analyzed the long-range correlations of residue embeddings to understand how different models capture sequential and spatial dependencies in proteins. For sequential correlations, we randomly selected ten residues per protein and computed the correlation between their embeddings and those

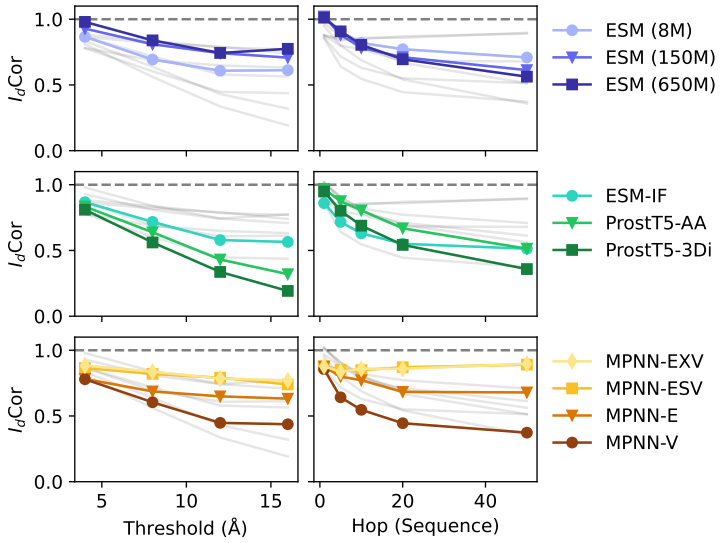


Figure 3: Long-range I_dCor values of residue embeddings with (left) the neighboring residues within the radius threshold and (right) the N-hop neighbors along the sequence.

of their N-hop neighbors along the sequence at distances of $N = 1, 2, 5, 10, 20$, and 50 residues. For spatial correlations, we identified neighboring residues within shells of increasing radii (<4 Å, 4–8 Å, 8–12 Å, and 12–16 Å) and calculated the correlations between their embeddings.

Our findings reveal a general decay in correlation with increasing residue distance for both sequential and spatial metrics. Notably, ProteinMPNN, an inverse folding model based solely on protein geometry, exhibits the highest long-range spatial correlations in its ESV and EXV-type embeddings. Interestingly, its correlations do not decay with sequential distance, reflecting its independence from sequence proximity. The MPNN-V (node) embeddings, however, show a rapid decay, suggesting they primarily capture local geometric features, whereas the MPNN-E (edge) embeddings retain correlations over longer distances, indicating they encode more global geometric relationships.

Similarly, ProstT5-3Di demonstrates a rapid decay in correlations akin to MPNN-V embeddings, and both models exhibit very high protein embedding I_dCor values (Figure 2 (a)), implying they encode similar local geometric information.

ESM-2 models, despite being based solely on sequence information, exhibit strong long-range spatial correlations. This aligns with previous observations that protein language models can unsupervisedly learn contact maps or coevolutionary statistics, which are crucial for predicting protein folds [17]. Notably, the largest ESM-2 model examined (650M) demonstrates the highest long-range spatial correlation among the ESM-2 models studied.

A slow decay in long-range correlation might suggest that a model retains meaningful information between residues over greater distances. However, we also note that consistently high correlations at very large distances may indicate the presence of redundant or unnecessarily correlated features. Determining whether these long-range correlations reflect meaningful interactions or redundancy would be an important direction for future research.

4.3 Understanding mutant embeddings

Understanding mutant embeddings is crucial, as predicting mutant properties is a key downstream task in protein engineering. To understand the structure of the embedding space for mutants, we analyzed 3,127 single and double substitution mutants of the SH3 domain of the Obscurin protein (PDB ID: 1V1C) from the mega-scale protein folding stability dataset in [28]. This dataset includes Deep Mutational Scanning (DMS) assays measuring folding stability.

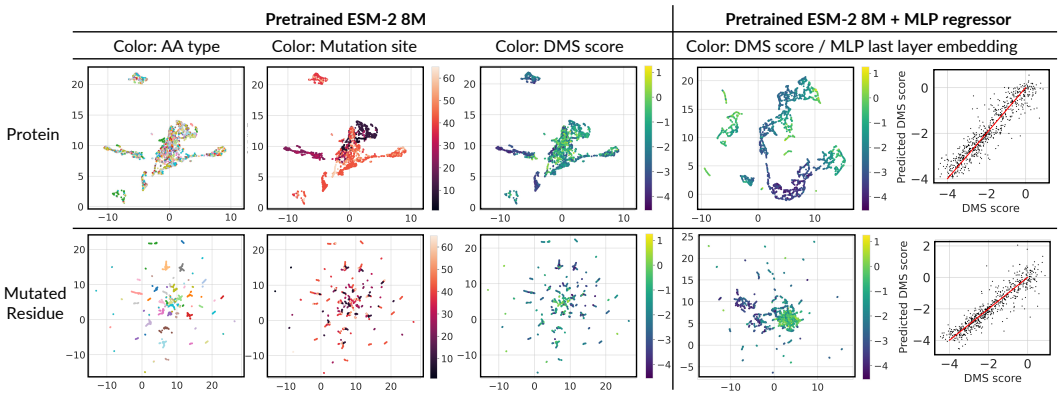


Figure 4: First two axes of UMAP projections [19] of protein and mutated residue embeddings. The scatter plots shows protein (top) and mutated residue (bottom) embeddings from mutants, colored by amino acid type, mutation site, and DMS score. The left column displays embeddings from the pretrained ESM-2 8M model, while the right column shows embeddings of the final layer of MLP regressor for mutation stability prediction. Clustering patterns reflect the model's ability to capture mutation site and DMS score information.

Mutant embeddings in the OBSCN family show significantly lower intrinsic dimensions (I_d = 7.8 on average for the three ESM-2 models) compared to embeddings from diverse proteins, reflecting the reduced variability within a family of mutants.

A key observation from Figure 4 is that mutant protein embeddings are primarily clustered by the mutation sites. Since ESM-2 models are trained with masked language modeling task, which likely encourages the models to learn which parts of the sequence deviate from the 'natural' wild-type sequence—the mutated regions in the mutants. In the OBSCN family, the mutation sites strongly correlates with the DMS score, as in many other mutant families. This alignment, even without specific training, helps explain the good zero-shot and supervised performance of ESM-2 models on the DMS score regression task, as seen in benchmarks like ProteinGym [20].

Embeddings of the mutated residues exhibit an intrinsic dimension of 7.1 (average of the three ESM-2 models), much smaller than the I_d=29.7 of residues from random sequences. In Figure 4, mutated residue embeddings are clustered primarily by amino acid type, consistent with the clustering observed in Figure 5. However, within each amino acid type cluster, embeddings are further aligned based on the mutation sites in the sequence. This suggests that, while amino acid identity is a dominant factor in structuring the embeddings, the context of the mutation—where it occurs in the protein—adds another dimension of high variability. Additionally, the average I_dCor between the mutant residue embeddings and the corresponding mutant protein embeddings is 0.77 across the three ESM-2 models, further indicating that the mutant residue embeddings are also aware of global protein context.

We trained a two-layer Multi-Layer Perceptron (MLP) regressor with ReLU activation (hidden dimension=16) to predict DMS scores from the frozen ESM-2 (8M) embeddings of mutant protein and mutated residues. After training, the I_d of the final layer embeddings dropped to 2.5, suggesting that the model converges on a highly compact and task-specific representation. As shown in Figure 4, after fine-tuning, the final layer embeddings align more smoothly with the DMS scores, in line with the reduced I_d.

5 Conclusion

In this study, we analyzed the intrinsic dimension (I_d) and I_d correlation (I_dCor) values of protein and residue embeddings generated by protein foundation models. First, we observed that the alignment of the foundation models with biological principles, with highly conserved residues consistently showing smaller I_d values across models. Second, we found that protein models tend to align with one another, showing high I_dCor and similar I_d scales. Third, we identified a degree of long-range awareness in the foundation models, where sequentially and spatially proximal residues show higher

correlation than distal ones. The varying rates of correlation decay could be attributed to the specific architectures and pretraining tasks of the models. These findings show that I_d and I_dCor analysis provide valuable insights into the internal structure of protein embeddings.

Our analysis of mutant family embeddings further reveals that mutant embeddings focus on the differences between mutants, resulting in smaller I_d values. The primary source of variability is the location of the mutation, which is consistent with the embedding clusters based on their mutation site. Fine-tuning experiments reveal that the I_d values of embeddings decrease further when models are optimized for specific downstream tasks, suggesting that fine-tuning allows models to simplify their representations to focus on task-relevant information.

Future works. We will expand the analysis to include more protein models as well as other biomolecular modalities, such as DNA and RNA models. Cross-modal I_dCor analysis could help quantify the information shared between protein and their corresponding DNA sequence embeddings, for example. Another interesting direction is to track the evolution of I_d during fine–tuning to gain insights into how models adapt to specific tasks and suggest training strategies accordingly.

References

[1] E. C. Alley, G. Khimulya, S. Biswas, M. AlQuraishi, and G. M. Church. Unified rational protein engineering with sequence-based deep representation learning. *Nature Methods*, 16 (12):1315–1322, Dec. 2019. ISSN 1548-7105. doi: 10.1038/s41592-019-0598-1. URL https://www.nature.com/articles/s41592-019-0598-1. Publisher: Nature Publishing Group.

[2] A. Ansuini, A. Laio, J. H. Macke, and D. Zoccolan. Intrinsic dimension of data representations in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://papers.nips.cc/paper_files/paper/2019/hash/cfcce0621b49c983991ead4c3d4d3b6b-Abstract.html.

[3] Y. Bahri, E. Dyer, J. Kaplan, J. Lee, and U. Sharma. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27):e2311878121, July 2024. doi: 10.1073/pnas.2311878121. URL https://www.pnas.org/doi/10.1073/pnas.2311878121. Publisher: Proceedings of the National Academy of Sciences.

[4] L. Basile, S. Acevedo, L. Bortolussi, F. Anselmi, and A. Rodriguez. Intrinsic Dimension Correlation: uncovering nonlinear connections in multimodal representations, June 2024. URL http://arxiv.org/abs/2406.15812. arXiv:2406.15812 [cs].

[5] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000.

[6] I. Borg and P. J. F. Groenen. *Modern Multidimensional Scaling*. Springer Series in Statistics. Springer, New York, NY, 2005. ISBN 978-0-387-25150-9. doi: 10.1007/0-387-28981-X. URL http://link.springer.com/10.1007/0-387-28981-X.

[7] J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. Wicky, A. Courbet, R. J. de Haas, N. Bethel, et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022.

[8] A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, D. Bhowmik, and B. Rost. ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Learning, July 2020. URL http://biorxiv.org/lookup/doi/10.1101/2020.07.12.199554.

[9] E. Facco, M. d'Errico, A. Rodriguez, and A. Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7(1):12140, Sept. 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-11873-y. URL https://www.nature.com/articles/s41598-017-11873-y. Publisher: Nature Publishing Group.

[10] M. Heinzinger, K. Weissenow, J. G. Sanchez, M. Steinegger, and B. Rost. ProstT5: Bilingual Language Model for Protein Sequence and Structure.

[11] C. Hsu, R. Verkuil, J. Liu, Z. Lin, B. Hie, T. Sercu, A. Lerer, and A. Rives. Learning inverse folding from millions of predicted structures, Apr. 2022. URL http://biorxiv.org/lookup/doi/10.1101/2022.04.10.487779.

[12] B. G. Iovino, H. Tang, and Y. Ye. Protein domain embeddings for fast and accurate similarity search, Feb. 2024. URL https://www.biorxiv.org/content/10.1101/2023.11.27.567555v2. Pages: 2023.11.27.567555 Section: New Results.

[13] M. Jazayeri and S. Ostojic. Interpreting neural computations by examining intrinsic and embedding dimensionality of neural activity. *Current opinion in neurobiology*, 70:113–120, 2021.

[14] B. Jing, S. Eismann, P. Suriana, R. J. L. Townshend, and R. Dror. Learning from protein structure with geometric vector perceptrons. In *International Conference on Learning Representations*, 2020.

[15] N. Konz and M. A. Mazurowski. The Effect of Intrinsic Dataset Properties on Generalization: Unraveling Learning Differences Between Natural and Medical Images, Feb. 2024. URL http://arxiv.org/abs/2401.08865. arXiv:2401.08865 [cs, eess, stat].

[16] S. Kornblith, M. Norouzi, H. Lee, and G. Hinton. Similarity of Neural Network Representations Revisited. In *Proceedings of the 36th International Conference on Machine Learning*, pages 3519–3529. PMLR, May 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html. ISSN: 2640-3498.

[17] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, and Y. Shmueli. Evolutionary-scale prediction of atomic level protein structure with a language model.

[18] P. Lorenz, R. L. Durall, and J. Keuper. Detecting images generated by deep diffusion models using their local intrinsic dimensionality. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 448–459, 2023.

[19] L. McInnes, J. Healy, and J. Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, Sept. 2020. URL http://arxiv.org/abs/1802.03426. arXiv:1802.03426 [cs, stat].

[20] P. Notin, A. W. Kollasch, D. Ritter, L. V. Niekerk, S. Paul, H. Spinner, N. J. Rollins, A. Shaw, R. Weitzman, J. Frazer, M. Dias, D. Franceschi, R. Orenbuch, Y. Gal, and D. S. Marks. ProteinGym: Large-Scale Benchmarks for Protein Fitness Prediction and Design. Nov. 2023. URL https://openreview.net/forum?id=URoZHqAohf&referrer=%5Bthe%20profile%20of%20Yarin%20Gal%5D(%2Fprofile%3Fid%3D~Yarin_Gal1).

[21] P. Pope, C. Zhu, A. Abdelkader, M. Goldblum, and T. Goldstein. The intrinsic dimension of images and its impact on learning, 2021. URL https://arxiv.org/abs/2104.08894.

[22] M. Raghu, J. Gilmer, J. Yosinski, and J. Sohl-Dickstein. SVCCA: Singular Vector Canonical Correlation Analysis for Deep Learning Dynamics and Interpretability. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://papers.nips.cc/paper_files/paper/2017/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html.

[23] R. Rao, J. Meier, T. Sercu, S. Ovchinnikov, and A. Rives. Transformer protein language models are unsupervised structure learners, Dec. 2020. URL https://www.biorxiv.org/content/10.1101/2020.12.15.422761v1. Pages: 2020.12.15.422761 Section: New Results.

[24] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, and R. 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, Apr. 2021. doi: 10.1073/pnas.2016239118. URL https://www.pnas.org/doi/full/10.1073/pnas.2016239118. Publisher: Proceedings of the National Academy of Sciences.

[25] R. Schmirler, M. Heinzinger, and B. Rost. Fine-tuning protein language models boosts predictions across diverse tasks. *Nature Communications*, 15(1):7407, Aug. 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-51844-2. URL https://www.nature.com/articles/s41467-024-51844-2. Publisher: Nature Publishing Group.

[26] U. Sharma and J. Kaplan. A neural scaling law from the dimension of the data manifold, 2020. URL https://arxiv.org/abs/2004.10802.

[27] G. J. Székely, M. L. Rizzo, and N. K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, Dec. 2007. ISSN 0090-5364, 2168-8966. doi: 10.1214/009053607000000505. URL https://projecteuclid.org/journals/annals-of-statistics/volume-35/issue-6/Measuring-and-testing-dependence-by-correlation-of-distances/10.1214/009053607000000505.full. Publisher: Institute of Mathematical Statistics.

[28] K. Tsuboyama, J. Dauparas, J. Chen, E. Laine, Y. Mohseni Behbahani, J. J. Weinstein, N. M. Mangan, S. Ovchinnikov, and G. J. Rocklin. Mega-scale experimental analysis of protein folding stability in biology and design. *Nature*, 620(7973):434–444, 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06328-6. URL https://doi.org/10.1038/s41586-023-06328-6.

[29] UniProt Consortium. UniProt: a hub for protein information. *Nucleic acids research*, 43(D1): D204–D212, 2015.

[30] L. Valeriani, D. Doimo, F. Cuturello, A. Laio, A. Ansuini, and A. Cazzaniga. The geometry of hidden representations of large transformer models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, pages 51234–51252, Red Hook, NY, USA, May 2024. Curran Associates Inc.

[31] X. Zhen, Z. Meng, R. Chakraborty, and V. Singh. On the Versatile Uses of Partial Distance Correlation in Deep Learning, Nov. 2022. URL http://arxiv.org/abs/2207.09684. arXiv:2207.09684 [cs].

A Appendix

A.1 Clustering analysis

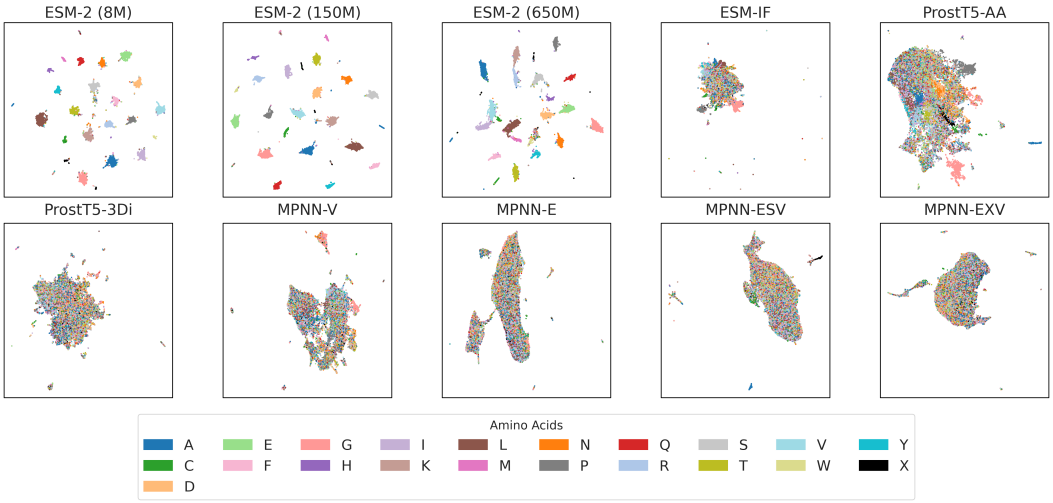


Figure 5: UMAP projection [19] of residue embeddings, colored by amino acid type.

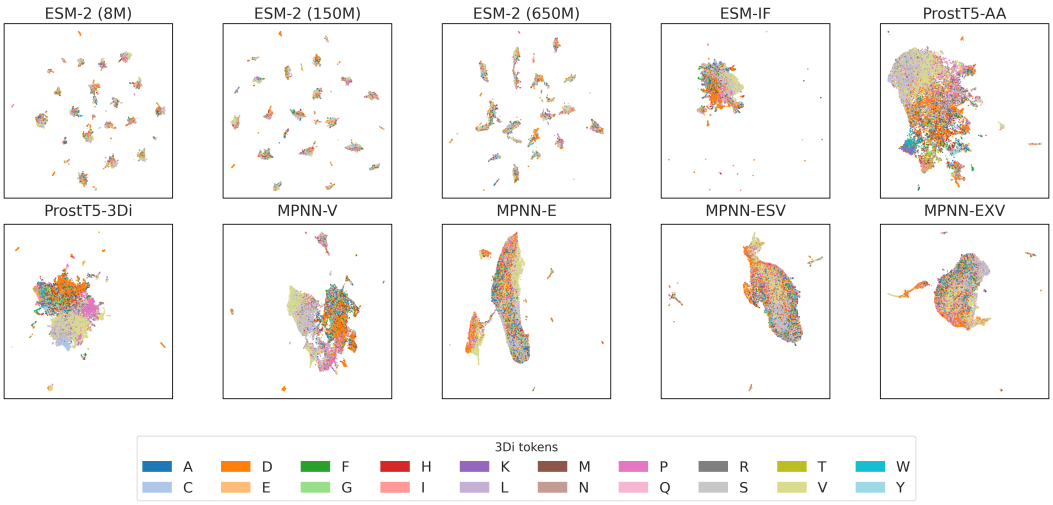


Figure 6: UMAP projection of residue embeddings, colored by 3Di token.

In this section, we show clustering results of the embeddings and discuss their agreement with the intrinsic dimension analysis.

In Figures 5 and 6, ESM-2 embeddings show clear clustering into distinct amino acid types. This suggests that the model has learned strong, distinct representations for each amino acid. This outcome can be attributed to the masked language model (MLM) training objective of ESM-2. The training process likely encourages the model to learn the differences between amino acids, as it needs to predict the identity of masked residues based on the sequence context. This predictive task forces the model to separate the embeddings of different amino acids effectively, which explains the clear and well-separated clusters.

On the contrary, models like ProteinMPNN or ProstT5 are trained with structure-based tasks focused on local geometry. These models prioritize learning the spatial relationships between residues and their structural environments, leading to clusters that are organized based more on local geometric features rather than strictly by amino acid type.