TOWARDS UNDERSTANDING THE SHAPE OF REPRESENTATIONS IN PROTEIN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

While protein language models (PLMs) are one of the most promising avenues of research for future de novo protein design, the way in which they transform sequences to hidden representations, as well as the information encoded in such representations is yet to be fully understood. Several works have attempted to propose interpretability tools for PLMs, but they have focused on understanding how individual sequences are transformed by such models. Therefore, the way in which PLMs transform the whole space of sequences along with their relations is still unknown. In this work we attempt to understand this transformed space of sequences by identifying protein structure and representation with square-root velocity (SRV) representations and graph filtrations. Both approaches naturally lead to a metric space in which pairs of proteins or protein representations can be compared with each other.

We analyze different types of proteins from the SCOP dataset and show that the Karcher mean and effective dimension of the SRV shape space follow a non-linear pattern as a function of the layers in ESM2 models of different sizes. Furthermore, we use graph filtrations as a tool to study the context lengths at which models encode the structural features of proteins. We find that PLMs preferentially encode immediate as well as local relations between residues, but start to degrade for larger context lengths. The most structurally faithful encoding tends to occur close to, but before the last layer of the models, indicating that training a folding model ontop of these layers might lead to improved folding performance.

1 INTRODUCTION

Protein language models (PLMs) are a novel and powerful approach for the modeling and design of new proteins with desired structural or functional features (Ferruz & Höcker, 2022). Using PLMs one can predict the folding of proteins in 3D space (Lin et al., 2023), generate new candidate sequences for viral vectors (Lyu et al., 2024), enzymes (Madani et al., 2023), biosensors (Hayes et al., 2025) and functional binders (Chen et al., 2025; Bryant & Elofsson, 2023). A particular setting in which PLMs have been found to be useful is in finding high-dimensional representations of protein sequences that are thought to reflect the physical, evolutionary or functional properties of a protein. This is especially useful since it allows one to efficiently evaluate and compare newly generated proteins without the need for expensive modeling or experiments. One of the most widespread models of this type are the Evolutionary Scale Modeling (ESM) models (Lin et al., 2023; Hayes et al., 2025), which we will analyze in this work.

It is empirically understood that the representations in these models form a good initialization for folding models (Lin et al., 2023) and can also be used as a reward function to guide other generative approaches (Wang et al., 2025). However, the precise features that these representations encode are still not fully understood. In a previous work by Zhang et al. (2024), the authors develop a categorical Jacobian approach to suggest that PLMs encode the pairwise statistics of coevolving residues. In another work by Simon & Zou (2024), the authors use sparse autoencoders to suggest that PLMs encode human-interpretable features such as binding, structural motifs and functional domains. In yet another work that leverages the power of sparse autoencoders (Gujral et al., 2025), the authors show that PLMs and the individual neurons in them also encode features related to biologically relevant terms within the Gene Ontology hierarchy (Aleksander et al., 2023).

What all of these works study is the transformation of individual sequences to high-dimensional latent representations. However, such approaches ignore how different proteins or their representations relate to each other in the latent space of a PLM. If structure determines function, then it is reasonable to assume that similar structure determines similar function. Furthermore, if protein representations in PLMs characterize structural, evolutionary and functional features, then one might expect that similar representations share such features. In addition, the representation of a protein in a PLM is a tensor of shape *amino acids*$\times$*model dimension*, however for many applications it is standard to take the average over the first dimension, which ignores the shape of the representation, thereby missing the full richness of the information present in PLM representations.

To state our motivation more formally, it is important to understand how the metric space of proteins compares to the metric space of complete shape representations in PLMs. Metric space approaches for the structural analysis of proteins have previously been considered within the rich field of shape analysis (Liu et al., 2010). The essential realization of these approaches is that two proteins can be structurally compared by finding the optimal way to superimpose them, allowing different types of transformations. This is the backbone for many standard tools in the field of rigid and flexible structural alignment such as root mean square deviation (RMSD), TM score (Zhang & Skolnick, 2004) and FATCAT (Li et al., 2020) among others. Despite their popularity, to our knowledge, such methods have yet to be applied to study the hidden representations of PLMs.

In this work, we adapt and extend the shape-analysis framework to study the layerwise metric representation spaces of eight different classes in the SCOP (Chandonia et al., 2022) dataset pushed through several ESM2 models. We consider different features such as the Fréchet radius and the effective dimensionality of these spaces and show that they follow a peculiar pattern as a function of layer, which is especially prominent for larger models. Furthermore, we introduce a graph-filtration method that allows us to separate and study the scale at which PLMs best maintain the structure of proteins. Using this analysis, we show that while structure is always encoded better than chance in deep PLM representations, it is optimally encoded at very short context lengths of 2 or at slightly longer context lengths at ~ 8 amino-acid neighbors.

2 BACKGROUND

In this section, we fix our notation and introduce the mathematical machinery necessary for comparing proteins and their corresponding representations in protein language models. We define two different ways to compare proteins with PLM representations. In the first one, each protein is a point in a high-dimensional (sometimes infinite-dimensional) space, and we define a metric that can be consistently applied to proteins made up of different numbers of amino acids. In the second approach, we define a filtration of metric spaces that only allows us to compare proteins of the same size. As we will argue, this is useful for understanding the context length that current PLMs are sensitive to, as well as the degree to which protein structure is encoded in PLMs.

We start by outlining several ways to mathematically formulate what proteins are. For each such definition of a "protein" we also propose a metric space in which such proteins can be sensibly embedded in.

1. *Proteins as sequences*: In this case, we define a protein P by its amino acid sequence. So, a protein of length L will live in a space $\mathcal{A}^L$, where $\mathcal{A}$ is an alphabet of the 20 canonical amino acids. Given that we want to compare proteins with different numbers of amino acids, we can define the space of all possible amino acid sequences as $\mathcal{A}^* = \bigcup_{L=0}^{\infty} \mathcal{A}^L$. An edit distance, such as the Levenshtein distance for example, can be used to define a metric on this space. Thus, the metric space defined by amino acid sequences will be the pair $(\mathcal{A}^*, d_{lev})$.
2. *Proteins as three dimensional point clouds*: If we consider the actual physical structure of a protein, we can define a protein as an ordered point cloud of size L in $\mathbb{R}^3$ equipped with the Euclidean metric. However, $\mathbb{R}^3$ contains points rather than point clouds, so point clouds of arbitrary sizes live in a different space, namely $\mathcal{P}_3^* = \bigsqcup_{n=0}^{\infty} (\mathbb{R}^3)^n$. While it is possible to define a metric on this space, such metrics (for example, Hausdorff or Wasserstein) often

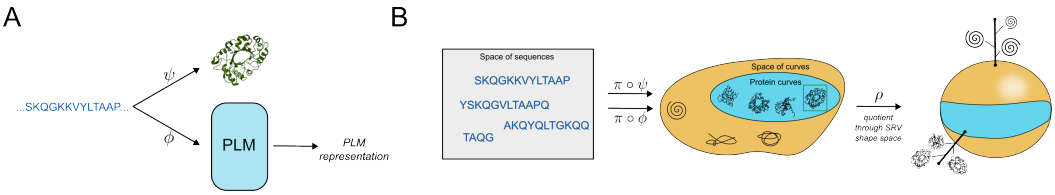

Figure 1: **A**) Depiction of a single sequence (1kr1) mapped to a 3d structure and a PLM representation. **B**) Illustration of how the space of sequences is first transformed into a space of curves and afterwards into a shape space. The lines sticking out of the sphere show the fibers of the map ρ.

ignore the ordering of the point cloud, which matters when studying proteins. Note that each protein can be described through a map $\psi : \mathcal{A}^* \to \mathcal{P}_3^*$ from the sequence space to L points in $\mathbb{R}^3$.

3. *Proteins as curves*: Instead, we can study structure by identifying proteins with curves in $\mathbb{R}^3$. This is done by identifying the space of proteins with the space of continuous curves $\Gamma_3 = \{\gamma : \gamma : [0,1] \to \mathbb{R}^3\}$. We further require that for every protein of length L and coordinates $\psi_l(P) \in \mathbb{R}^3$ we have $\gamma(tL) = \psi_l(P)$ if $l = tL \in \mathbb{N}$. As we will discuss later, this definition of a protein treats proteins of any length as the same object and makes it easier to define a metric that compares them.
4. *Proteins as graphs*: Finally, proteins can also often be though of as graphs. This usually requires a contact map, which is a binary matrix that encodes whether two residues are closer than a specified threshold (6-12Å being a standard choice). We define this approach more rigorously in subsection 2.2.

Notice that while we have defined different ways to think of proteins, the second and third definitions can also be generalized to any ordered point cloud. Therefore, we can use them to study 3d protein structure as well as the embeddings of sequences in protein language models. We will denote the map from an amino acid sequence to the embedding space of a language model by the function $\phi : \mathcal{A}^* \to \mathcal{P}_m^*$, where $m >> 3$ is the embedding dimension of the model. Thus, our approach can be summarized by Figure 1 and the following diagram,

$$\begin{array}{ccc} \mathcal{A}^* & \xrightarrow{\phi} & \mathcal{P}_m^* \\ \psi \downarrow & & \downarrow \pi \\ \mathcal{P}_3^* & \xrightarrow{\pi} & \Gamma_m \end{array} \tag{1}$$

where π is a map that sends point clouds in different spaces to a shared metric space of curves in $\mathbb{R}^m$. Practically, one can think of π as a choice of how to interpolate ordered point clouds so that they become curves in $\mathbb{R}^m$.

2.1 Shape spaces and the square-root velocity representation

A fundamental feature of protein structure is invariance to rotations and translations. What this implies is that protein structure is invariant under isometries of $\mathbb{R}^3$ or the special Euclidean group $SE(3)$. This is the principle behind models such as the SE(3)-transformer (Fuchs et al., 2020), which is equivariant to transformations from this group and is thought to play an important role in the success in folding models such as AlphaFold2 (Jumper et al., 2021) and RoseTTAFold (Baek et al., 2021).

In the present work we use the square-root velocity (SRV) framework introduced in (Srivastava et al., 2010) to enforce invariance to translation. For any ordered point cloud, we interpolate the points with quadratic splines, thereby generating a curve γ which has the same length independently of the number of amino-acids in a protein sequence. Following the interpolation step, the SRV representation is defined as,

$$q(t) = \dot{\gamma}(t)/\sqrt{||\dot{\gamma}(t)||}. \tag{2}$$

From the norm in the denominator one can see that this approach projects curves to an infinite-dimensional sphere S^∞, which makes computing geodesics, and thereby measuring distances, straightforward. Since translations are already accounted for, we quotient out the remaining actions of the $SE(m)$ group, namely the rotations generated by $SO(m)$. This is done using SVD to solve the optimization problem for any two SRV curves q_1 and q_2,

$$\hat{R} = \underset{R \in SO(n)}{\arg\min} ||q_1 - Rq_2||. \tag{3}$$

Given this, the distance between any two curves is defined in terms of the L2 norm as $d(q_1, q_2) = ||q_1 - \hat{R}q_2||_2$. It is often the case that one also removes different reparameterizations of curves, but we have avoided that step given that interpolating between residues forms a consistent way of parameterizing proteins, and this additional step is not worth the computational cost.

In summary, we map sequences to their 3d structure and to their representations in a PLM. These form point clouds which we interpolate to map to curves that are then transformed to their SRV representation. Finally, we quotient out rotations to form the shape space, which inherits a Riemannian structure and allows for the efficient computation of distances. We denote this procedure by a quotient map $\rho : \Gamma_m \to \Gamma_m/SE(m)$. Notice that the space of protein curves under this map form a submanifold of the quotient space $H = S^\infty/SO(m)$ as illustrated in panel **B** of Figure 1. All curves that are the same up to actions of $SE(m)$ end up at a single point y on this submanifold. The set of these curves is called a fiber and is defined by $\rho^{-1}(y) = \{\gamma | \rho(\gamma) = y\}$.

2.2 Graph filtrations

When making predictions, language models have to consider the context within which a token exists. The context length to which a model is sensitive is unknown apriori. Furthermore, while using a contact map with a threshold of 6-12Å makes sense in real 3d protein structure, it is less clear how to choose such a threshold in PLM representations where such a unit is not defined. For this reason, one needs to work with methods that are capable of considering many possible context lengths. One such example is the concept of a filtration, which is fundamental in the topological data analysis literature (Edelsbrunner & Harer, 2010) and has been extended to graphs within the field of graph learning (Hofer et al., 2020; O'Bray et al., 2021).

In a filtration, one defines a parameter t and a family of objects $S_t \subset S_{t'}$ whenever $t < t'$. In this specific case, we consider filtrations of graphs composed of a set of vertices $V = \{v_1, v_2, ..., v_N\}$ with coordinates $v_i \in \mathbb{R}^m$ and edges whenever $d_2(v_i, v_j) < t$, with d_2 being the Euclidean metric enforced by the ambient space. This induces a filtration on their adjacency matrices A^t, since $A^t_{i,j} \leq A^{t'}_{i,j}$ for all i, j when $t < t'$.

Comparing the coordinates of real proteins in $\mathbb{R}^3$ to those of a PLM in $\mathbb{R}^m$ is not a well-defined procedure. However, for the same protein P of length L, both adjacency matrices live in $\{0, 1\}^{L \times L}$. This allows us to compare the structure of a protein with the structure of its embedding in a PLM. A natural choice of a metric in this space is the entry-wise 1-norm of the difference between the two adjacency matrices or $d_A(\psi(P), \phi(P)) = ||\psi(A^t) - \phi(A^t)||_1$, where by abuse of notation $\psi(A^t)$ and $\phi(A^t)$ respectively indicate the flattened adjacency matrix of the three-dimensional structure and the PLM representation at the t-th filtration value.

For our analysis we construct the k nearest neighbor graphs of the true protein structure and PLM representation as shown in Figure 2. For small k, the graphs can only differ at a few locations, and the distance between their adjacency matrices is small. On the other hand, as one increases the number of neighbors, the graphs converge to cliques, and therefore this distance converges to 0. In general, these distances tend to follow a hypergeometric distribution over k. To counteract this, we normalize the distance by the empirical distribution of distances between real proteins and random point clouds $R_i \subset \mathbb{R}^m$. Therefore, for a family of proteins $\mathbf{P} = \{P_1, P_2, ..., P_N\}$, we get a filtration of distance histograms $\{d_A(P_i, \phi(P_i))\}_i^k$. At each level of the filtration we study the normalized first moment, which we will refer to as the *graph filtration moment*. It is given by the equation

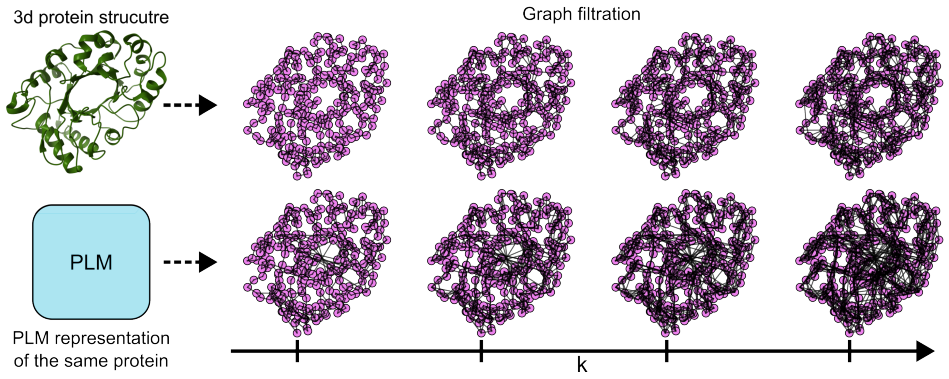

Figure 2: Illustration of how each protein or PLM representation is transformed to a filtration of graphs with a different number of neighbors. The connectivity in the PLM representation is superimposed over the 3d protein graph for clarity.

$$\mathbb{E}_{P_i \in \mathbf{P}}[d(P_i, \phi(P_i))]_{norm} = \frac{\mathbb{E}_{P_i \in \mathbf{P}}[d_A(P_i, \phi(P_i))]}{\mathbb{E}_{P_i \in \mathbf{P}, R_i \in \mathbf{R}}[d_A(P_i, R_i)]}. \tag{4}$$

3 Results

All of the following analysis is based on 1098 randomly sampled protein structures from the SCOP dataset (Chandonia et al., 2022). We sampled up to 200 proteins (some proteins were excluded due to missing a pdb file) from each of the following protein classes – [Alpha, Beta, Alpha/Beta, Alpha+Beta, Alpha and Beta, Membrane and cell surface proteins and peptides, Small proteins and Designed proteins]. More information about these protein classes can be found on the SCOP website `https://scop.berkeley.edu/` or in Chandonia et al. (2022).

3.1 Geometry of PLM shape spaces

In section 2.1 we defined a notion of a shape space which inherits a Riemannian structure that allows us to compute distances and also carry out computations in the tangent space of each fiber. With this in hand, we can define and estimate statistics of curves on the shape space as well as make use of tools such as tangent PCA. We therefore track two measures of PLM shape space geometry, namely effective dimension and Fréchet radius. For all computations we use the Geomstats package (Miolane et al., 2020).

For a set of proteins $\mathbf{P} = \{P_1, P_2, ..., P_N\}$, we first interpolate them at 1000 equally spaced points with quadratic splines and then project them to the SRV shape space giving us the set $\mathbf{Y} = \{y_1, y_2, ..., y_N\}$. There is no a priori correct choice of spline order, but quadratic splines are the simplest option that is still differentiable as required by the SRV framework. Very high order splines are expected to add a lot of additional structure to the PLM point clouds and could thus bias the results. For an evaluation of the robustness of our results to spline order and interpolation samples see Figure 6. We compute the Fréchet radius by finding the Fréchet mean $p_F = \underset{x \in H}{\arg\min} \sum d(x, y_i)$ through gradient descent after which the Fréchet radius is defined as,

$$r_F = \mathop{\mathbb{E}}_{y_i \in \mathbf{Y}}[d(y_i, p_F)]. \tag{5}$$

An intuitive way to think of this object is as a measure for how spread out PLM representations are with respect to each other on the shape space. Therefore, a small value indicates that different proteins are represented by similar shapes, whereas a larger value indicates the opposite. As one can see in Figure 3 the Fréchet radius tends to decrease with depth and is much smaller for PLM representations than for real 3d protein structures. Surprisingly enough, it does not seem to vary with model size, indicating that the variability among shapes is low for all models.

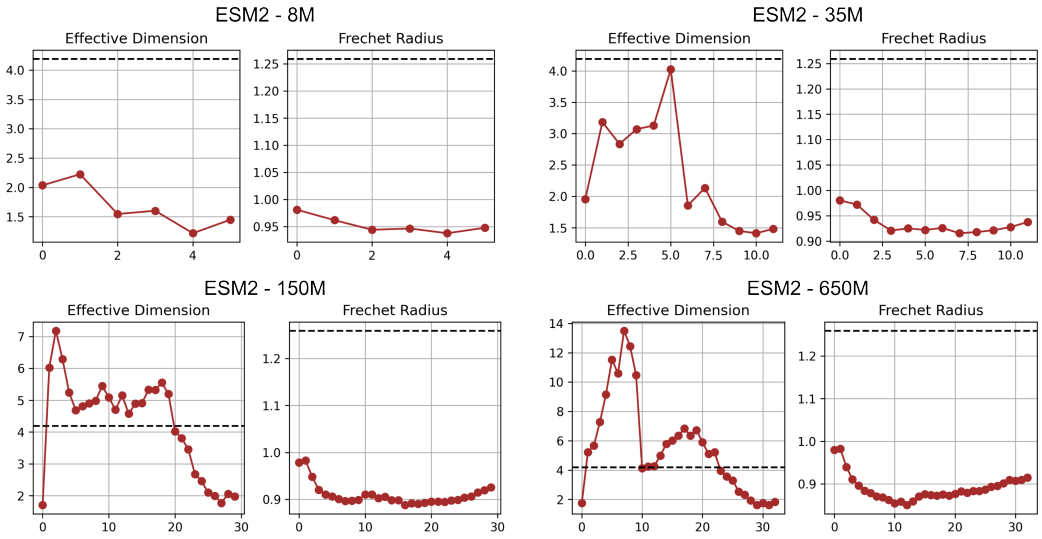

Figure 3: Effective dimension and Fréchet radius for each of the four models as a function of the layers. The black line indicates the value of each metric for the 3d protein structure. As one can see larger models exhibit dimension expansion in the initial layers and contraction later on.

Another measure describing the geometry of shape spaces in PLMs is the dimension of the submanifold on which they live. In Euclidean data this is usually measured by the effective dimension defined through the eigenvalues $\{\lambda_1, \lambda_2, ..\}$ of the covariance matrix of the data or

$$\lambda_{eff} = \frac{(\sum \lambda_k)^2}{\sum \lambda_k^2} \ . \tag{6}$$

This procedure can be extended to curved data through tangent PCA (Abboud et al., 2020) where one first uses the log map to project all data to the tangent space of the Fréchet mean by $z_i = \log_{p_F}(y_i) = \frac{d(y_i,p_F)}{\sin[d(y_i,p_F)]}(y_i - \cos[d(y_i,p_F)]p_F)$ and afterwards applies PCA in this tangent space. A large effective dimension implies that PLM representations explore many different variations in shape, whereas a small effective dimension implies that a few specific variations are enough to describe the differences in PLM representation shapes. Judging by Figure 3, it seems like PLMs go through two regimes – a dimension expansion in the first layers followed by dimension contraction towards the end. Larger models expand the dimension more than the 3d structure baseline, whereas the smaller models stay below it. The largest models even show a second peak in dimension expansion.

It is also worth pointing out that, especially in later layers, this dimensionality is very low and severely different from the dimensionality found through standard PCA on the flattened PLM representation (see Figure 5 in the Appendix). We interpret this to mean that while shapes are encoded in a high-dimensional ambient space, the differences in shapes of PLM representations can be described by just a few shape descriptors. In other words, PLMs encode proteins by similar shapes while spanning many different directions within their ambient space. To see if this trend only occurs in the ESM2 models, we also ran the same analysis on the general purpose Ankh model Elnaggar et al. (2023). As one can see in Figure 7, a similar pattern, though with a more extreme initial dimension expansion phase, can be seen in that model.

3.2 Context length sensitivity of PLMs

While it is interesting to understand the global geometric features of PLM representation shape spaces, language models are known to encode contextual features of text which are hard to relate to the aforementioned measures. In order to better understand how structural context is encoded in PLMs, we use the graph filtration moment defined in Equation 4. This measure indicates how

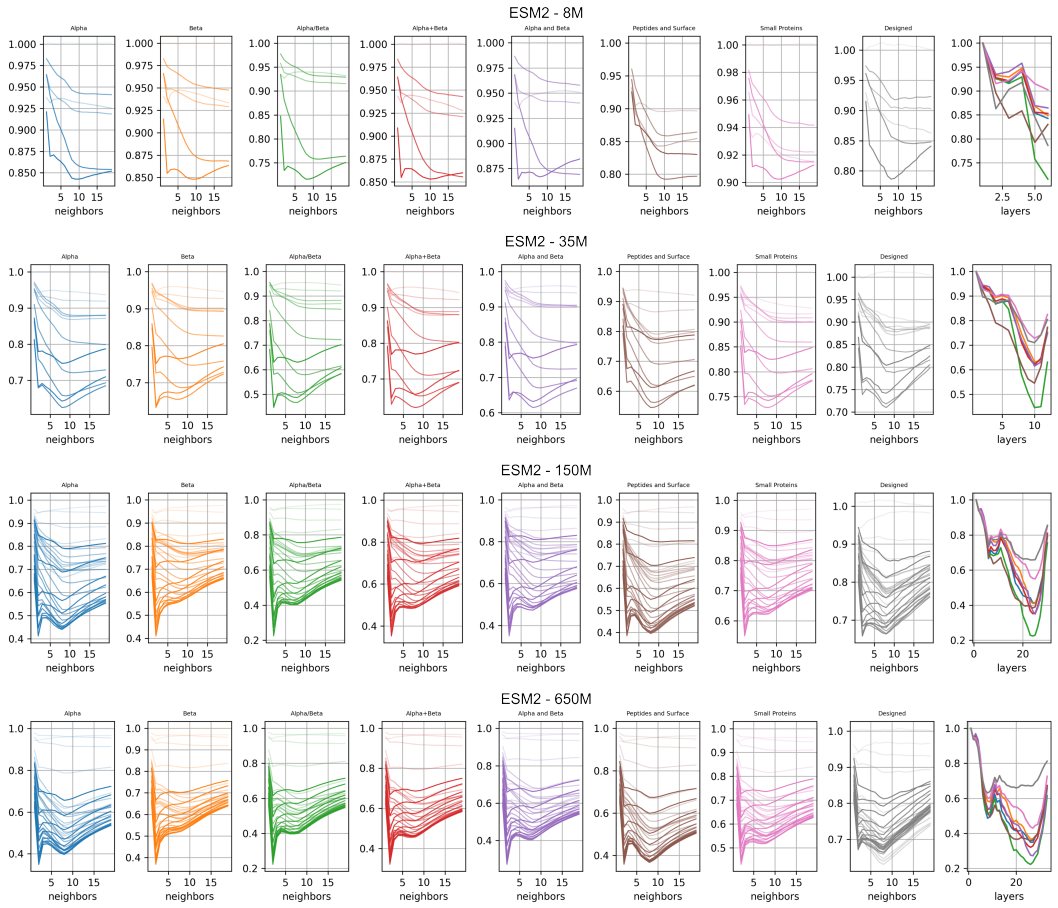

Figure 4: Graph filtration moments for all models and all layers evaluated on different protein classes from the SCOP database. Curves increase in transparency with increasing layer number. The rightmost plots show the minimal value of the filtration as a function of the layer. Each curve is color-coded for protein class in correspondence with the other plots.

similar the shapes of PLM representations are to their 3d protein structure counterparts. A value of 1 or above means that the PLM arranges all residues in a random point cloud, whereas lower values imply that 3d protein structure is encoded in the PLM representation.

In Figure 4 we show the graph filtration moments across all layers of the different ESM2 models on the SCOP protein classes. The curves for shallow layers are more transparent than deep layers and the rightmost plot shows the minimum value across the filtration as a function of depth. Several peculiar patters arise throughout the filtration. The first thing to note is that larger models have a better similarity to 3d protein structure in their intermediate layers, but model size seems to have less of an impact at the last layer. This implies that while encoding protein structure emerges as an important intermediate processing step for unmasking, it is not as important at the last classification step.

The second striking observation is that in all models many of the curves have a bimodal shape throughout the filtration. This implies that PLM representations encode 3d protein structure at both a very local level, at about 2 neighbors, as well as at a slightly less local level, at about 8 neighbors. The improved encoding at 2 neighborhoods has a simple interpretation as PLM representations having similar immediate neighbors to 3d protein structure. However, the second valley at around 8 neighbors is harder to interpret. Given the fact that it is less pronounced in the Beta class, one might speculate that it might be related to representations of Alpha helices, but more work is needed to understand the precise meaning of this feature.

Finally, it is also clear that certain protein classes like Alpha/Beta are represented by shapes that are much more structurally similar to their 3d structure compared to other classes such as small and designed proteins. Furthermore, while initial layers showed a correlation between protein length and the minimum graph filtration moment, later layers did not show such a pattern (see Figure 9). This indicates that local context structure is represented independently of protein length. The results for the Ankh model show a similar pattern as can be seen in Figure 8.

4 DISCUSSION

We have applied two approaches in order to better understand the geometry of shape spaces generated by PLM representations. The first uses SRV representations and quotients out rotations to create a space with Riemannian structure, which allows us to define a metric and generalize statistical methods such as PCA to shapes. The second uses graph filtrations to study how protein structure is encoded in the layers of a PLM at as many levels of resolution as one desires. Given the abstract nature of our results, here we provide a discussion of how they can be understood more intuitively and propose several future directions that would be exciting to pursue.

4.1 EXPANSION AND CONTRACTION IN PLM REPRESENTATION SHAPE SPACES

As shown in Figure 3, the initial layers in PLMs exhibit an expansion in effective dimensionality, whereas later layers contract the shape space to a very low dimensional subspace. Previous work on traditional language models has used the notion of intrinsic dimensionality (Li et al., 2018) and has shown that language models have a remarkably low intrinsic dimensionality relative to model size (Aghajanyan et al., 2020). Similar measures of dimensionality are also thought to relate to task performance (Marbut et al., 2024), training convergence and generalization (Ruppik et al., 2025).

The specific expansion-contraction pattern observed in our estimate of dimensionality is very similar to the behavior seen in Cheng et al. (2024) and Valeriani et al. (2023). The universal appearance of this pattern is thought to correspond to a general high-abstraction regime in the dimension-expansion phase and a specific semantically rich regime in the contraction phase. These properties of language model layers can be effectively used to solve any task.

Our approach looks at the dimensionality of the data within the shape space manifold rather than directly looking at all residue representations in the embedding space. This leads to an arguably more clear interpretation, the initial layers of a PLM represent proteins by shapes that can be flexibly deformed to each other by combining many different non-linear shape transformations. The higher the dimensionality, the more such transformations there are. Therefore, the sharp reduction in dimensionality in later layers means that there are remarkably few transformations (less than in the space of 3d protein shapes) that are needed to efficiently navigate PLM representation shape spaces. Understanding the precise nature of these transformations would be an exciting direction for future work.

4.2 STRUCTURAL ENCODING IN PLM REPRESENTATIONS

In addition to studying the geometry of PLM representation shape spaces, we also looked at how 3d protein structure is encoded in the layers of a PLM by using graph filtrations. We observe a bimodal pattern in which protein structure is optimally encoded at the resolution of very short context lengths of about 2 residues as well as at slightly longer context lengths of around 8 residues. The PLM representations of Alpha/Beta proteins showed by far the highest similarity to their 3d structure, whereas small and designed proteins were represented by more distinct shapes. While outside of the scope of this work, understanding what features of these protein classes determine how much of the structure is encoded by PLMs is an exciting direction for future research.

The finding that PLMs encode 3d protein structure at all is surprising given that PLMs are given masked sequences and are then trained to predict the most likely missing amino acids. There is no point at which PLM representations are incentivized to encode protein structure. Therefore, either learning protein structure is beneficial for the unmasking process or the function learned during unmasking shares some properties (one might say ”correlates”) with the function used in the folding of a sequence to its 3d structure.

Finally, our findings further explain why folding models such as ESMFold benefit from starting with a pretrained PLM that has already partially learned protein structure. Our observation that structure is not optimally encoded in the last layer, but rather in the layers that immediately precede it, can be used to improve initializations for protein models. Given our results, we expect that better folding performance can be achieved if one uses the layers with the optimal structural PLM representation rather than the whole model. Initial attempts to show this with linear models or small networks failed to generalize (data not shown) and verifying this hypothesis would require training and testing larger models. The way in which layerwise representations can be used for folding, along with other functional tasks, is another avenue of future research that would be exciting to explore.

REFERENCES

Michel Abboud, Abdesslam Benzinou, and Kamal Nasreddine. A robust tangent pca via shape restoration for shape variability analysis. *Pattern Analysis and Applications*, 23(2):653–671, 2020.

Armen Aghajanyan, Luke Zettlemoyer, and Sonal Gupta. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. *arXiv preprint arXiv:2012.13255*, 2020.

Suzi A Aleksander, James Balhoff, Seth Carbon, J Michael Cherry, Harold J Drabkin, Dustin Ebert, Marc Feuermann, Pascale Gaudet, Nomi L Harris, et al. The gene ontology knowledgebase in 2023. *Genetics*, 224(1):iyad031, 2023.

Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N Kinch, R Dustin Schaeffer, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021.

Patrick Bryant and Arne Elofsson. Peptide binder design with inverse folding and protein structure prediction. *Communications Chemistry*, 6(1):229, 2023.

John-Marc Chandonia, Lindsey Guan, Shiangyi Lin, Changhua Yu, Naomi K Fox, and Steven E Brenner. Scope: improvements to the structural classification of proteins–extended database to facilitate variant interpretation and machine learning. *Nucleic acids research*, 50(D1):D553–D559, 2022.

Leo Tianlai Chen, Zachary Quinn, Madeleine Dumas, Christina Peng, Lauren Hong, Moises Lopez-Gonzalez, Alexander Mestre, Rio Watson, Sophia Vincoff, Lin Zhao, et al. Target sequence-conditioned design of peptide binders using masked language modeling. *Nature Biotechnology*, pp. 1–9, 2025.

Emily Cheng, Diego Doimo, Corentin Kervadec, Iuri Macocco, Jade Yu, Alessandro Laio, and Marco Baroni. Emergence of a high-dimensional abstraction phase in language transformers. *arXiv preprint arXiv:2405.15471*, 2024.

Herbert Edelsbrunner and John Harer. *Computational Topology: An Introduction*. American Mathematical Soc., 2010.

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.

Noelia Ferruz and Birte Höcker. Controllable protein design with language models. *Nature Machine Intelligence*, 4(6):521–532, 2022.

Fabian Fuchs, Daniel Worrall, Volker Fischer, and Max Welling. Se (3)-transformers: 3d roto-translation equivariant attention networks. *Advances in neural information processing systems*, 33:1970–1981, 2020.

Onkar Gujral, Mihir Bafna, Eric Alm, and Bonnie Berger. Sparse autoencoders uncover biologically interpretable features in protein language model representations. *Proceedings of the National Academy of Sciences*, 122(34):e2506316122, 2025.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025.

Christoph Hofer, Florian Graf, Bastian Rieck, Marc Niethammer, and Roland Kwitt. Graph filtration learning. In *International Conference on Machine Learning*, pp. 4314–4323. PMLR, 2020.

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.

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. *arXiv preprint arXiv:1804.08838*, 2018.

Zhanwen Li, Lukasz Jaroszewski, Mallika Iyer, Mayya Sedova, and Adam Godzik. Fatcat 2.0: towards a better understanding of the structural diversity of proteins. *Nucleic acids research*, 48 (W1):W60–W64, 2020.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Wei Liu, Anuj Srivastava, and Jinfeng Zhang. Protein structure alignment using elastic shape analysis. In *Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology*, pp. 62–70, 2010.

Suyue Lyu, Shahin Sowlati-Hashjin, and Michael Garton. Variational autoencoder for design of synthetic viral vector serotypes. *Nature Machine Intelligence*, 6(2):147–160, 2024.

Ali Madani, Ben Krause, Eric R Greene, Subu Subramanian, Benjamin P Mohr, James M Holton, Jose Luis Olmos Jr, Caiming Xiong, Zachary Z Sun, Richard Socher, et al. Large language models generate functional protein sequences across diverse families. *Nature biotechnology*, 41 (8):1099–1106, 2023.

Anna C Marbut, John W Chandler, and Travis J Wheeler. Exploring the impact of a transformer's latent space geometry on downstream task performance. *arXiv preprint arXiv:2406.12159*, 2024.

Nina Miolane, Nicolas Guigui, Alice Le Brigant, Johan Mathe, Benjamin Hou, Yann Thanwerdas, Stefan Heyder, Olivier Peltre, Niklas Koep, Hadi Zaatiti, Hatem Hajri, Yann Cabanes, Thomas Gerald, Paul Chauchat, Christian Shewmake, Daniel Brooks, Bernhard Kainz, Claire Donnat, Susan Holmes, and Xavier Pennec. Geomstats: A python package for riemannian geometry in machine learning. *Journal of Machine Learning Research*, 21(223):1–9, 2020. URL `http://jmlr.org/papers/v21/19-027.html`.

Leslie O'Bray, Bastian Rieck, and Karsten Borgwardt. Filtration curves for graph representation. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 1267–1275, 2021.

Benjamin Matthias Ruppik, Julius von Rohrscheidt, Carel van Niekerk, Michael Heck, Renato Vukovic, Shutong Feng, Hsien-chin Lin, Nurul Lubis, Bastian Rieck, Marcus Zibrowius, et al. Less is more: Local intrinsic dimensions of contextual language models. *arXiv preprint arXiv:2506.01034*, 2025.

Elana Simon and James Zou. Interplm: Discovering interpretable features in protein language models via sparse autoencoders. *bioRxiv*, pp. 2024–11, 2024.

Anuj Srivastava, Eric Klassen, Shantanu H Joshi, and Ian H Jermyn. Shape analysis of elastic curves in euclidean spaces. *IEEE transactions on pattern analysis and machine intelligence*, 33 (7):1415–1428, 2010.

Lucrezia Valeriani, Diego Doimo, Francesca Cuturello, Alessandro Laio, Alessio Ansuini, and Alberto Cazzaniga. The geometry of hidden representations of large transformer models. *Advances in Neural Information Processing Systems*, 36:51234–51252, 2023.

Yeji Wang, Minghui Song, Fujing Liu, Zhen Liang, Rui Hong, Yuemei Dong, Huaizu Luan, Xiaojie Fu, Wenchang Yuan, Wenjie Fang, et al. Artificial intelligence using a latent diffusion model enables the generation of diverse and potent antimicrobial peptides. *Science Advances*, 11(6): eadp7171, 2025.

Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, 2004.

Zhidian Zhang, Hannah K Wayment-Steele, Garyk Brixi, Haobo Wang, Dorothee Kern, and Sergey Ovchinnikov. Protein language models learn evolutionary statistics of interacting sequence motifs. *Proceedings of the National Academy of Sciences*, 121(45):e2406285121, 2024.

A Appendix

A.1 Additional figures

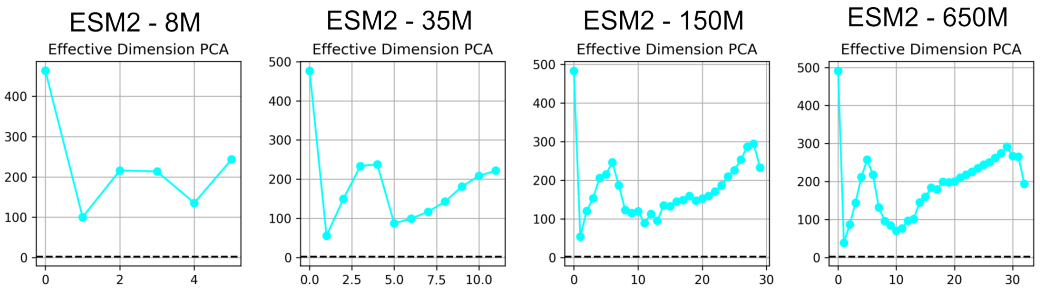

Figure 5: Effective dimension esimated using PCA directly on PLM representations. Each PLM representation is first interpolated at 1000 points. Afterwards each tensor of shape $1000 \times$ *PLM dimension at a layer* is flattened, meaning that the maximum dimension of this space is 1000 times the dimension at the layer.

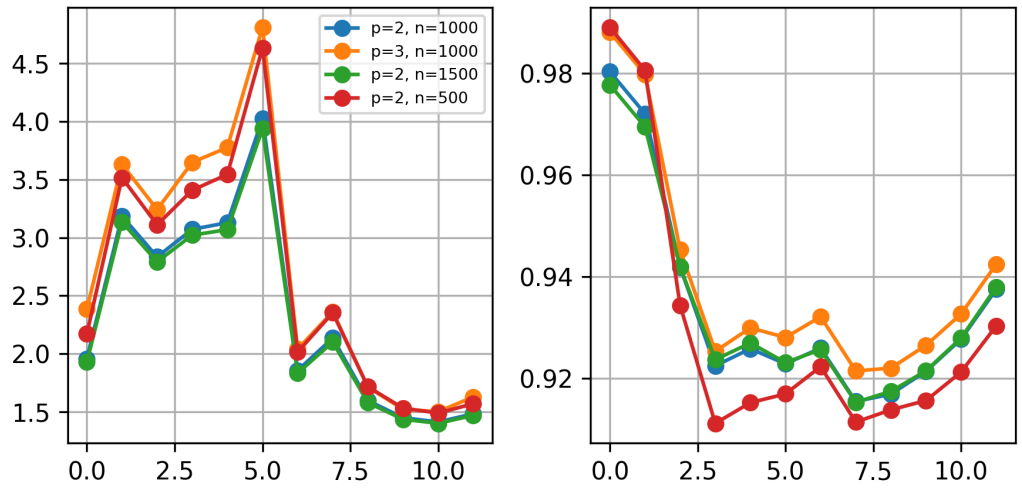

Figure 6: Robustness with respect to interpolation order (indicated by p) and number of sampled points (indicated by n) evaluated on the ESM2 - 35M model.

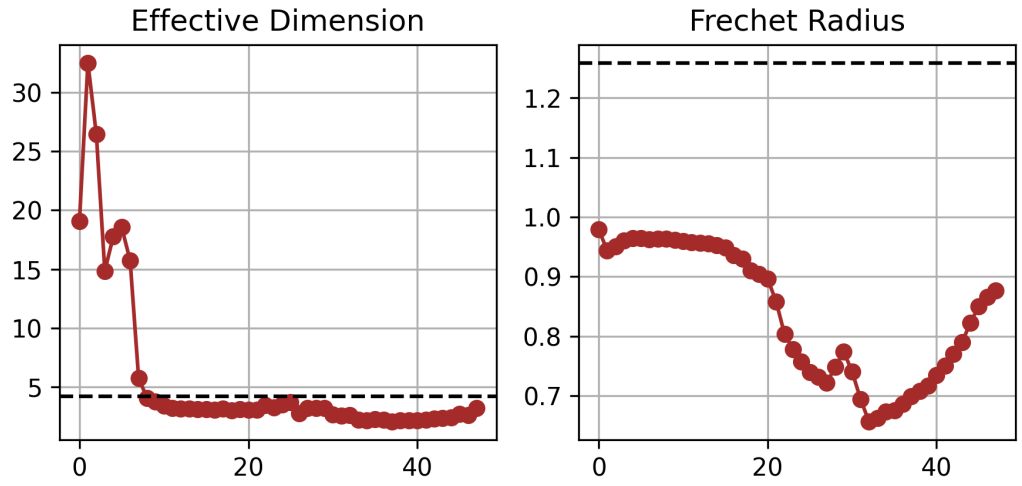

Figure 7: Effective dimension and Fréchet radius for the base Ankh model as a function of the layers.

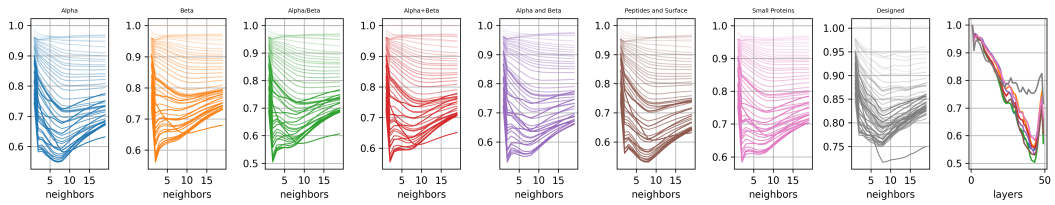

Figure 8: Graph filtration moments for the base Ankh model.

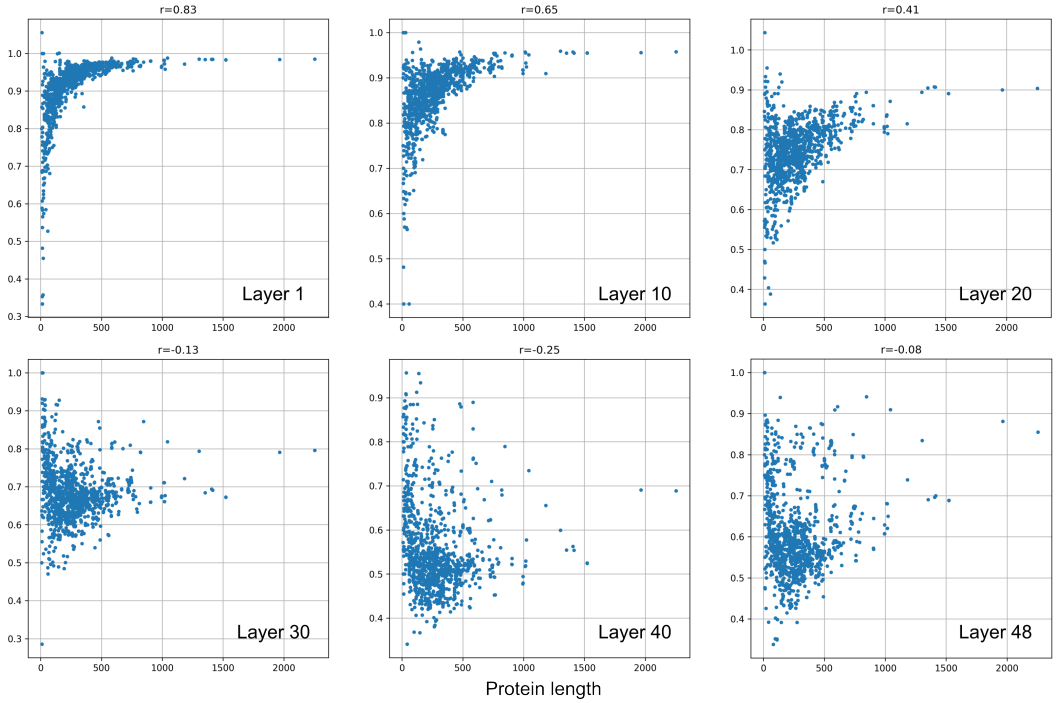

Figure 9: Protein length vs minimum graph filtration moment for several layers of the Ankh model. Values above the plots indicate the Spearman correlation values.