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

LifeLU reading group

presented by Özdeniz Dolu

19.12.2024

Motivation / Core Idea

- Embeddings that are generated by protein foundation models have been successful in a wide variety of downstream tasks.
- The “task-independent” relationship between these embeddings is underexplored.
- Although some attempts at comparison are made, they are generally compared based on the performance on downstream tasks.

Motivation / Core Idea

- This work provides an analysis of embedding spaces based on the estimations of Intrinsic Dimension (Id for short).
- Intrinsic Dimension is defined as the minimum number of variables to represent a dataset. Therefore, it’s a measure of “information content”.
- Authors also used Intrinsic Dimension Correlation (IdCor for short) as a measure of “mutual information” between two datasets.

Key Findings

Universality: Id’s of residue or protein embeddings are consistent across various protein foundation models.

Redundancy: Id’s of residue or protein embeddings are much smaller than embedding dimension.

Local and long-range awareness: Residue embeddings are more correlated with close (spatially or sequentially) proximity residues.

Understanding Mutant Embeddings: A case study of the analysis on mutant embeddings of ESM2 models is provided.

Protein Foundation Models

Embeddings from 4 popular models have been analyzed.

ESM-2 (8M, 150M, 650M): Transformer based encoder. Embeddings are generated based on sequence.

ESMInverse Folding (ESM-IF): GVP-Transformer encoder and transformer decoder. Embeddings are generated based on structure.

ProstT5: Bidirectional transformer encoder-decoder. Can generate embeddings based on either structure or sequence.

ProteinMPNN (MPNN): GNN-based encoder-decoder. Embeddings are generated based on structure.

Methods (Data Preparation)

Structures with missing information is eliminated and out of 16380 (30% sequence identity) cluster representatives from PDB, 4591 are selected.

For each model, embeddings are generated and cached. Depending on the model either the PDB structure or the sequence information is used.

Protein-level embeddings are generated through mean pooling.

Methods (Intrinsic Dimension Estimation)

Various methods for estimation of Id exist. In this work, TwoNN method is employed.

TwoNN is selected because it has good speed and performance and does not requires you to project the data into a lower dimensional space.

A note here: Authors do not provide any analysis or arguments about TwoNN’s assumptions/shortcomings.

Methods (Intrinsic Dimension Correlation)

A measure of “mutual information” introduced by Basile et al.

If id1 and id2 are Id of a dataset calculated from embeddings 1 and 2, idC is the Id calculated from the vector that is the concatenation of embeddings 1 and 2.

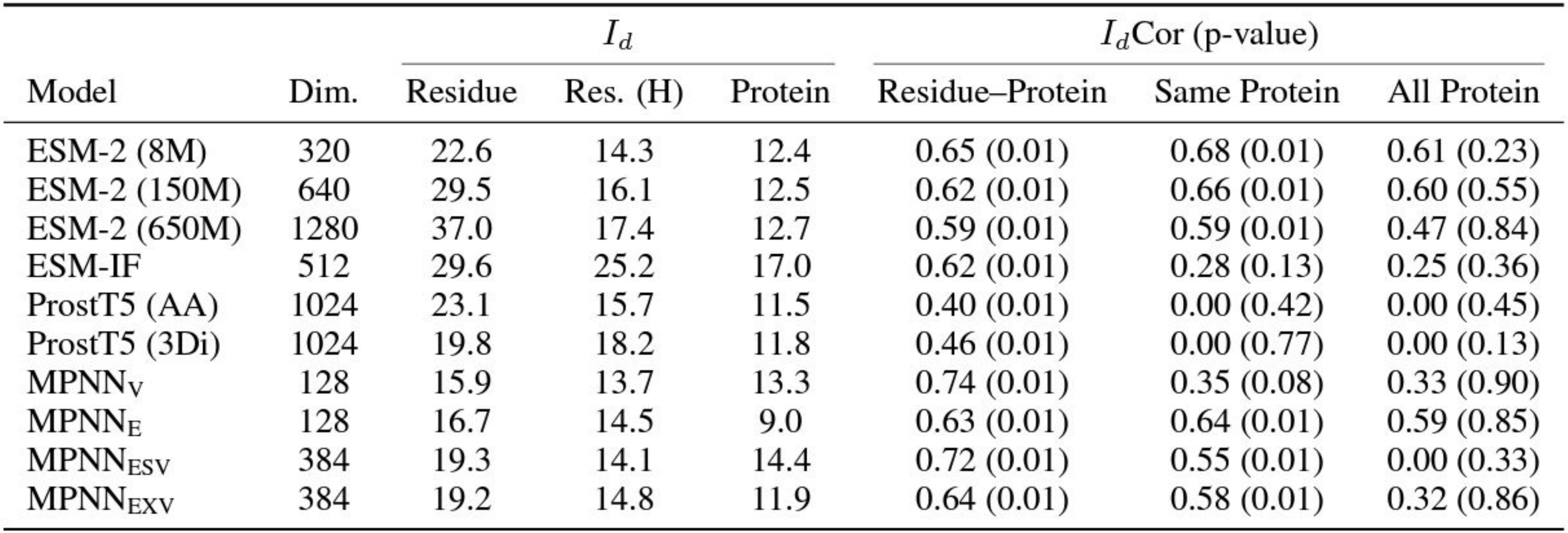
Note that since Id calculation is an estimation, IdCor is also an estimation. Therefore, a p-value calculation is provided for IdCor values.

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

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].

Results

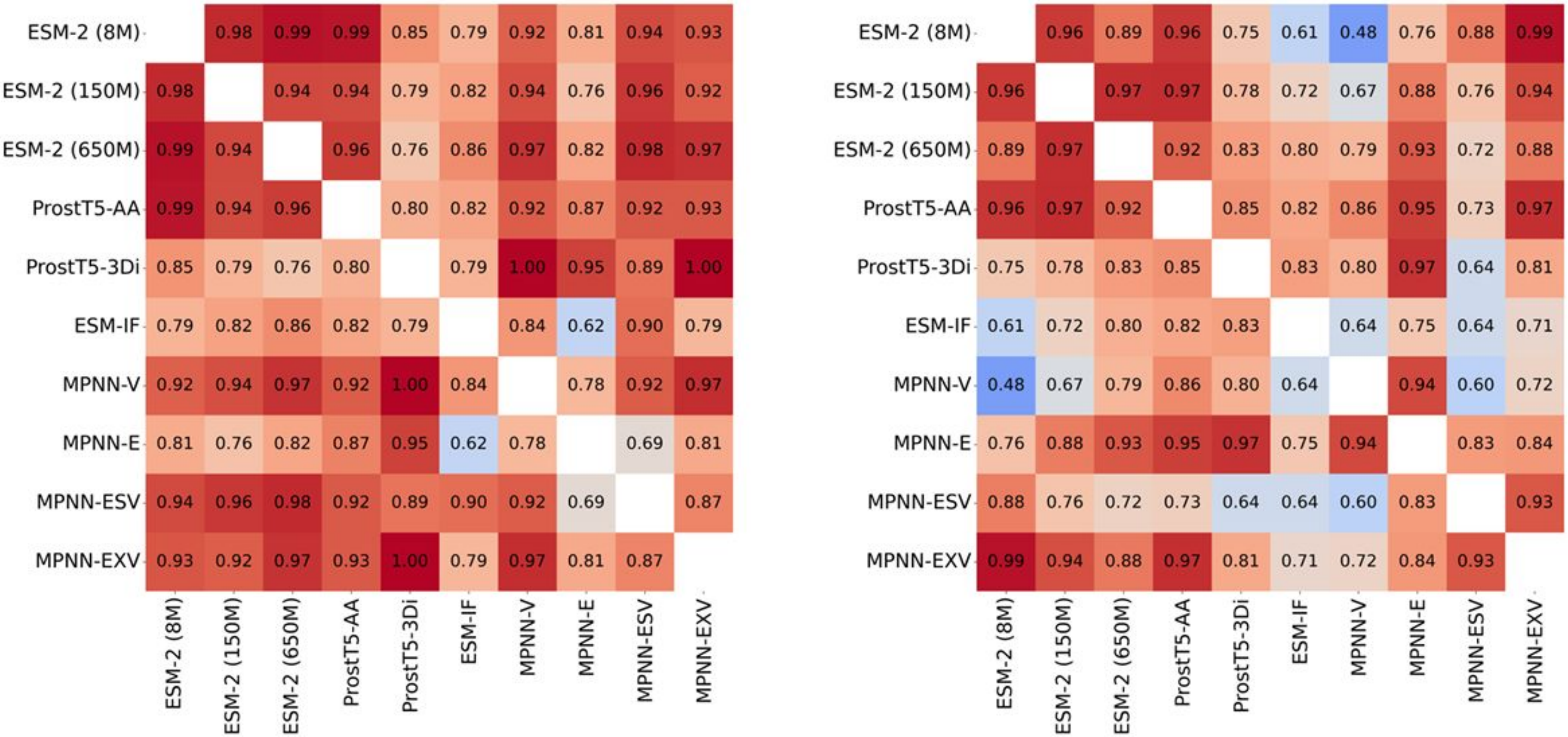
All Protein: two sets of randomly selected residue embeddings from all proteins
Same Protein: two sets of randomly selected residue embeddings from “a protein”
Residue-Protein: between residue embeddings their corresponding protein embeddings



		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$_{\text{V}}$	128	15.9	13.7	13.3	0.74 (0.01)	0.35 (0.08)	0.33 (0.90)
MPNN$_{\text{E}}$	128	16.7	14.5	9.0	0.63 (0.01)	0.64 (0.01)	0.59 (0.85)
MPNN$_{\text{ESV}}$	384	19.3	14.1	14.4	0.72 (0.01)	0.55 (0.01)	0.00 (0.33)
MPNN$_{\text{EXV}}$	384	19.2	14.8	11.9	0.64 (0.01)	0.58 (0.01)	0.32 (0.86)

IdCor values between embeddings of protein foundation models. All of them are significant with $p = 0.01$.

On the left: Protein-level embeddings and on the right: Residue(H) level embeddings



	ESM-2 (8M)	ESM-2 (150M)	ESM-2 (650M)	ProstT5-AA	ProstT5-3Di	ESM-IF	MPNN-V	MPNN-E	MPNN-ESV	MPNN-EXV
ESM-2 (8M)		0.98	0.99	0.99	0.85	0.79	0.92	0.81	0.94	0.93
ESM-2 (150M)	0.98		0.94	0.94	0.79	0.82	0.94	0.76	0.96	0.92
ESM-2 (650M)	0.99	0.94		0.96	0.76	0.86	0.97	0.82	0.98	0.97
ProstT5-AA	0.99	0.94	0.96		0.80	0.82	0.92	0.87	0.92	0.93
ProstT5-3Di	0.85	0.79	0.76	0.80		0.79	1.00	0.95	0.89	1.00
ESM-IF	0.79	0.82	0.86	0.82	0.79		0.84	0.62	0.90	0.79
MPNN-V	0.92	0.94	0.97	0.92	1.00	0.84		0.78	0.92	0.97
MPNN-E	0.81	0.76	0.82	0.87	0.95	0.62	0.78		0.69	0.81
MPNN-ESV	0.94	0.96	0.98	0.92	0.89	0.90	0.92	0.69		0.87
MPNN-EXV	0.93	0.92	0.97	0.93	1.00	0.79	0.97	0.81	0.87	

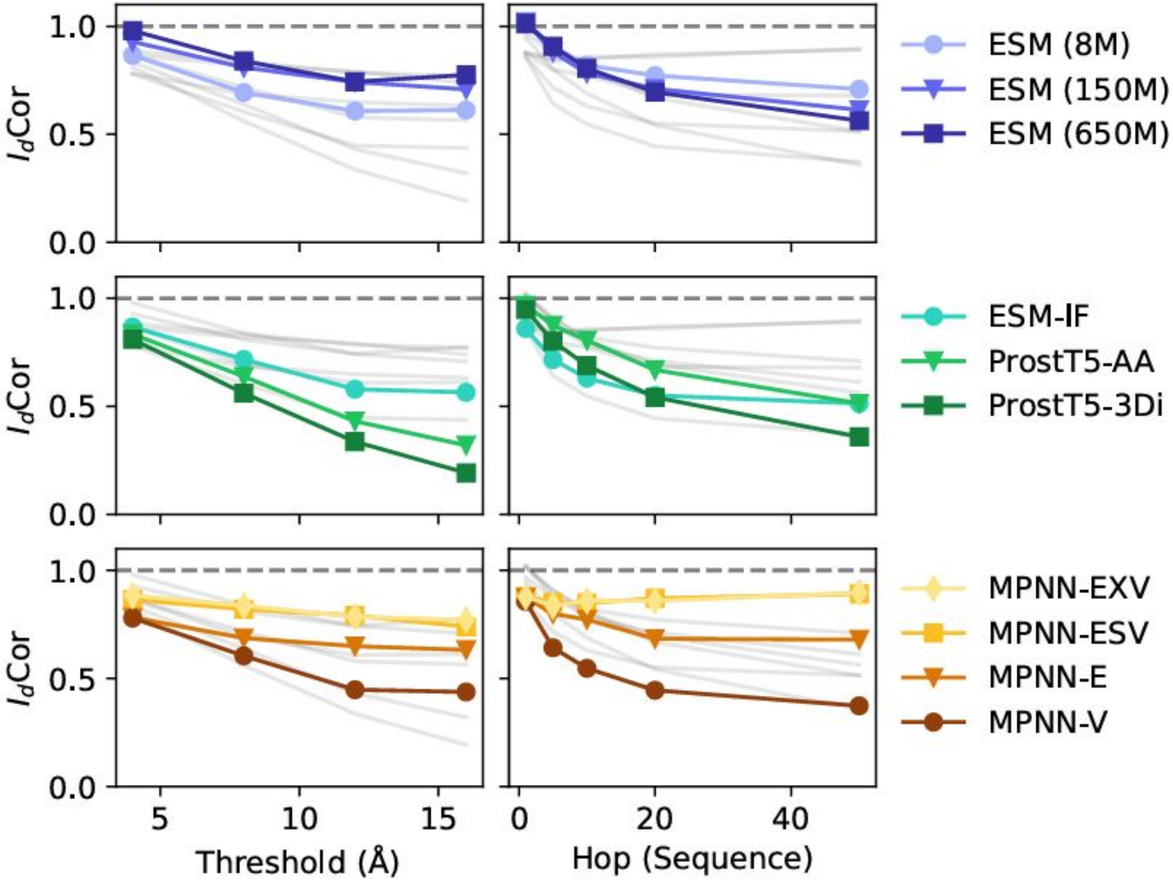
	ESM-2 (8M)	ESM-2 (150M)	ESM-2 (650M)	ProstT5-AA	ProstT5-3Di	ESM-IF	MPNN-V	MPNN-E	MPNN-ESV	MPNN-EXV
ESM-2 (8M)		0.96	0.89	0.96	0.75	0.61	0.48	0.76	0.88	0.99
ESM-2 (150M)	0.96		0.97	0.97	0.78	0.72	0.67	0.88	0.76	0.94
ESM-2 (650M)	0.89	0.97		0.92	0.83	0.80	0.79	0.93	0.72	0.88
ProstT5-AA	0.96	0.97	0.92		0.85	0.82	0.86	0.95	0.73	0.97
ProstT5-3Di	0.75	0.78	0.83	0.85		0.83	0.80	0.97	0.64	0.81
ESM-IF	0.61	0.72	0.80	0.82	0.83		0.64	0.75	0.64	0.71
MPNN-V	0.48	0.67	0.79	0.86	0.80	0.64		0.94	0.60	0.72
MPNN-E	0.76	0.88	0.93	0.95	0.97	0.75	0.94		0.83	0.84
MPNN-ESV	0.88	0.76	0.72	0.73	0.64	0.64	0.60	0.83		0.93
MPNN-EXV	0.99	0.94	0.88	0.97	0.81	0.71	0.72	0.84	0.93	

Results (Long Range / Local Awareness)

1st group:
randomly select 10 residue embedding per protein

2nd group (threshold):
for each residue in 1st group, select embeddings of residues within <threshold> radius of that residue

2nd group (hop):
for each residue in 1st group, select embeddings of residues at <hop> distance in the sequence



Results (Mutant Embeddings)

Embedding space (only of ESM-2 8M) of mutants have been analyzed.

3127 single and double substitution mutants of SH3 domain in Obscurin (PDB ID: 1V1C) is taken from dataset in Tsuboyama et Al.

Dataset also contains DMS scores. DMS (Deep Mutational Scan) score reflects the effect of the mutation on the protein’s general function (and maybe some other properties?).

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.

Results (Mutant Embeddings)

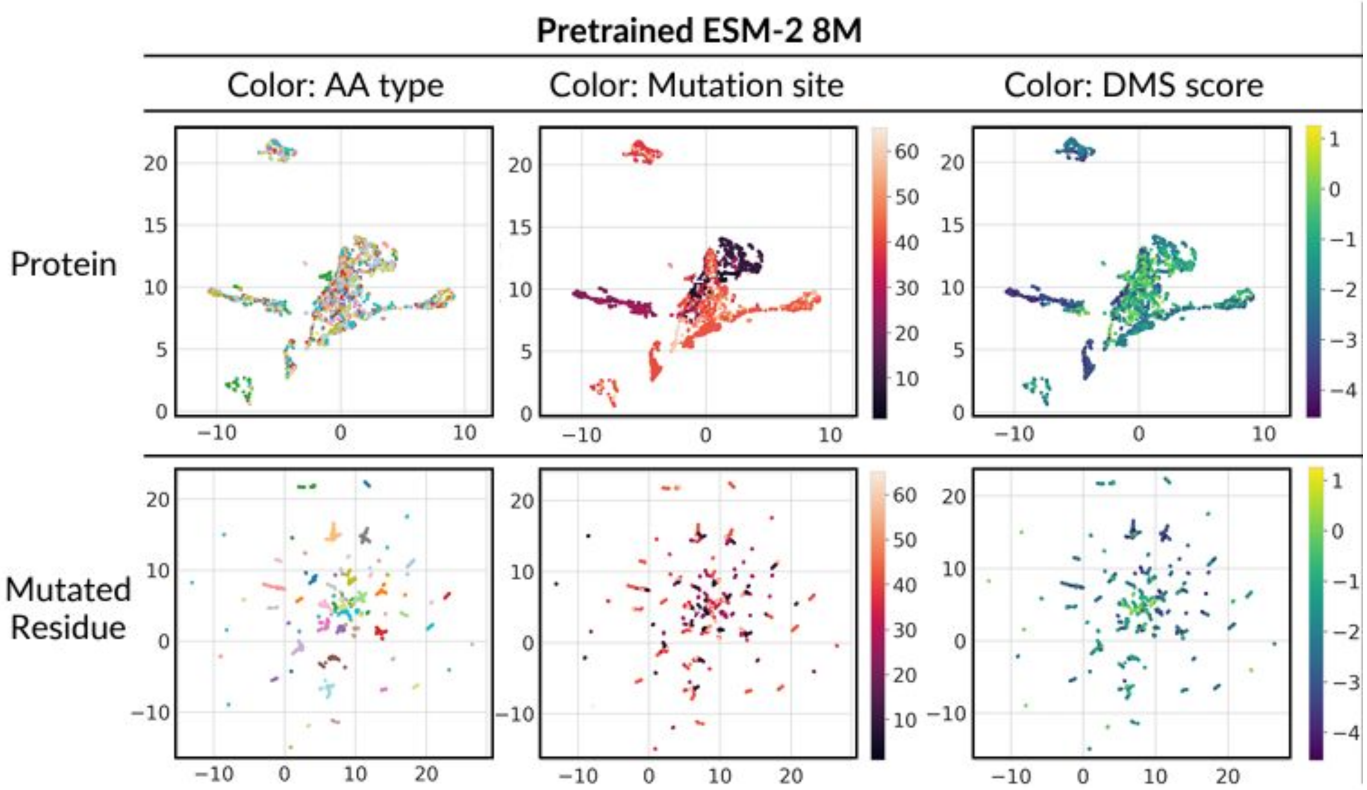
Mutant embeddings (Protein-level) in OBSCN family have generated an Id value of 7.8 averaged over three ESM2 models. This is lower than Id values for the more varied protein dataset.

Mutant residue embeddings generated an Id value of 7.1 (again, average of 3 models). This is significantly lower than the value of 29.7 (this value can be inferred from Table 1)

IdCor between mutant residue embeddings and mutant protein embeddings was 0.77 on average. (No p-value provided)

Results (Mutant Embeddings)

UMAP for mutant embeddings.



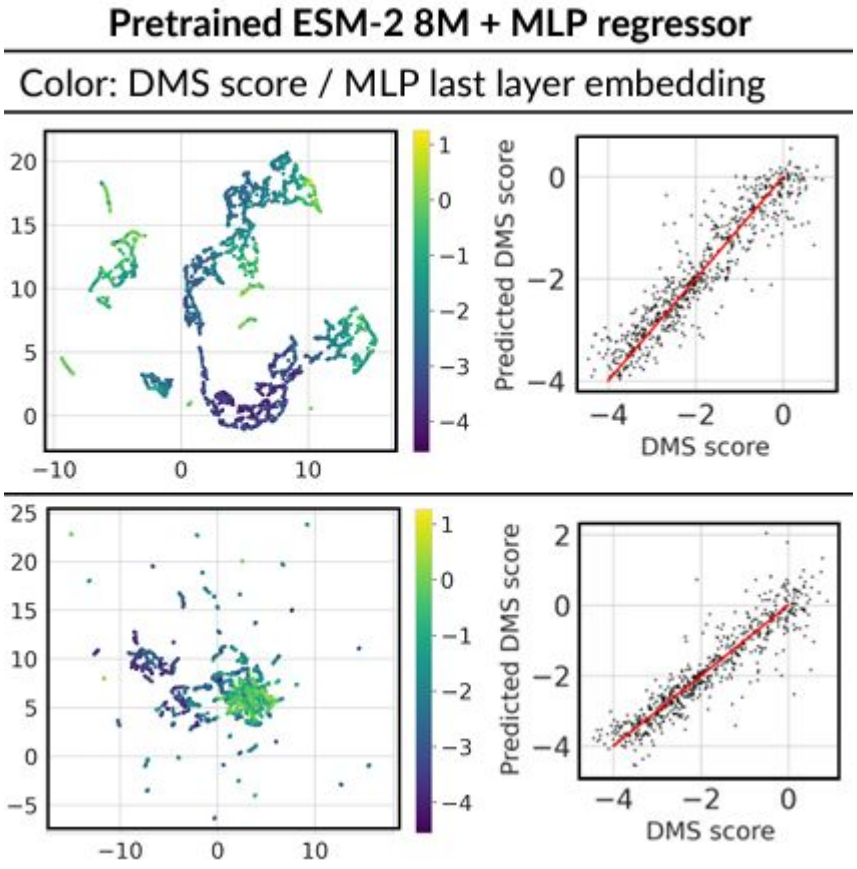
Results (Mutant Embeddings)

A simple MLP regressor to predict DMS scores from embeddings is trained. (16 hidden dim, ReLU activation)

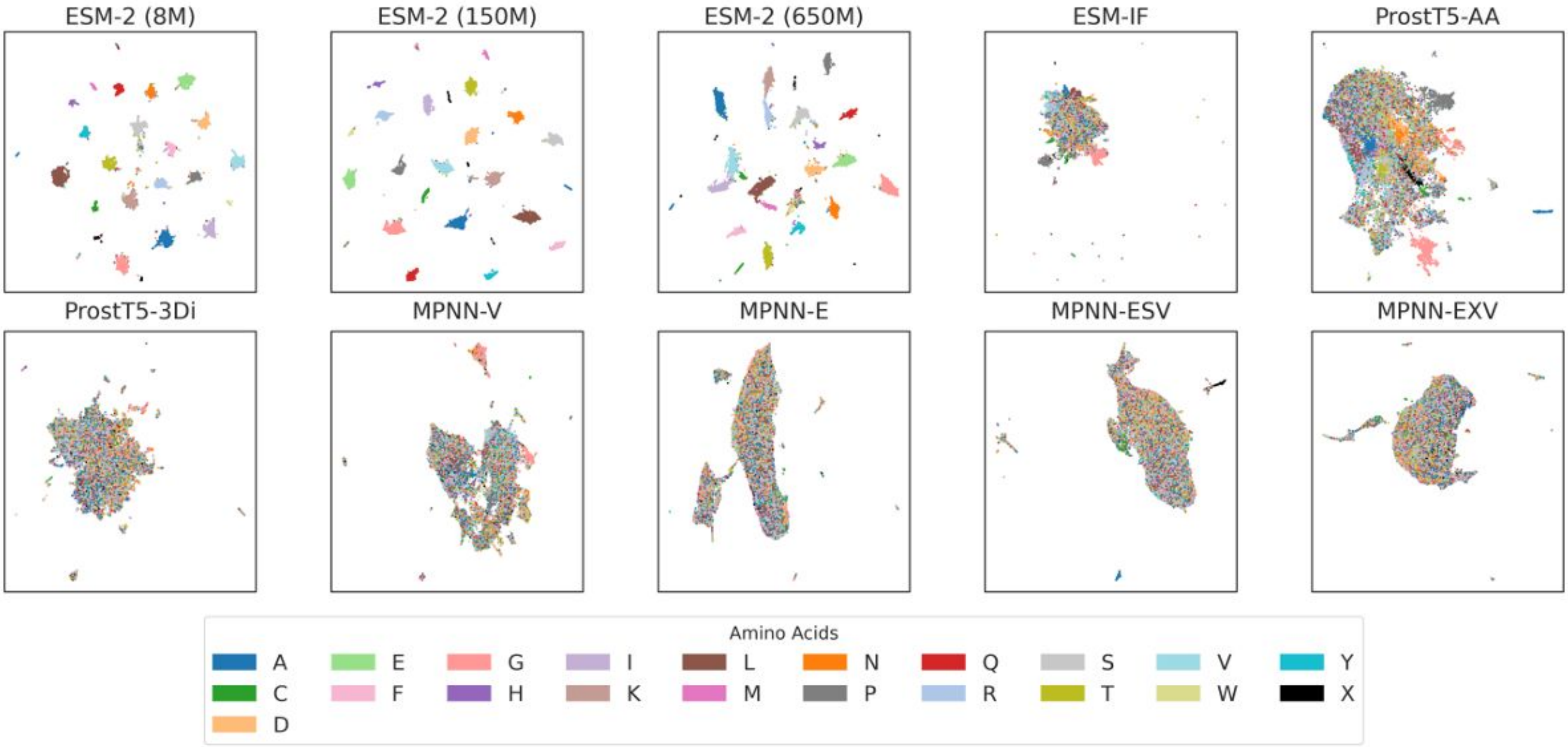
Final layer of MLP have exhibited an Id value of 2.5.

UMAP for the final layer embeddings is shown.

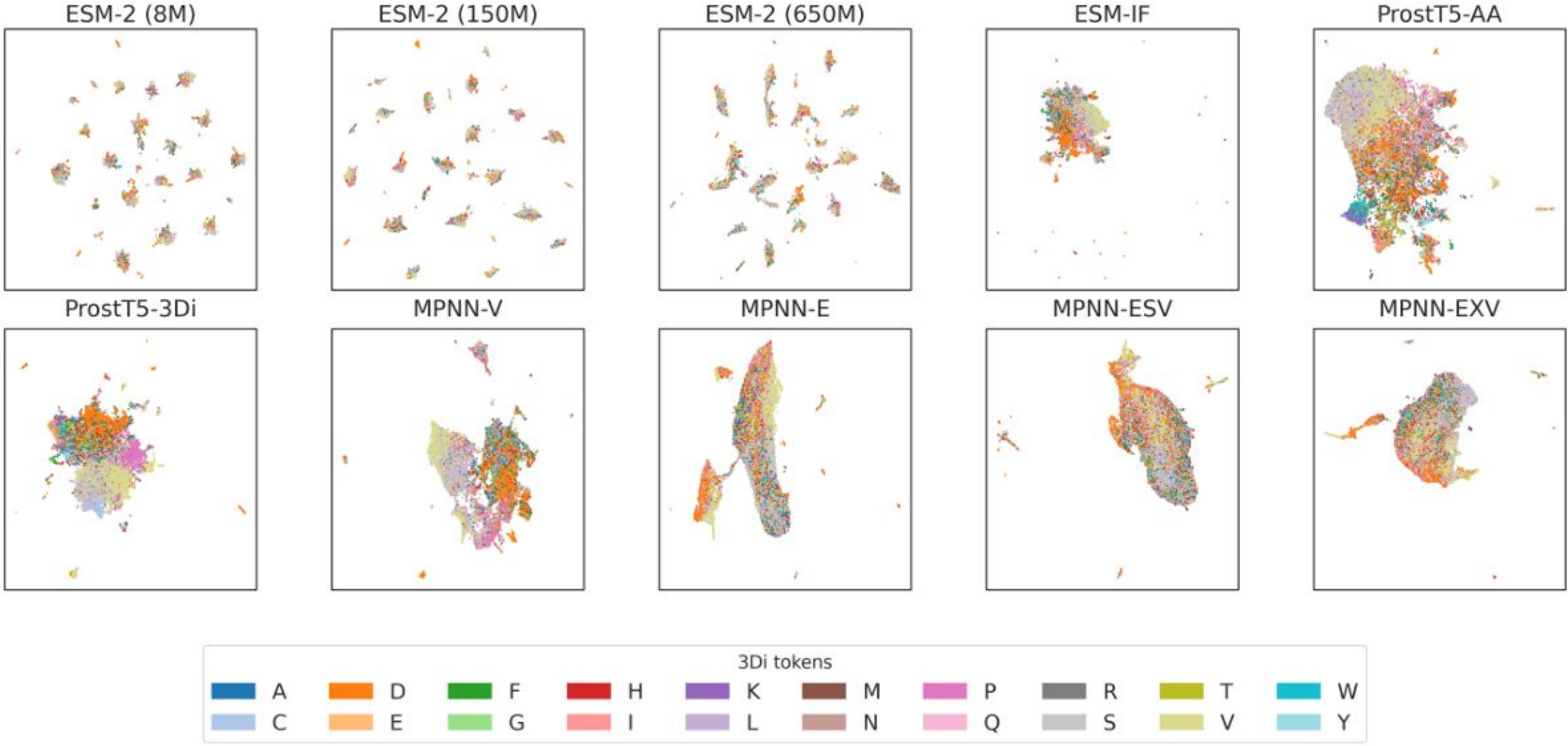
Again, the top row is protein-level and bottom row is residue-level.



Appendix (Clustering of amino acids)



Appendix (Clustering of amino acids)



Thanks for listening