Tokenized and Continuous Compressed Embeddings of Protein Sequence and Structure for Multimodal Generation

Amy X. Lu[1,2], Wilson Yan[1], Vladimir Gligorijevic[2], Pieter Abbeel[1], Kevin K. Yang[3], Nathan Frey[2]

[1]UC Berkeley [2]Prescient Design, Genentech [3]Microsoft Research

BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Prescient Design
A Genentech Accelerator

Abstract

- Existing protein generation and representation models sequence OR structure, with the other modality implicit
- The latent space of protein sequence-to-structure predictors (e.g. ESMFold [1]) offer a **joint representation of structure and sequence**, and can be tamed to obtain explicit decoding to sequence and structure.
  - However, the latent space, similar to large language models, is pathological and contains **massive activations** [2]
- We introduce **CHEAP** (Compressed Hourglass Embedding Adaptations of Proteins) representations, and find that the channel dimension of ESMFold latent spaces can be compressed by up to 256x, while retaining rich structural and functional information
- Our work paves the way towards enabling two-stage latent diffusion generation that has been successful in images [3] for the *simultaneous multimodal generation of sequence and structure*.

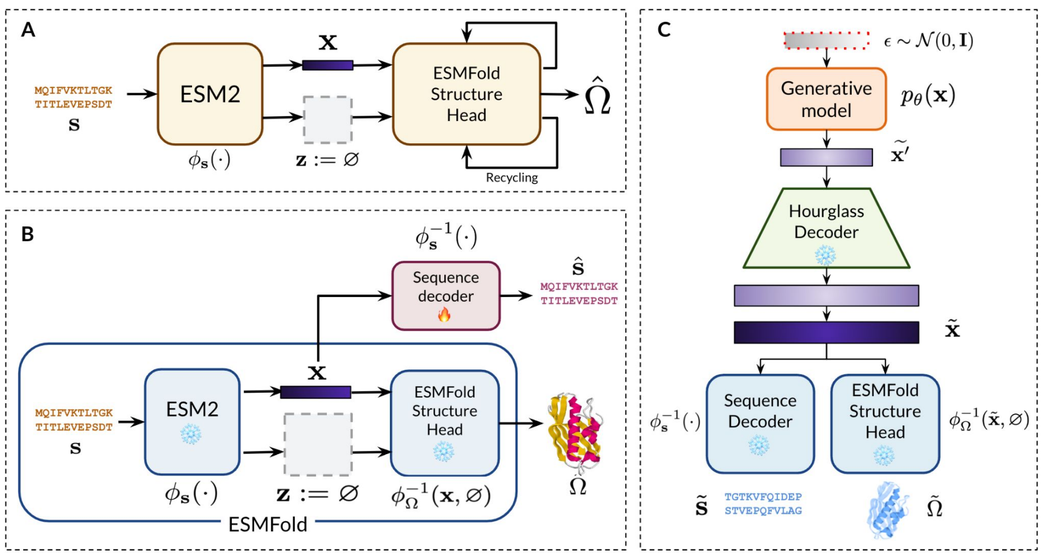


Figure 1. Visualizing how to obtain a joint embedding of sequence and structure from ESMFold.

Massive Activation in Protein Language Models

- ESMFold is derived from a language model, and latent space contains *massive activations:*

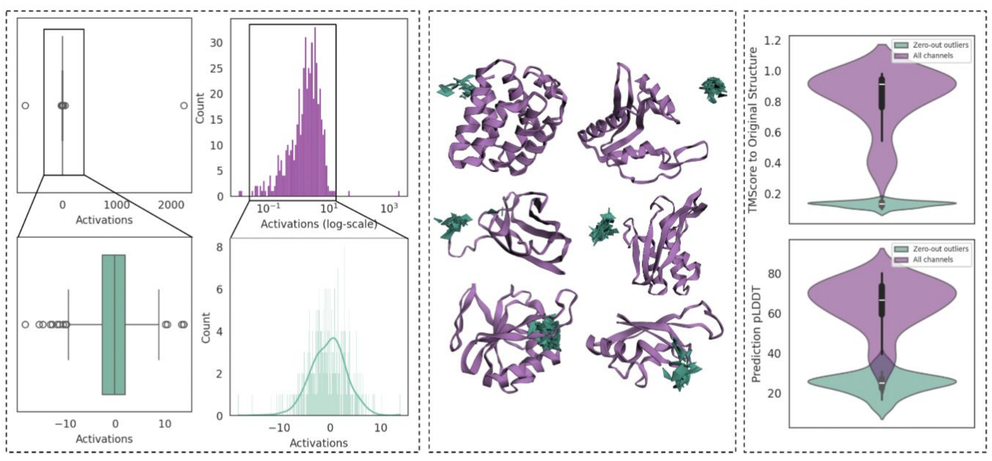


Figure 2. (L) Histogram of per-channel means, where some channels have drastically larger means. Removing three outlier channels with mean absolute values >20 creates a smoother latent space. (M, R) Removing outlier channels cause structure prediction to disintegrate as observed qualitatively and on structure prediction accuracy metrics.

Methods: Defining a Multimodal Latent Space

Define sequence $\mathbf{s} := \{r_i\}_{i=1}^{L}$ 3D positions of atoms in a residue $\mathbf{r} := \{\mathbf{a}_i\}_i^M$, $\mathbf{a} \in \mathbb{R}^3$ and its all-atom structure $\Omega := \{\mathbf{r}_i\}_{j=1}^{L}$. During inference for ESMFold, the pairwise input is initialized to zero. Then:

$$\mathbf{x} = \phi_{\mathbf{s}}(\mathbf{s}) \quad \text{ESM2 Language Model}$$
$$\Omega = \phi_{\Omega}^{-1}(\mathbf{x}, \varnothing) \quad \text{ESMFold Structure Module}$$

We can therefore sample from a learned distribution $\tilde{\mathbf{x}} \sim p_\theta(\mathbf{x}) = p_\theta(s, \Omega)$ and use the above mappings to decode to both structure $\tilde{\Omega} = \phi_{\Omega}^{-1}(\tilde{\mathbf{x}})$ and sequence $\tilde{\mathbf{s}} = \phi_{\mathbf{s}}^{-1}(\tilde{\mathbf{x}})$ (Fig. 1).

Methods: Continuous Compression

- Inspiration: to obtain a VQGAN [4] like autoencoder for latent diffusion models [3] that compresses salient information and improve diffusion efficiency and quality
- Use an *Hourglass Transformer* architecture to create a bottleneck that is shortened length-wise and down-projected channel-wise
- To fix massive activations, normalize by channel statistics:

$$\mathbf{x}' = \frac{\mathbf{x} - \mathbf{x}_{\min}}{\mathbf{x}_{\max} - \mathbf{x}_{\min}} \times \left((c_{\max} - c_{\min}) + c_{\min}\right)$$

Methods: Discrete Compression

- Inspiration: autoregressively modeling tokenized image representations is a successful paradigm in image generation
  - aim to compress salient information into discrete tokens
  - examine different codebook sizes and discretization method
- VQVAE: nearest neighbor search for a codebook vector:

$$L_{\text{VQ}} = \log p(\mathbf{x}|h_q(\mathbf{z})) + ||\text{sg}[h_e(\mathbf{x})] - \mathbf{z}||_2^2 + \beta||h_e(\mathbf{x}) - \text{sg}[\mathbf{z}]||_2^2$$

- FSQ: directly bin into a discrete number

$$\mathbf{z} = h_e(\mathbf{x}), \mathbf{z} \in \mathbb{R}^d \quad \text{Continuous encoder representation} \quad (3)$$
$$\hat{\mathbf{z}} = \tanh(\mathbf{z}) \quad \text{Bound to } [-1, 1] \quad (4)$$
$$\hat{\mathbf{z}} = \text{round}(\lfloor (L/2) \rfloor \cdot \hat{\mathbf{z}}) \quad \text{Discretize to } L \text{ bins and round to nearest integer}$$

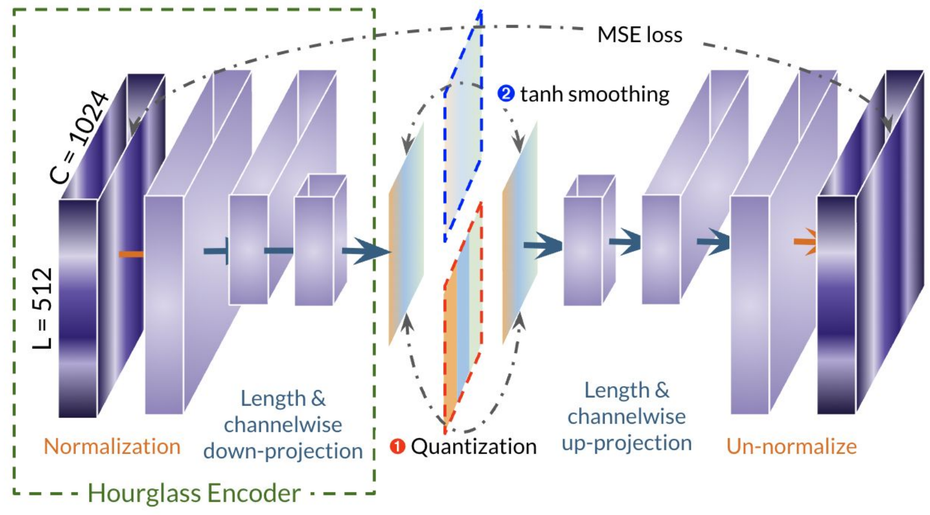


Figure 3. Obtaining a compressed embedding that might continuous (with tanh bounding) or discrete

Results

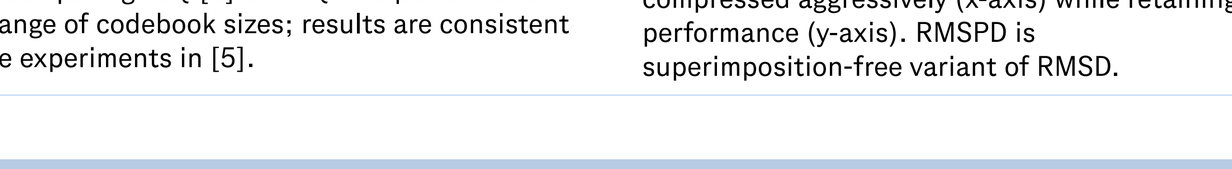
Dim.	4	8	16	32	64	128	256	512	1024
									
TMScore	0.789	0.920	0.968	0.995	0.991	0.999	1.000	0.997	1.000

Figure 4. Visualization and TMScore for embeddings at different channel compression levels. A shortening factor of 2 was applied to all structures (original dimension: 512 x 1024). Despite aggressive compression at the bottleneck, performance can still be retained to a high degree.

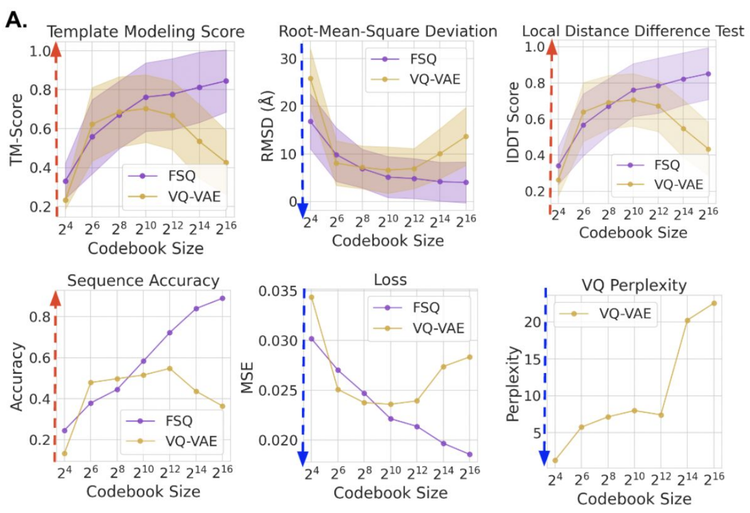


Figure 6. Comparing FSQ [5] and VQ-VAE performance across a range of codebook sizes; results are consistent with image experiments in [5].

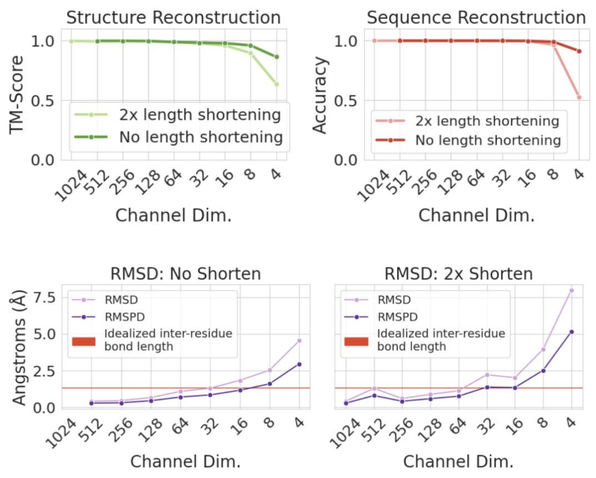


Figure 5. Continuous embeddings can be compressed aggressively (x-axis) while retaining performance (y-axis). RMSPD is superimposition-free variant of RMSD.

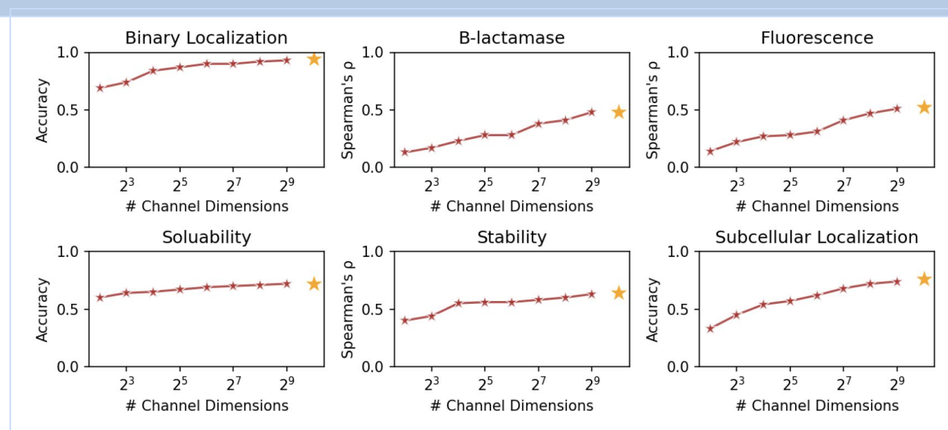


Figure 7. On PEER [6] benchmark tasks, function performance degrades more drastically as dimensions are reduced, as compared to structure and sequence tasks. Some tasks are more affected by compression than others.

References

[1] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, et al., Evolutionary-scale prediction of atomic- level protein structure with a language model, 2023.

[2] M. Sun, X. Chen, J. Z. Kolter, Z. Liu, Massive activations in large language models, 2024.

[3] P. Esser, R. Rombach, B. Ommer, Taming transformers for high-resolution image synthesis, 2021.

[4] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, 2022.

[5] Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: Vq-vae made simple, 2023.

[6] Minghao Xu, Zuobai Zhang, Jiarui Lu, Zhaocheng Zhu, Yangtian Zhang, Ma Chang, Runcheng Liu, and Jian Tang. Peer: a comprehensive and multi-task benchmark for protein sequence understanding, 2022.