The Unreasonable Compressibility of Protein Folding Models

Paper: bit.ly/cheap-proteins
GitHub: github.com/amyxlu/cheap-proteins

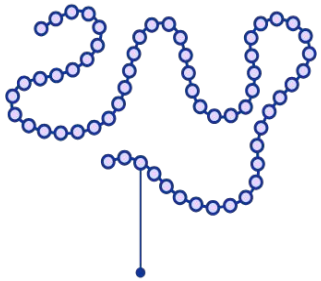
Paper GitHub

background:
protein folding 101

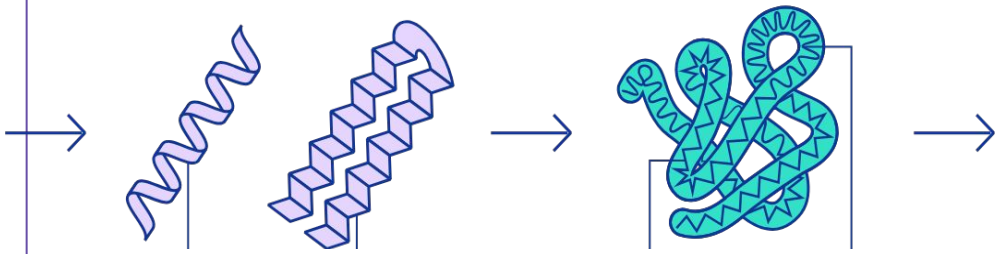
What is a protein?

sequence

Every protein is made up of a sequence of amino acids bonded together

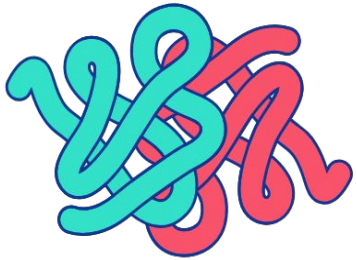


Amino acids

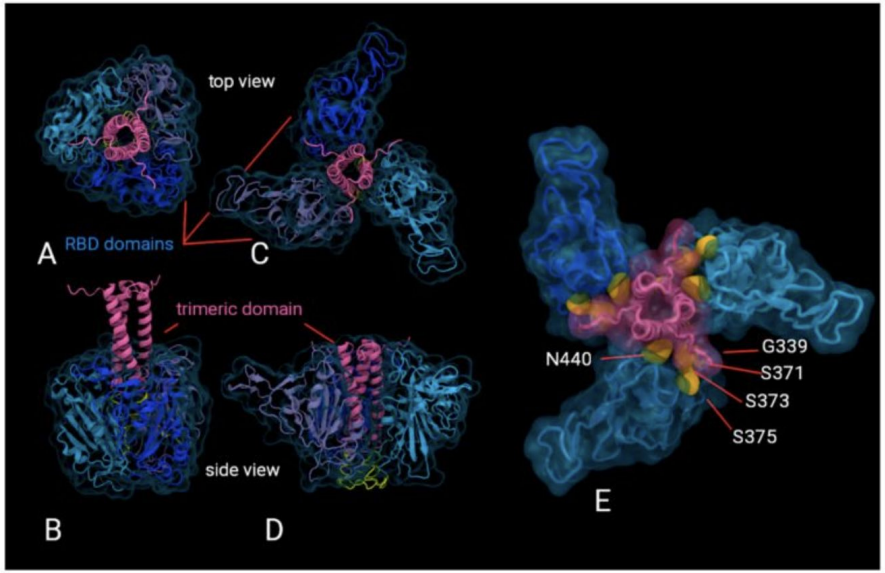


structure

Proteins can interact with other proteins, performing functions such as signalling and transcribing DNA

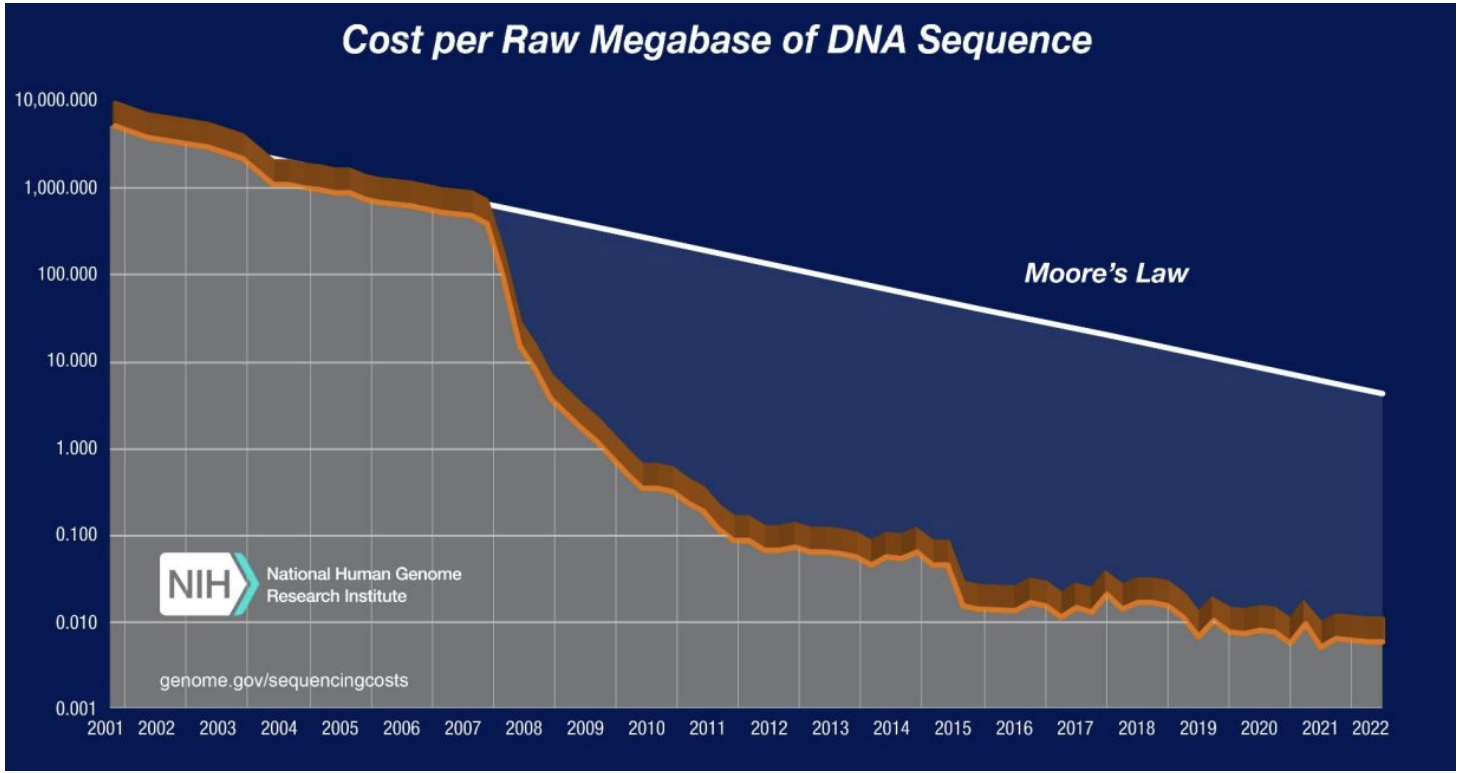


protein structure 🤝 drug discovery

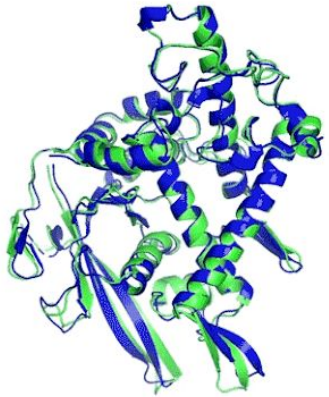


Example: AlphaFold-predicted structures help us hypothesize how sequence-level mutations in the SARS-CoV2 Omicron variant impacts its mechanism.
(Source: van Vuren et al., 2022)

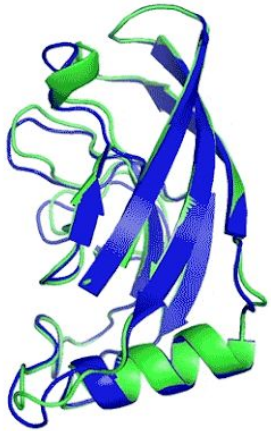
protein structure prediction 🤝 drug discovery



Enter: protein folding models



T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)



T1049 / 6y4f
93.3 GDT
(adhesin tip)

Experimental result
Computational prediction

AlphaFold tutorial:
https://bit.ly/amyxlu-af2

Example: AI for binder design

Demo from Google DeepMind’s AlphaProteo, released September 5 (today!)

CHEAP

(Compressed Hourglass Embedding Adaptations of Proteins)

Amy X. Lu, Wilson Yan, Kevin K. Yang, Vladimir Gligorijevic, Kyunghyun Cho, Pieter Abbeel, Richard Bonneau, Nathan Frey

Full paper: bit.ly/cheap-proteins

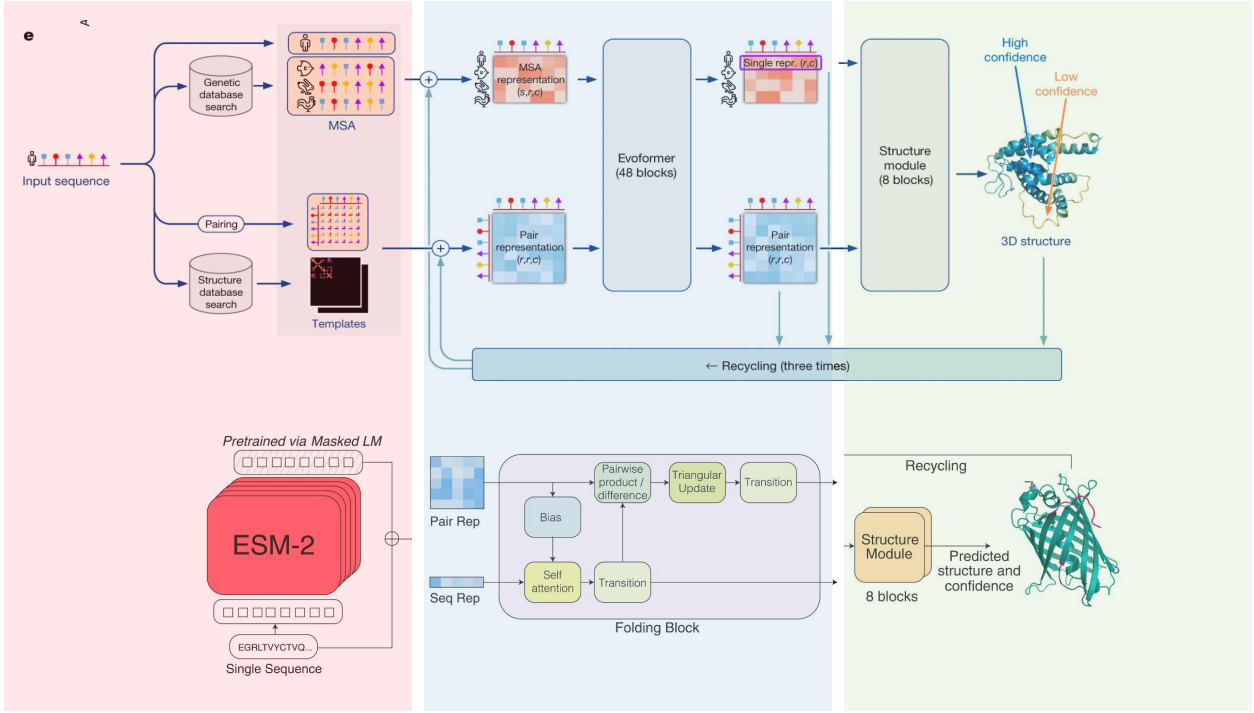
Protein language models can learn structure…

AlphaFold2:

Uses an explicit retrieval step

ESMFold:

Replaces retrieval step with a **language model**



harness additional sequence-based priors

learn structural features from sequence latents

generate structures

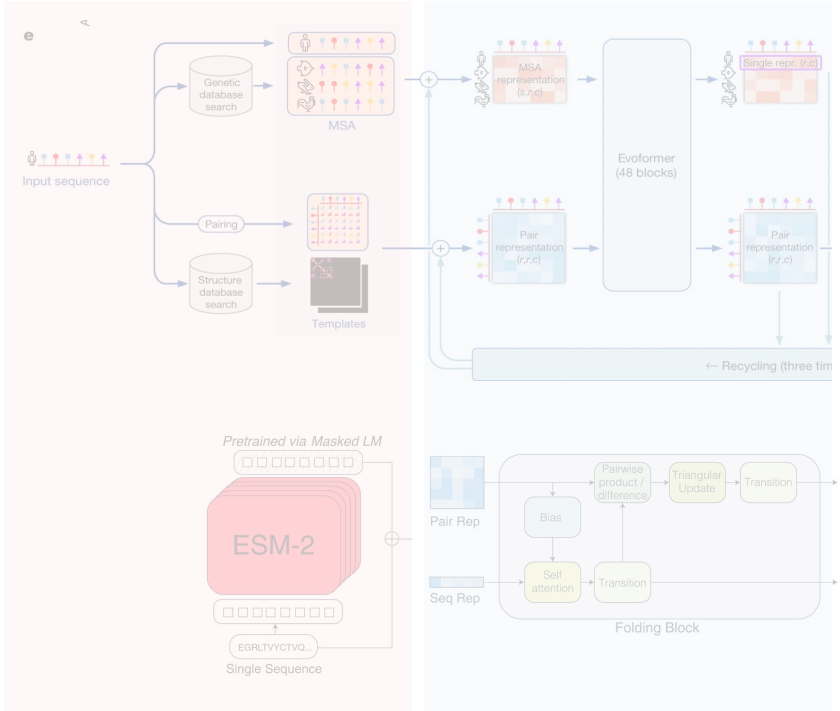
…from large prediction models to foundation models?

AlphaFold2:

Uses an explicit retrieval step

ESMFold:

Replaces retrieval step with a **language model**



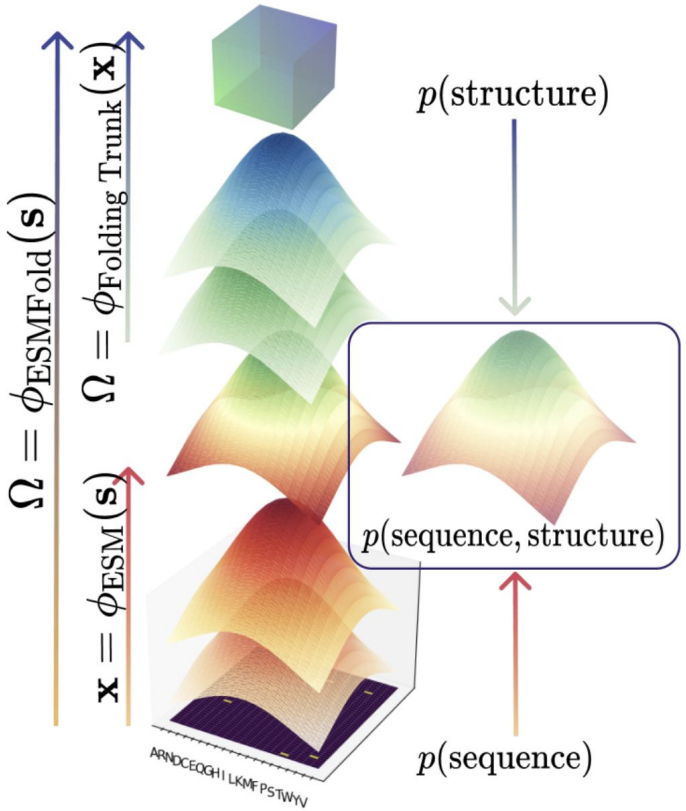
harness additional sequence-based priors

learn structural features from sequence latents

Motivation:

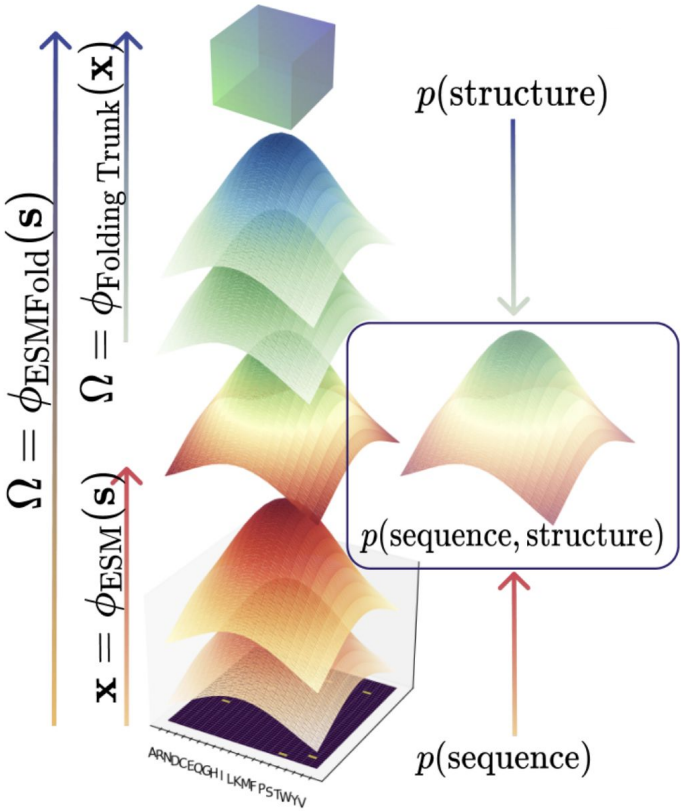
Can we reuse parts of large protein folding models to decrease training costs for other biological tasks?

…from large prediction models to foundation models?



Let's extract the inner layer representations as a multimodal embedding of sequence and structure…

Characterizing a multimodal latent space



Can we extract the latent space of sequence-to-structure models as a multimodal embedding?

→ multimodal **generation**

One-shot generation of drug binders!

→ multimodal **search**

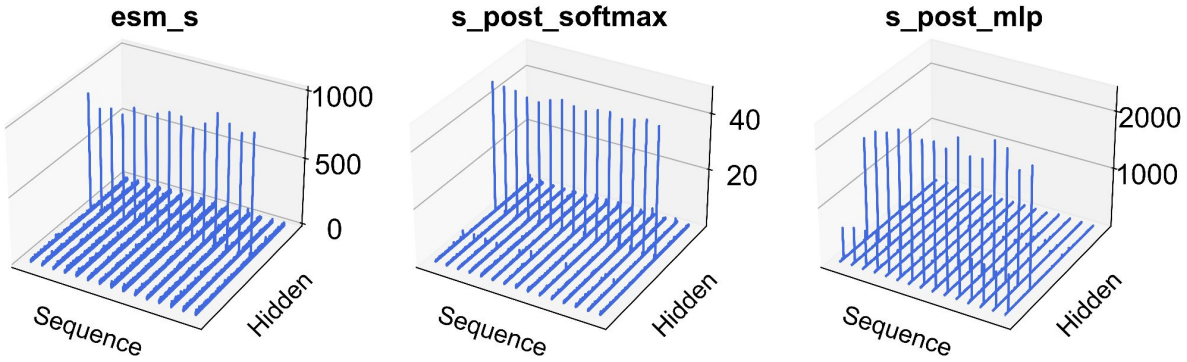
Find new gene editing enzymes from obscure bacteria!

→ ...

...turns out that it's not so straightforward :(

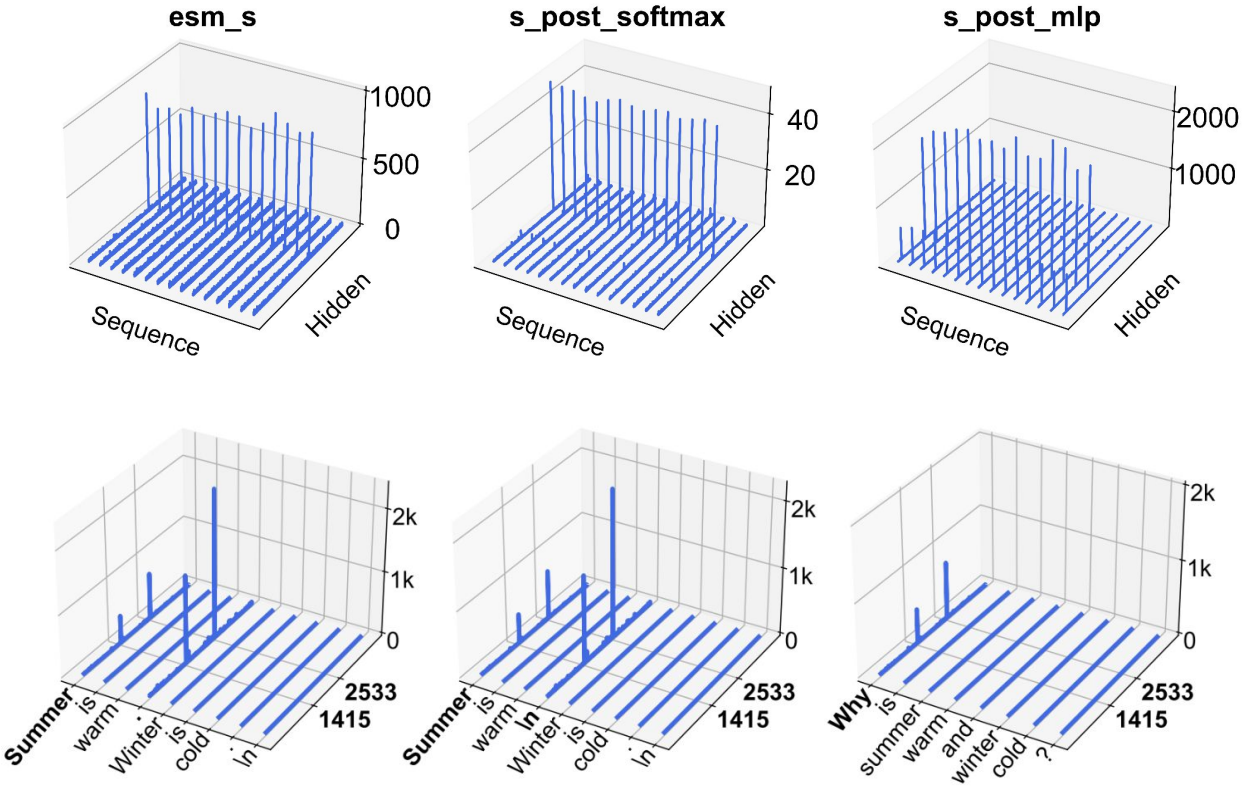
some mechanistic interpretability findings:

Protein language models have massive activations



ESMFold

~~Protein~~ language models have massive activations

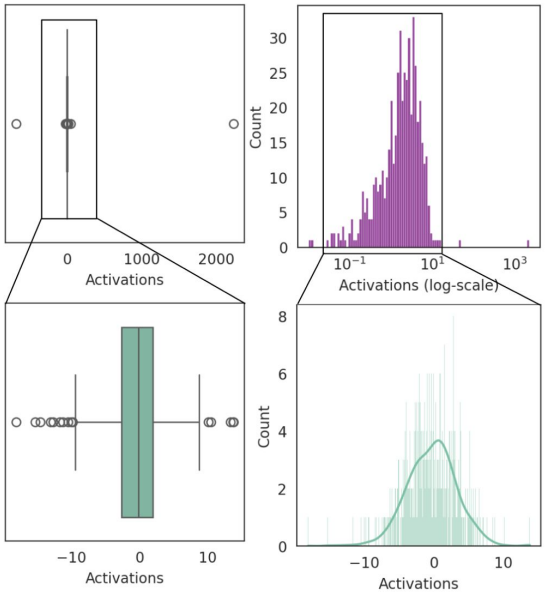


ESMFold

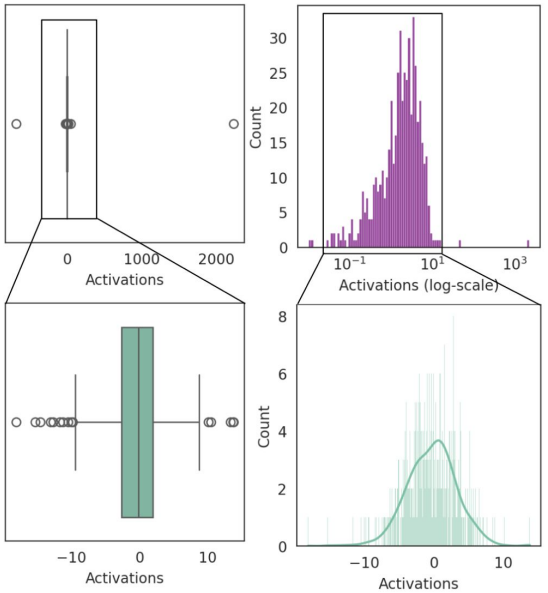
LLaMA2-7B

Image source: Sun et al.
https://arxiv.org/pdf/2402.17762

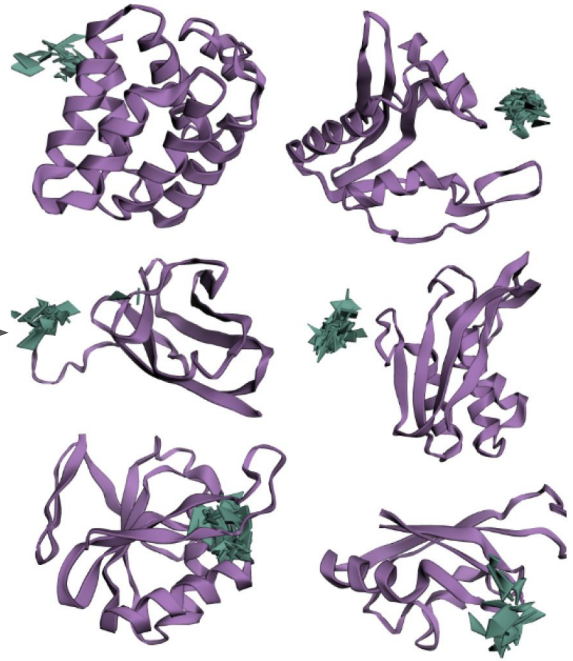
Protein language models have massive activations, and these wacky features are really important



Protein language models have massive activations, and these wacky features are really important

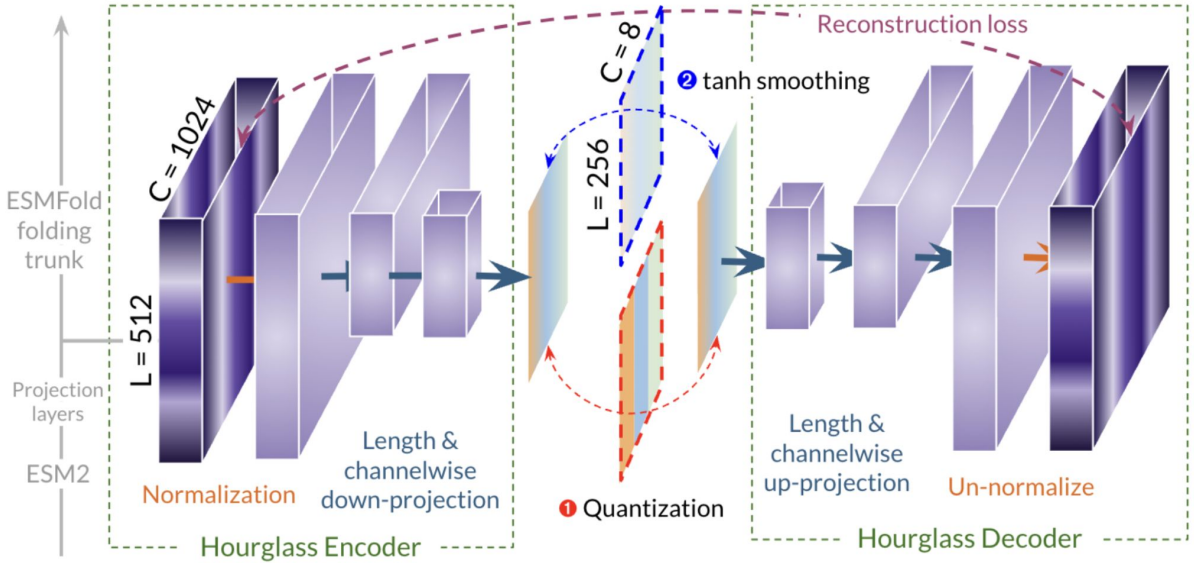


By removing the outliers, the **original performance entirely deteriorates**

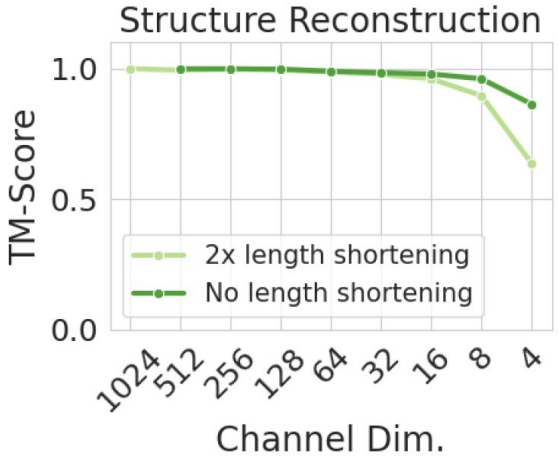


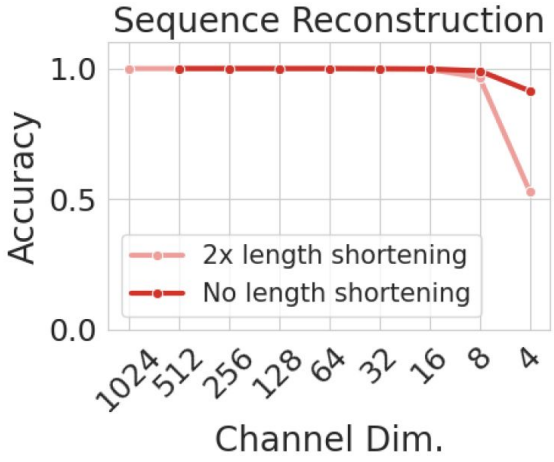
Since we don’t need all the features, can we compress the latent space?

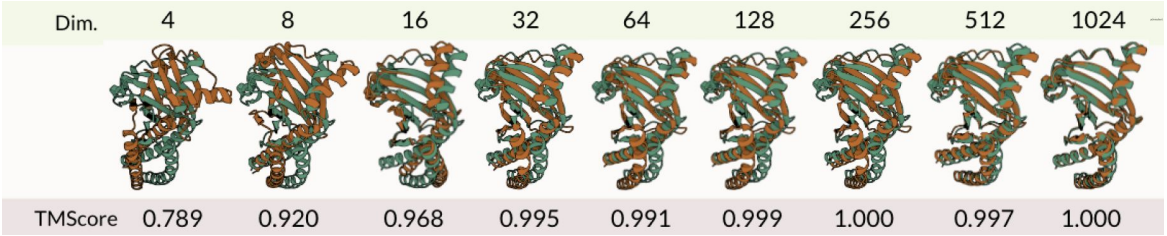
→ Train an autoencoder to squeeze the latent space from 1024 features to *[insert dimension here]*



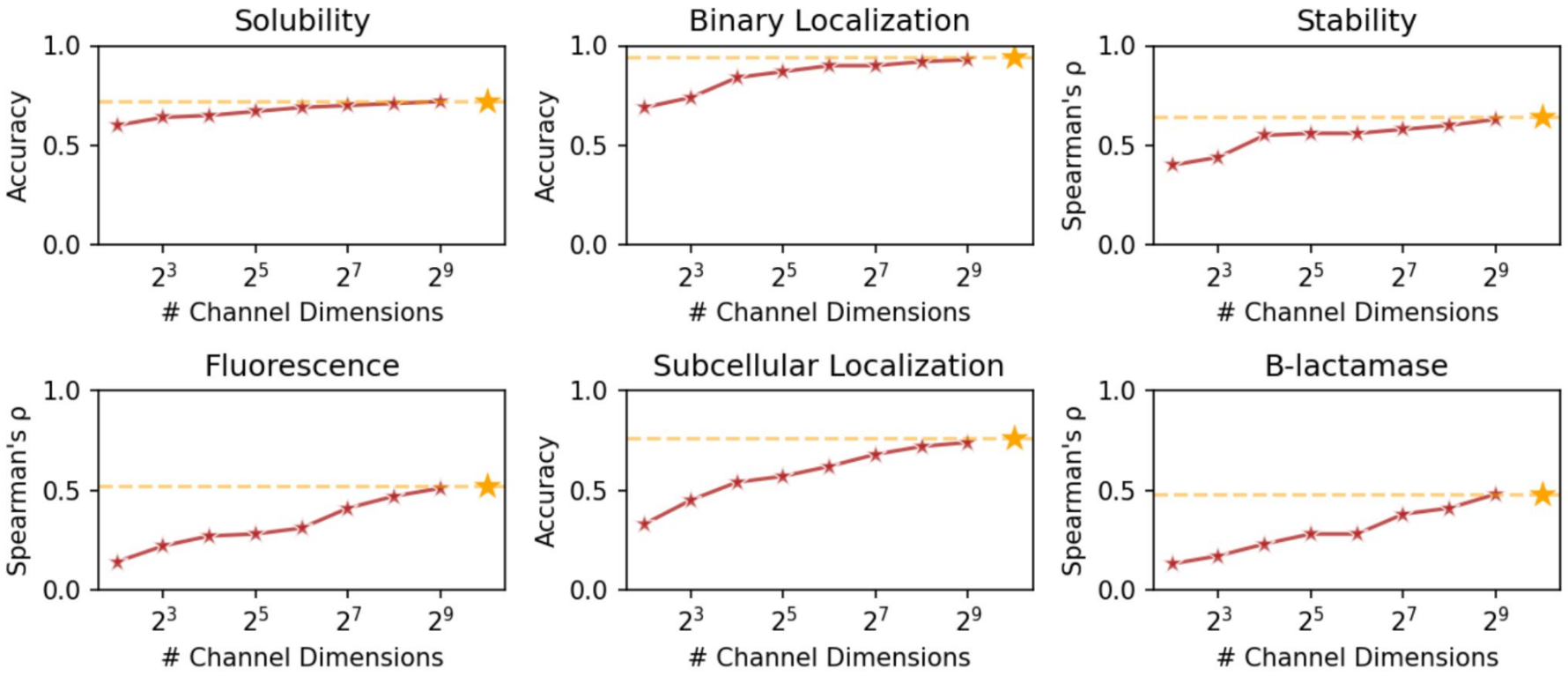
yes, we can compress up to 128x 😲







…but not quite the case for function prediction 🤔

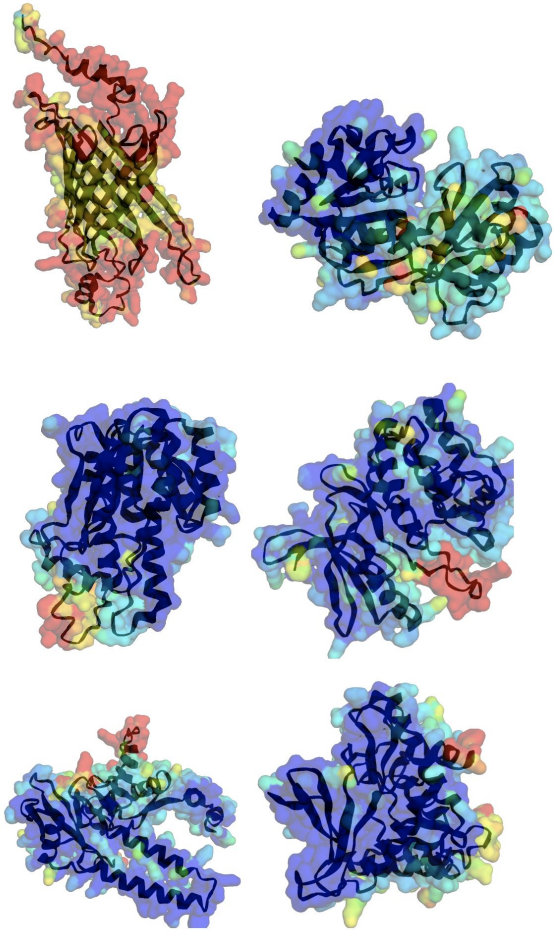


Takeaways

- Background: protein folding models are getting really good
  - Good for drug discovery b/c structure resolution is expensive and sequencing is cheap
- Motivation: repurposing them as foundation models?
  - Extract the latent space for downstream {generation, search, …}
- Findings: latent space is disorganized with massive activations
  - (Like other LMs)
- By compressing the latent space, we find that ***many*** channels are extraneous for structure prediction
  - We can perhaps build a much more compact “foundation model” from these protein folding models 👀

Plus some other findings: see full paper

bit.ly/cheap-proteins



Preview...

Compressing this latent space offers a lower-resolution representation, akin to latent diffusion models (aka Stable Diffusion)

Makes diffusion learning much easier! Allows us to do function and organism conditioned generation – paper to come 👀