DeepMind

Mapping ML methods to protein problems

John Jumper

CASP15

How ML has changed as a field

- Diffusion is the generative model of choice for anything but text
- Extremely large text and image models based on the “scaling hypothesis”
- Language models now incorporate “retrieval”
- Fine-tuning and prompts are keys to deployed models
- Transformer is the main architecture for large models

Outline

- Generative models and diffusion
- Protein language models and the scaling hypothesis
- Next problems

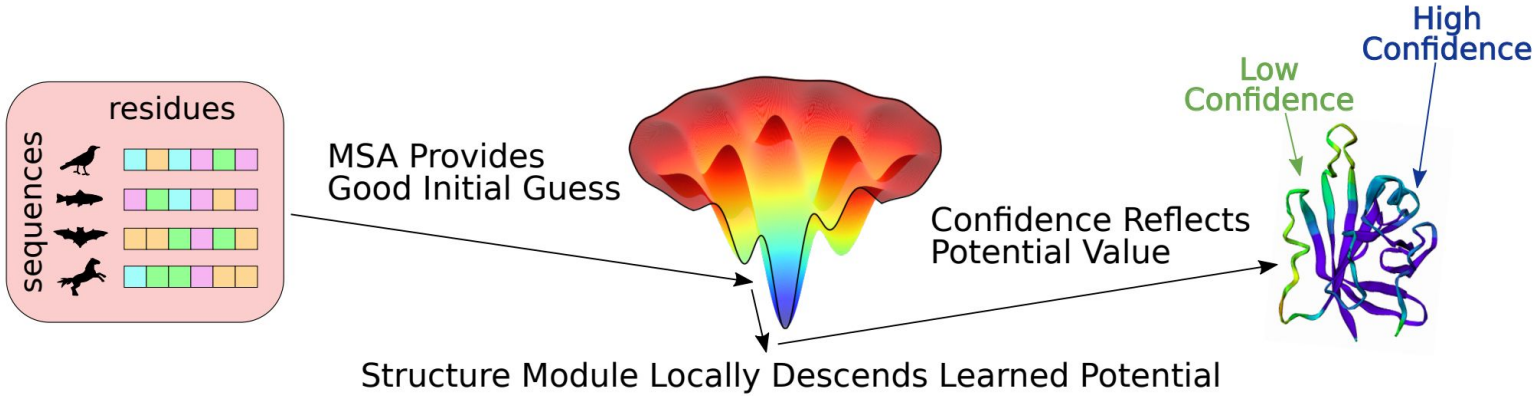
Multi-state modelling

Protein structure modelling from sequence is under-specified

Current state of the art is MSA subsampling then standard structure prediction

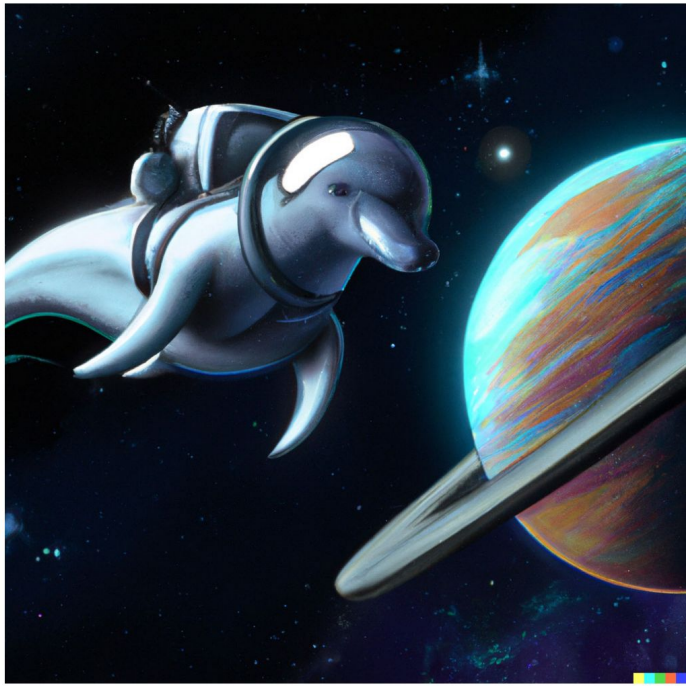
- Del Alamo et al., 2022
- Wayment-Steele et al., 2022

Probable explanation is given in Roney and Ovchinnikov -- smaller MSAs descend into a random basin



Conditional generative modelling for images

Standard answer in ML for how to get many answers from a single input

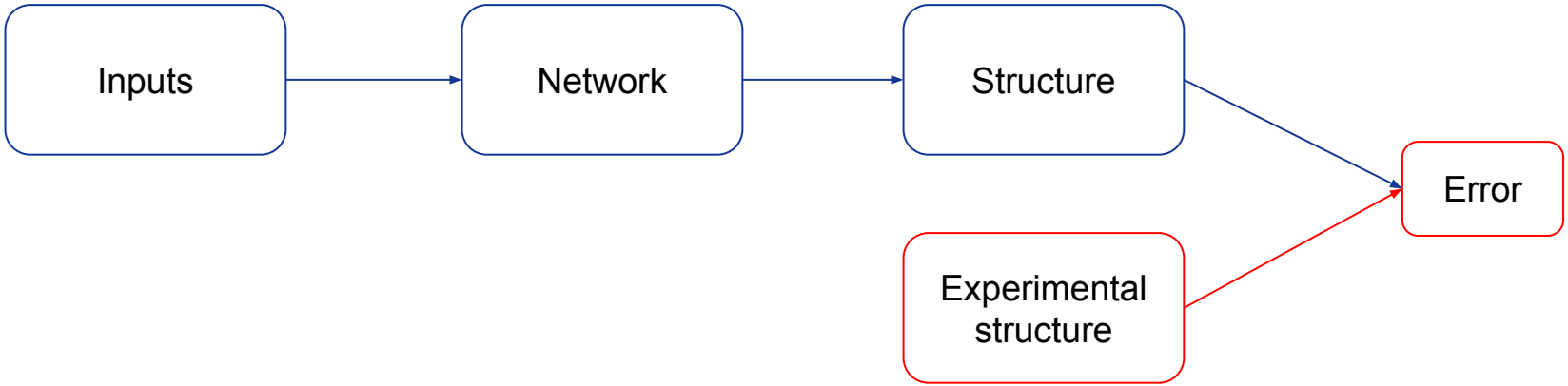


a dolphin in an astronaut suit on saturn, artstation

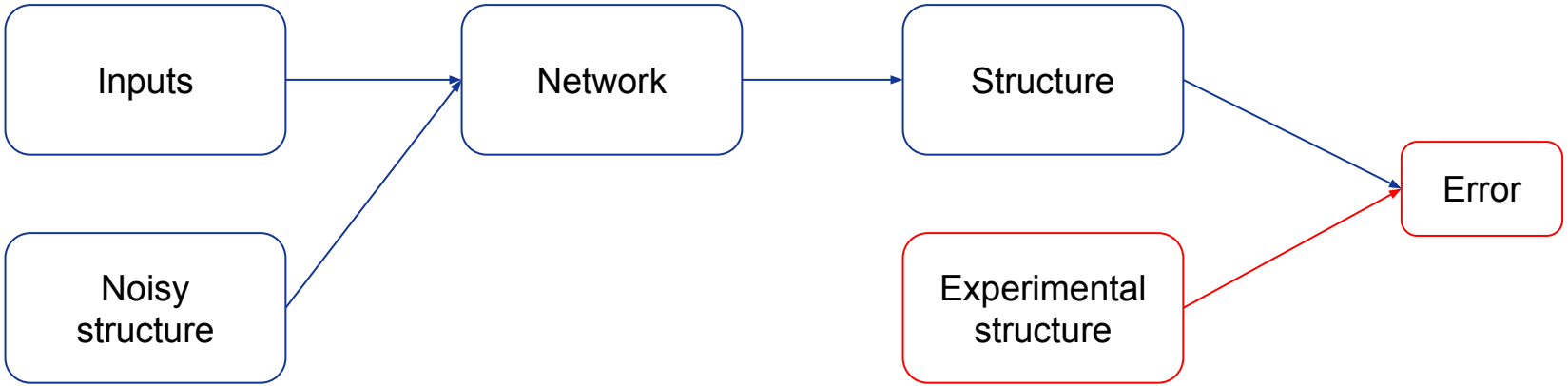
Ramesh et al., 2022

Protein structure networks, a simplified view

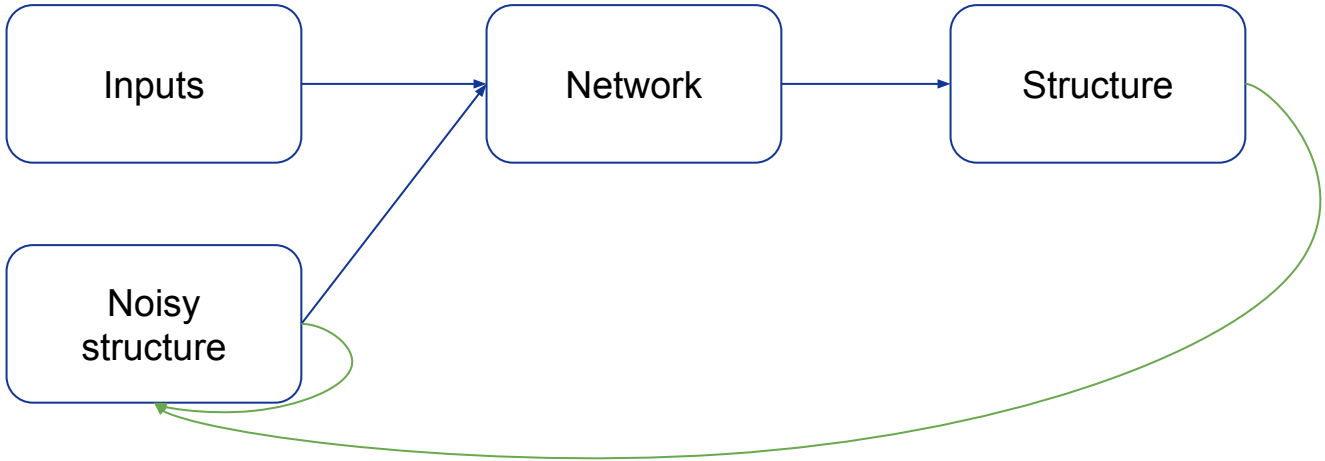
Protein structure training, a simplified view



Diffusion training, a simplified view



Diffusion sampling

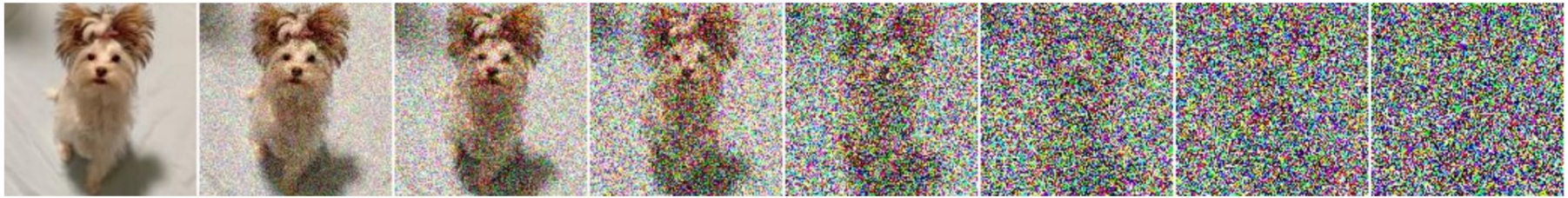


A lot like recycling but with noise

Diffusion sampling for images

Forward SDE (data → noise)

$\mathbf{x}(0)$ — $\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}$ → $\mathbf{x}(T)$



score function

$\mathbf{x}(0)$ ← $\mathrm{d}\mathbf{x} = \left[\mathbf{f}(\mathbf{x}, t) - g^2(t)\nabla_{\mathbf{x}} \log p_t(\mathbf{x})\right] \mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}}$ — $\mathbf{x}(T)$

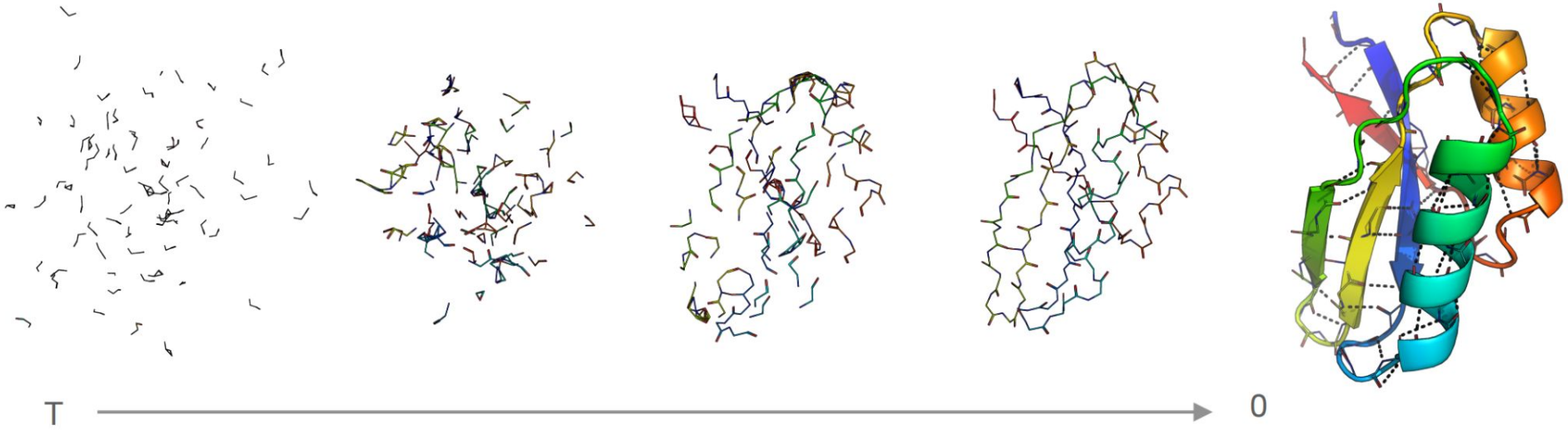
Reverse SDE (noise → data)

Song et al, 2021

Diffusion for proteins

Several recent papers, among others

- Anand and Achim, 2022 (equivariant diffusion)
- Watson et al, 2022 (RFdiffusion)
- Ingraham et al, 2022 (Chroma diffusion)



Language modelling in ML

Two big approaches in language

- Masked language modelling / BERT for understanding text
- Autoregressive language modelling for generating text

Key point of each is using simple, game-like tasks on large amounts of text

Masked language modelling

Given a very simple task to networks: fill in the blank

Example sentences:

- The car was ____ on the road and hit a speed bump
- They took a picture of the Eiffel tower when vacationing in _____.
- Since Alice wasn't hungry she took the ____ with her.

Solving this task *really well* requires knowing a lot about both text and the world

Very large networks are trained on huge amounts of text for this

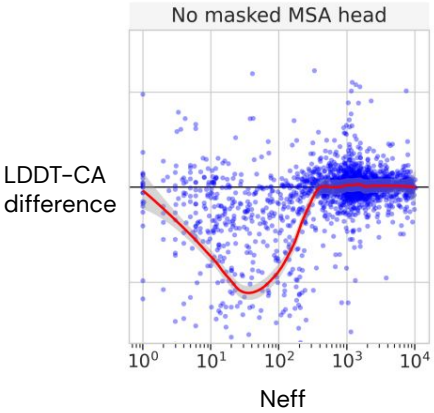
Masked language models as covariation

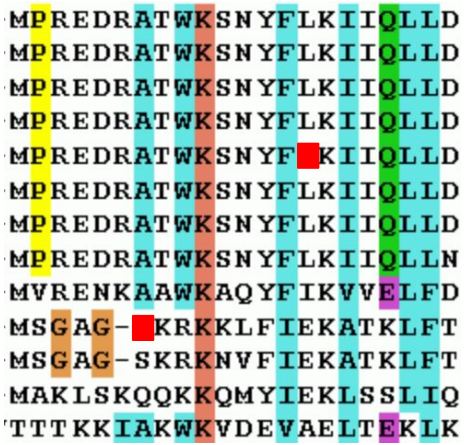
Masked language models and coupling analysis are highly analogous -- predict one part of the sequence from the rest

In theory, networks can pick up higher-order correlations on MSAs

Two approaches

- Classic pretraining for MSA Transformer (Rao et al, 2021)
- Co-training as another objective for AlphaFold





https://en.wikipedia.org/wiki/Multiple_sequence_alignment#/media/File:RPLP0_90_ClustalW_aln.gif

Single sequence masked language modelling

Why do we need the MSA at all?

- These sequences are in the training set anyway

Expect to need much larger models for MSA-free masked languaging modelling

- Memorizing protein families is a part of the task, just like memorizing place names
- Learning structure must eventually be useful
- Ultimately is a model of “this sequence could appear in UniProt / Mgnify / etc”

Several projects to train very large models including ESM-2, Prot-T5-XL, AminoBERT, OmegaFold, and others

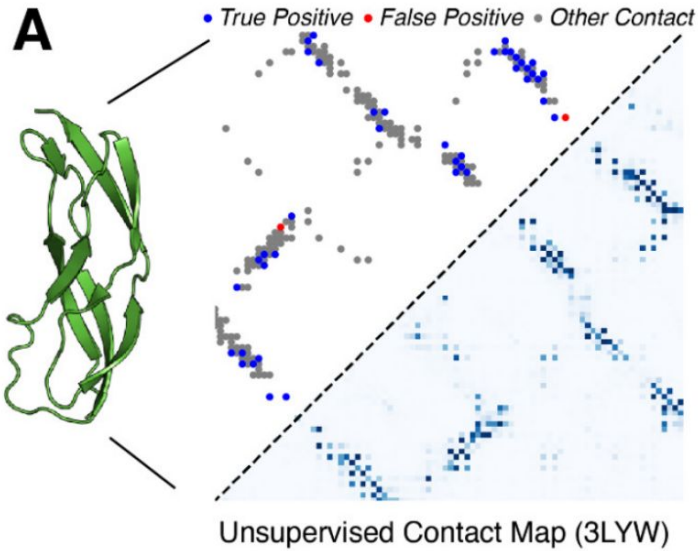
Natural applications to predicting variant effects

Can protein language models learn structure?

Look at the attention maps -- do they look like contacts?

Formalizing this procedure, the answer is definitely “yes” in favorable cases

- As expected since coupling analysis works

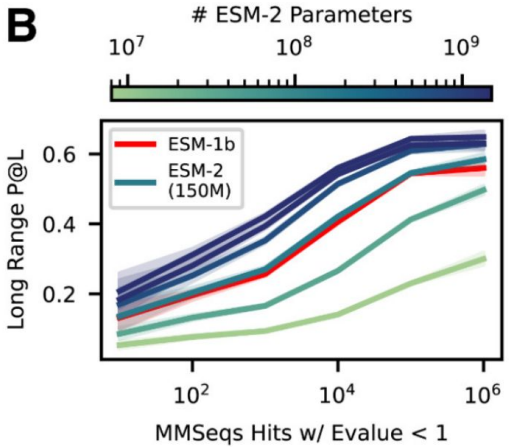


ESM-2 prediction

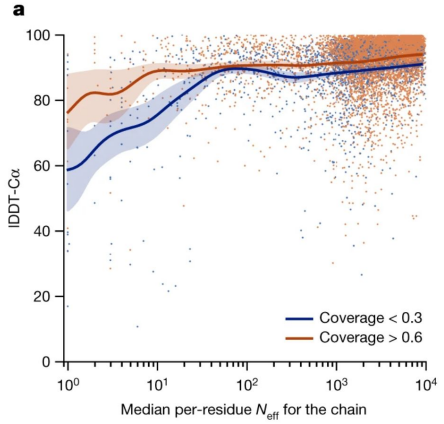
Is this evolutionary structure prediction or physical?

"Single sequence structure prediction" is ill-defined

- AlphaFold the software takes a single sequence (just happens to start with a Jackhmmer layer)
- Real question is how these networks behave with MSA depth
- While fair comparisons are hard, protein LMs look *more* MSA depth sensitive than MSA-based networks
- Very likely though that protein LMs could pick up unusual conservation patterns that MSAs miss
- Other reasons like speed and ease-of-use can come into play as well



ESM-2 unsupervised accuracy

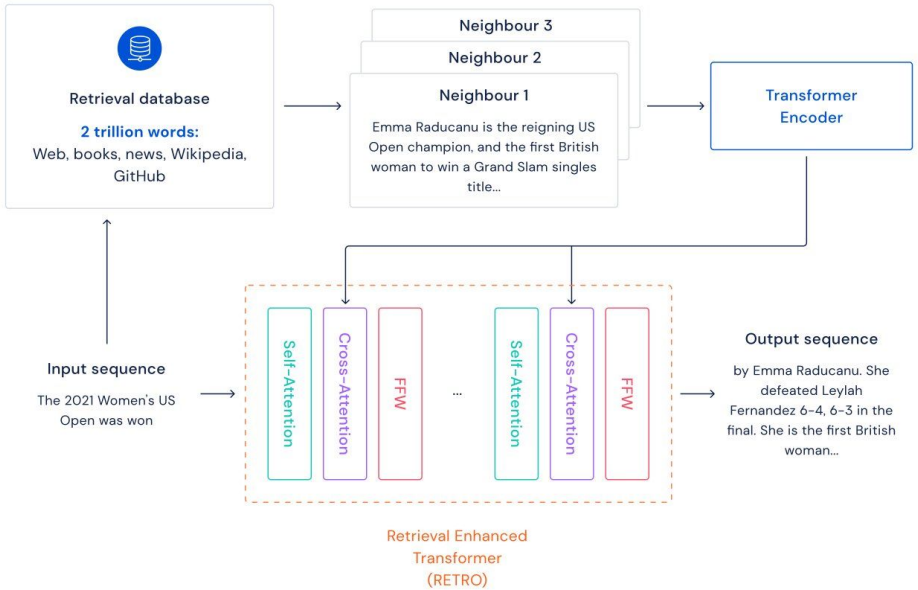


AF2 accuracy vs MSA depth

An analogy to MSAs in natural language modelling

Anecdotal evidence is that even large language models have trouble with rare topics

Humans are pretty similar, so we look things up or search the internet -- let's do that for the network



Basically, this is language modelling *introducing* MSAs to their networks

https://www.deepmind.com/blog/improving-language-models-by-retrieving-from-trillions-of-tokens

Autoregressive language modelling

Very similar to masked language modelling, except the model always predicts the last word

Examples:

- Alice was very sleepy. It was time to go to _____
- They were out of milk. It was time to go to _____

Some of the most famous AI models right now are autoregressive language models, e.g. ChatGPT

To use them, you ask the network to pick the next word over and over again which generates text

Error is measured in *perplexity* which is rough the 2^bits of entropy per word

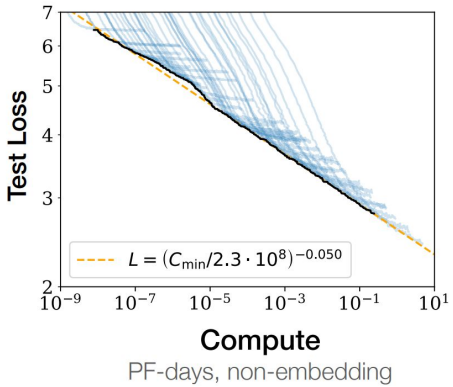
Scaling hypothesis

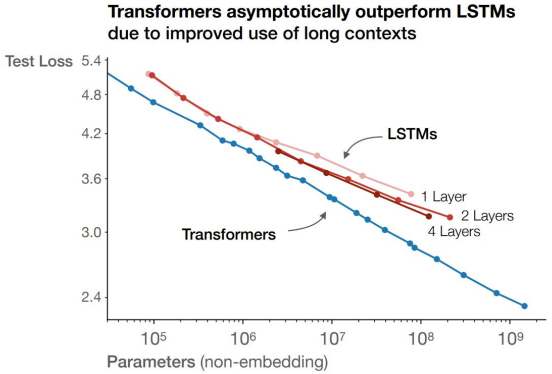
Three key inputs to a language model

- Number of parameters
- Amount of compute used at training
- Number of words seen during the whole training

The scaling hypothesis (Kaplan 2020) is roughly that you can fit a power law of perplexity to these three inputs

- Architectural decisions only shift the curve a bit
- Real power is in more data





Effects of scaling

Perplexity gets better in a smooth way, but capabilities *emerge*

Recent research has been training very, very, very large language models

- Chinchilla (Hoffmann et al) is typical at 70B parameters -- 3,000x larger than AlphaFold
- Trained on huge clusters with huge amounts of text

Result is models that start to show not just textual skills but also reasoning

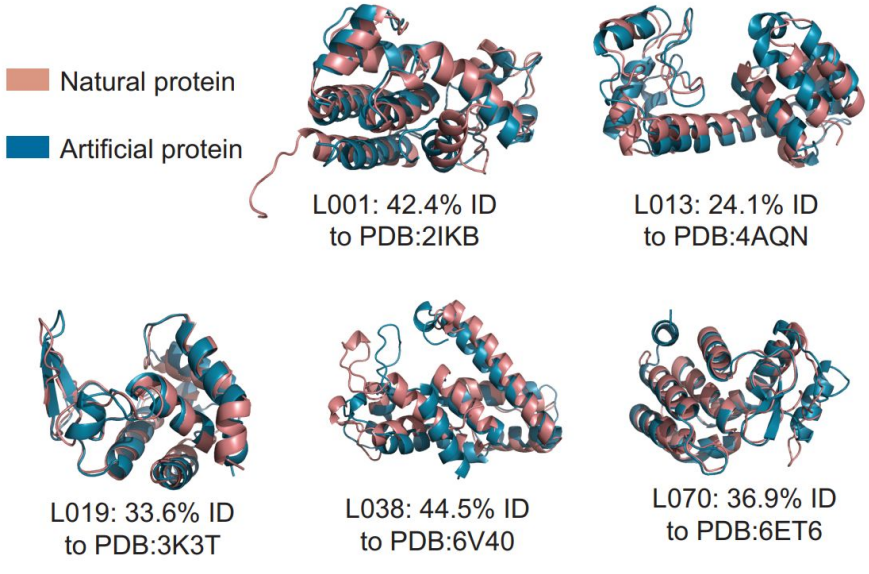
Key question is whether the same trick would work with protein sequences

Status of protein LMs

Large protein LMs have been trained, such as RITA and ProGen-2

Such models are beginning to show functional generation of proteins *without structural information*

Still depends heavily on having a common family, which makes sense since structural knowledge is weak without a very common family



ProGen designs from Madani et al

Key problems that will need ML solutions

- How are we going to interface with all the sparse experimental data that biologists have?
  - E.g. how do we use chemical cross-linking?
- How are we going to handle huge complexes that are becoming common?
- What is our role in interpreting cryo-ET data?
- How can we handle “negatives” structurally so that we can perform interaction screening?
- Can we say anything useful about mutation effects that shift populations?
- Sampling seems to be returning for AF-related methods -- is that a feature or an opportunity for improvement?
- How can we handle PTMs and associated state changes? The data will be sparse