Generative Models for Real-World Drug Discovery

Amy X. Lu
April 30th, 2025
PhD Dissertation Talk
BAIR Seminar

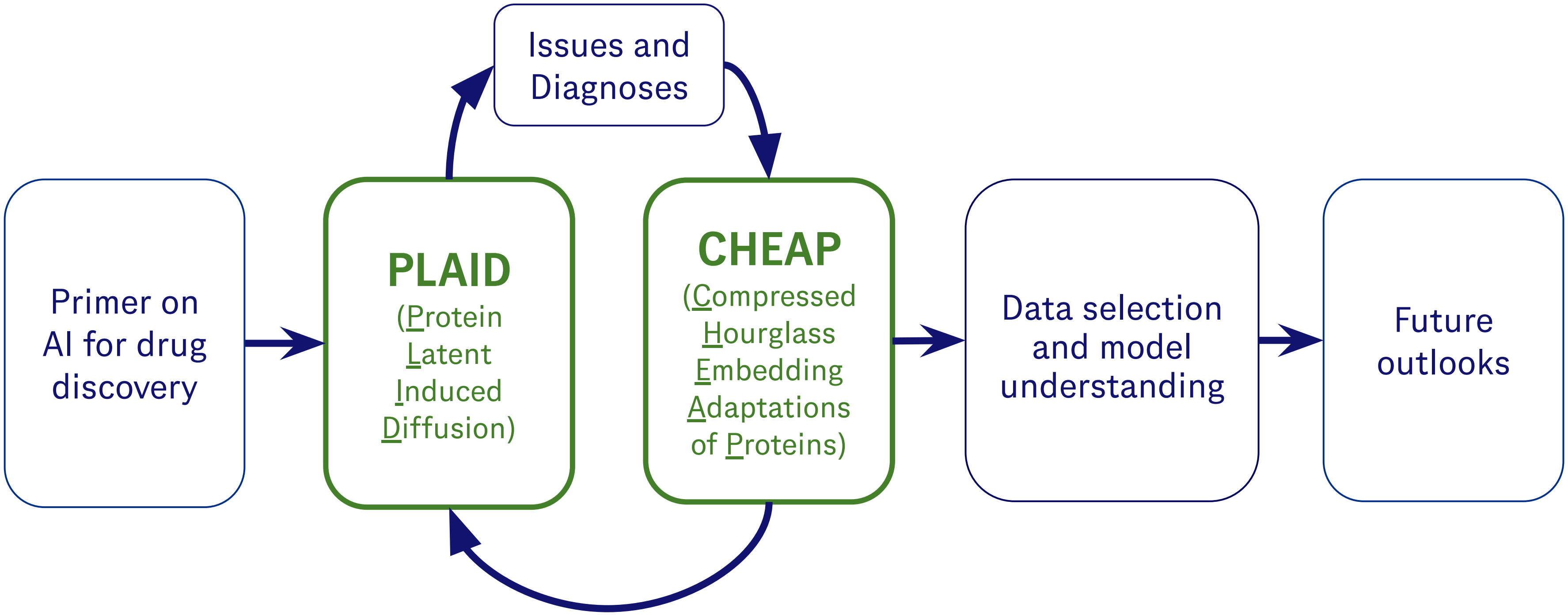
Genentech
A Member of the Roche Group

→ biology as a data modality for generative modeling

Generative Models for Real-World Drug Discovery

→ evaluations/tasks anchored around drug discovery for protein design

Agenda



Primer on AI for drug discovery

PLAID

(Protein Latent Induced Diffusion)

Issues and Diagnoses

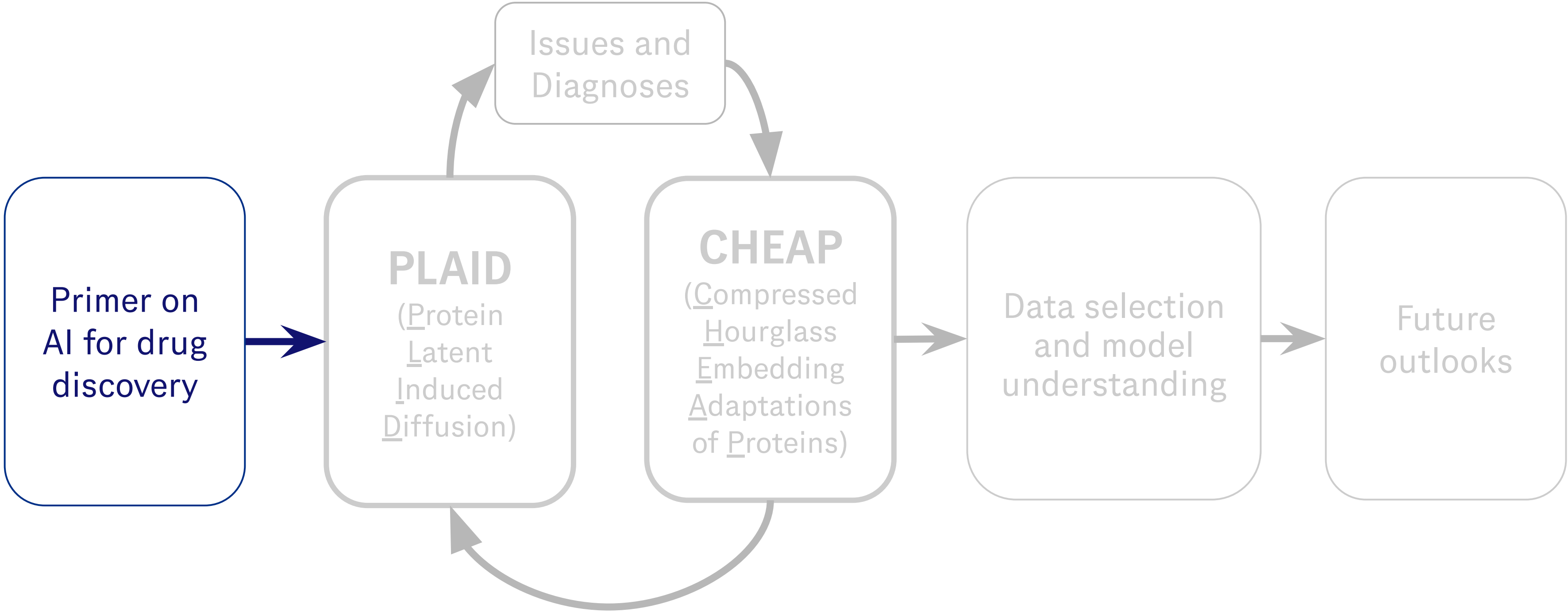
CHEAP

(Compressed Hourglass Embedding Adaptations of Proteins)

Data selection and model understanding

Future outlooks

Agenda



Primer on AI for drug discovery

Issues and Diagnoses

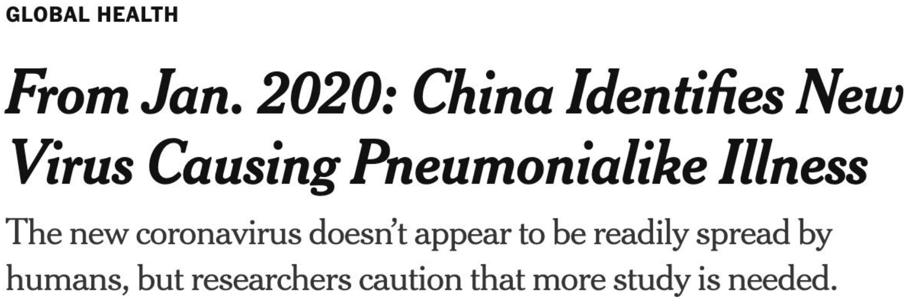
PLAID (Protein Latent Induced Diffusion)

CHEAP (Compressed Hourglass Embedding Adaptations of Proteins)

Data selection and model understanding

Future outlooks

A simplified look at modern drug discovery…

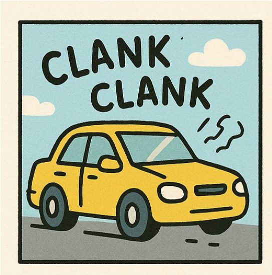


GLOBAL HEALTH

From Jan. 2020: China Identifies New Virus Causing Pneumonialike Illness

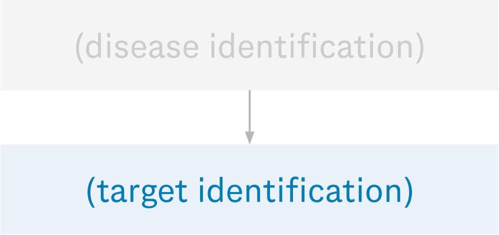
The new coronavirus doesn't appear to be readily spread by humans, but researchers caution that more study is needed.

A simplified look at modern drug discovery…



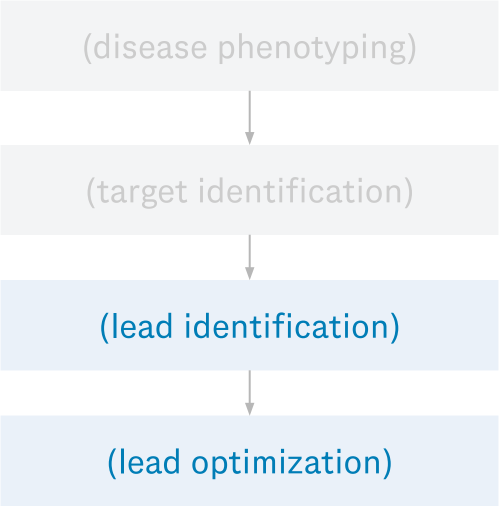
(disease identification)

A simplified look at modern drug discovery…


(disease identification)

(target identification)

A simplified look at modern drug discovery…



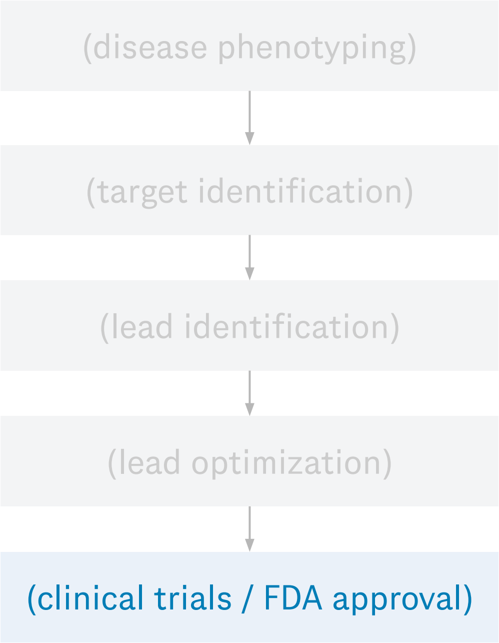
(disease phenotyping)

(target identification)

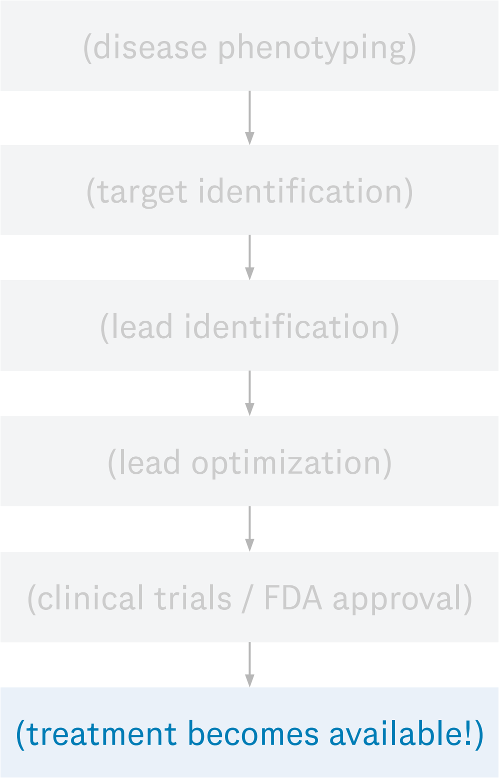
(lead identification)

(lead optimization)

A simplified look at modern drug discovery…



A simplified look at modern drug discovery…



(disease phenotyping)

(target identification)

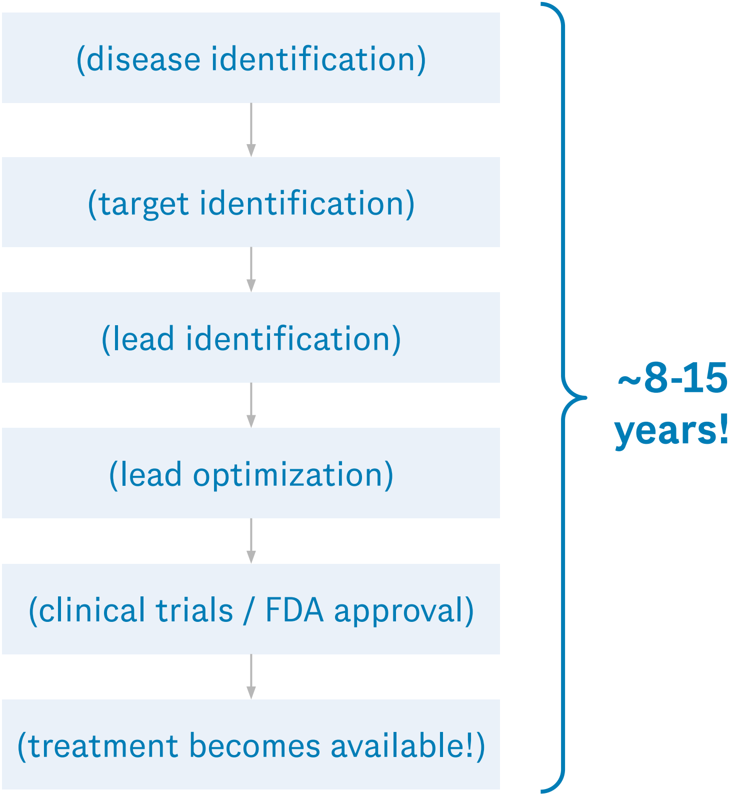
(lead identification)

(lead optimization)

(clinical trials / FDA approval)

(treatment becomes available!)

Drug discovery is time-consuming



(disease identification)

(target identification)

(lead identification)

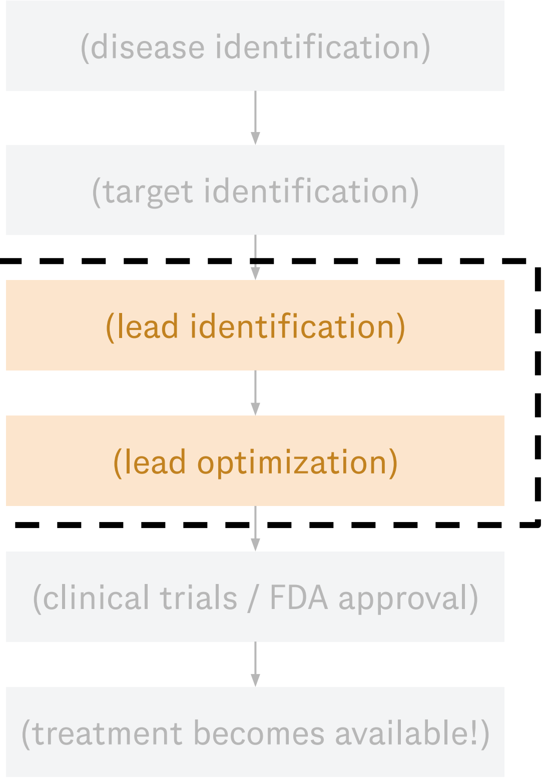
(lead optimization)

(clinical trials / FDA approval)

(treatment becomes available!)

~8-15 years!

Accelerating drug discovery with AI?


(disease identification)

(target identification)

(lead identification)

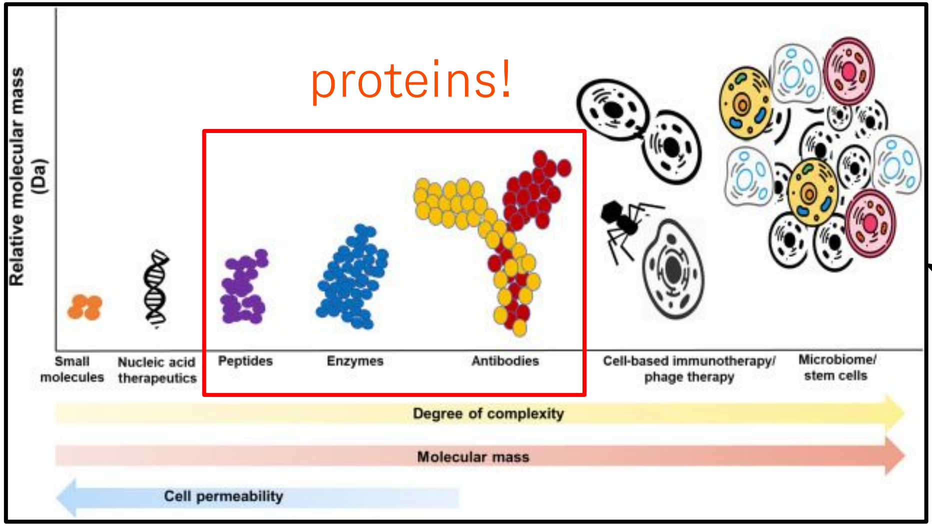
(lead optimization)

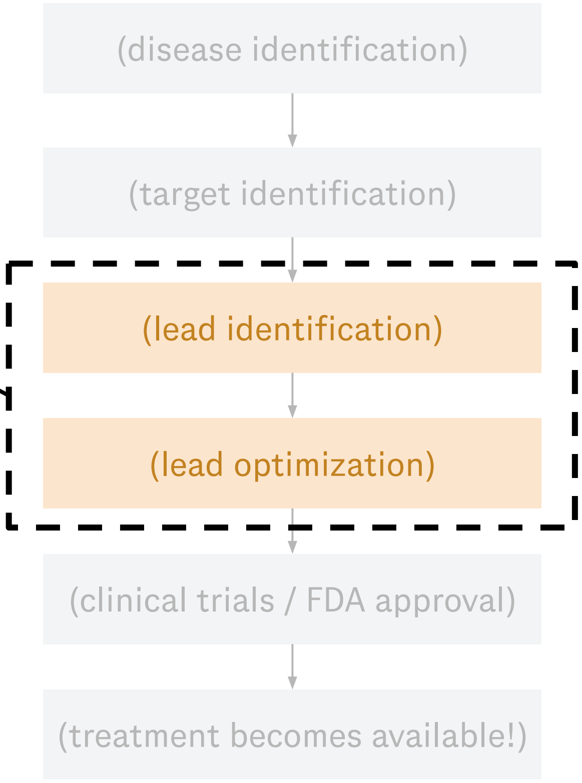
(clinical trials / FDA approval)

(treatment becomes available!)

🧠🤖
?

Accelerating protein design with AI?





(disease identification)

(target identification)

(lead identification)

(lead optimization)

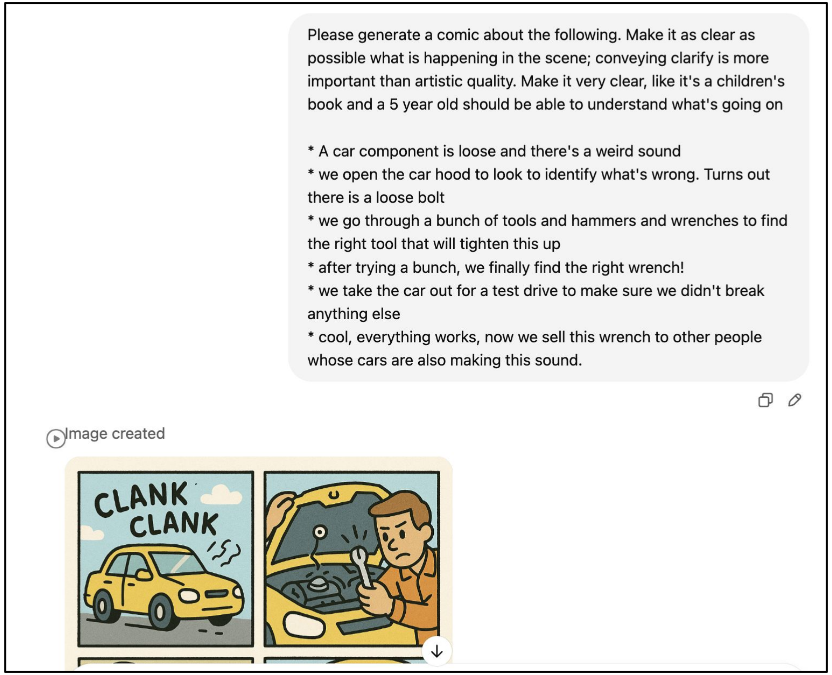
(clinical trials / FDA approval)

(treatment becomes available!)

🧠🤖

?

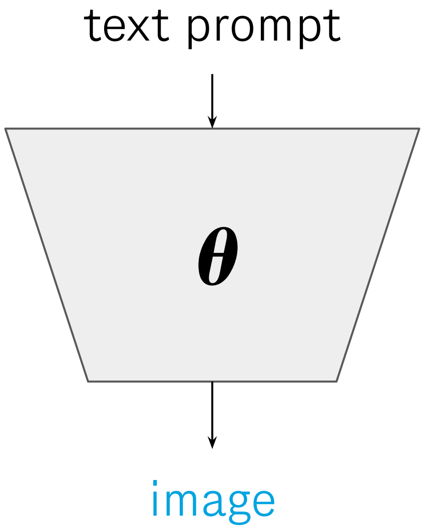
Accelerating protein design with AI?



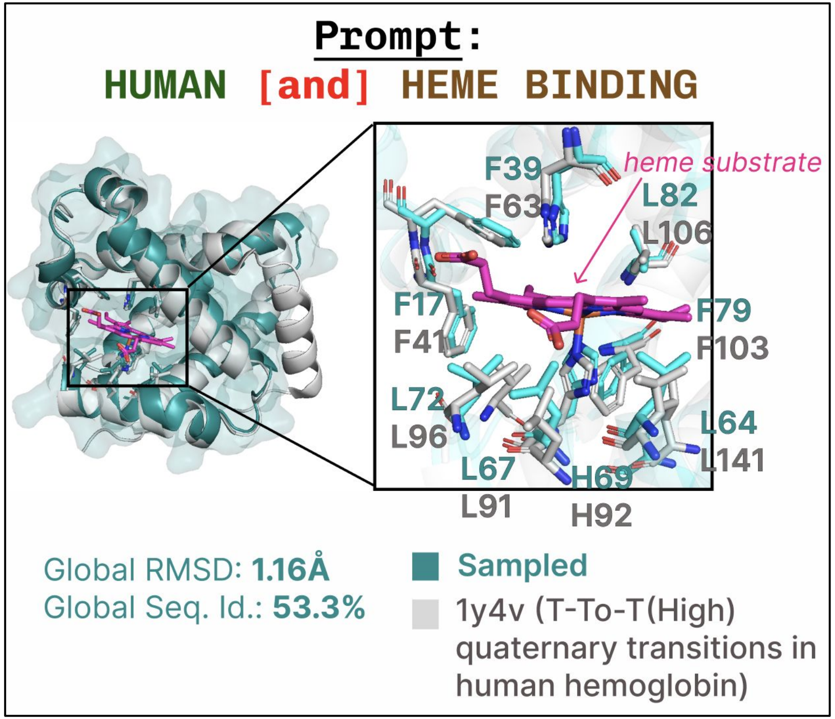
Please generate a comic about the following. Make it as clear as possible what is happening in the scene; conveying clarify is more important than artistic quality. Make it very clear, like it's a children's book and a 5 year old should be able to understand what's going on

* A car component is loose and there's a weird sound
* we open the car hood to look to identify what's wrong. Turns out there is a loose bolt
* we go through a bunch of tools and hammers and wrenches to find the right tool that will tighten this up
* after trying a bunch, we finally find the right wrench!
* we take the car out for a test drive to make sure we didn't break anything else
* cool, everything works, now we sell this wrench to other people whose cars are also making this sound.

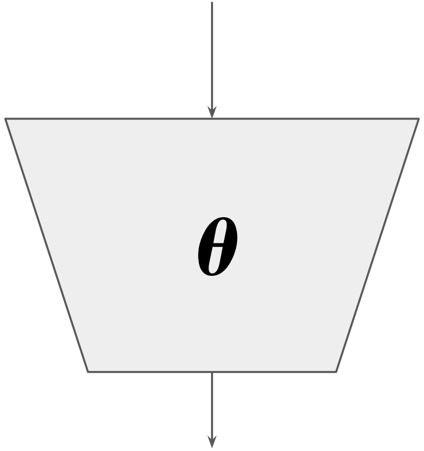
Image created



Accelerating protein design with AI?

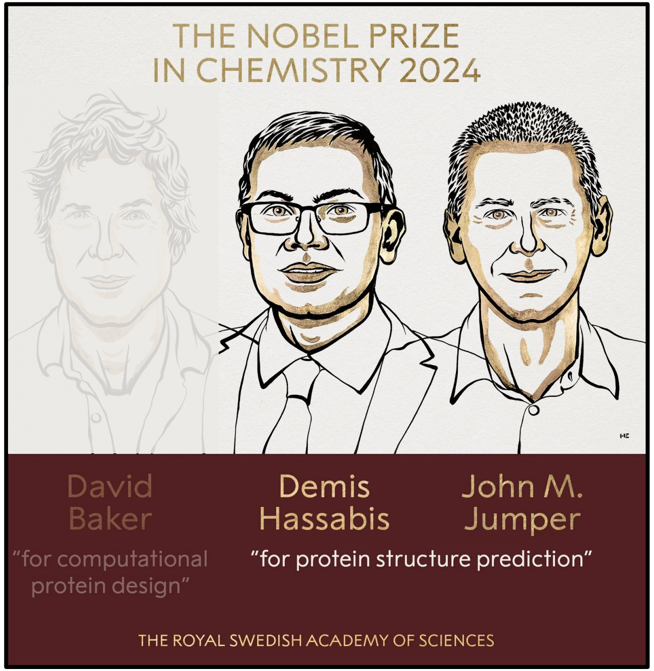


Desired attributes
(ex. “heme binding”)



protein / molecule!

The potential of deep learning for protein structure prediction

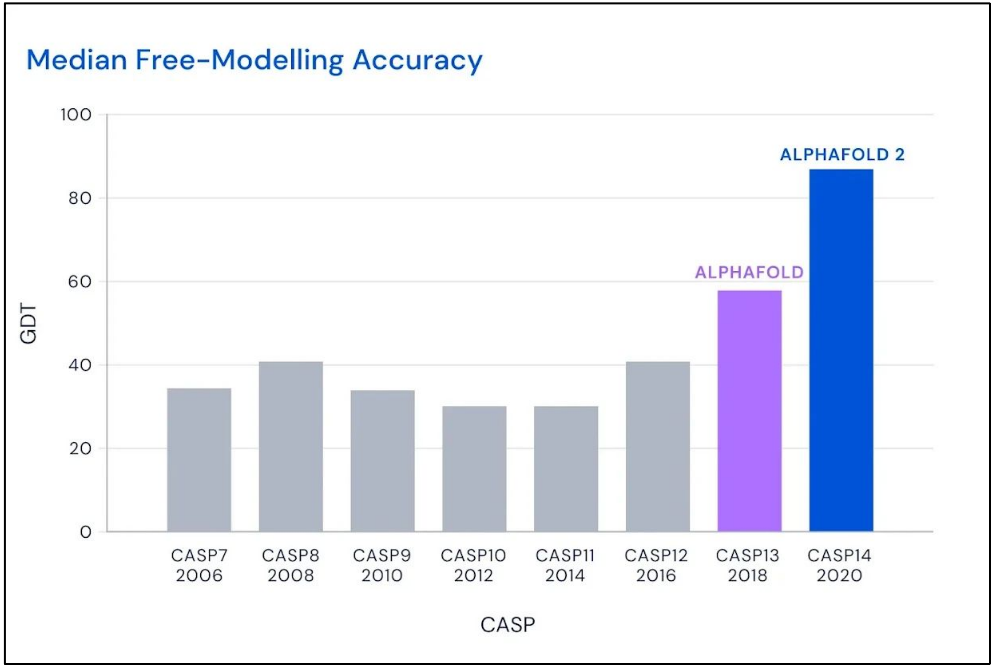


THE NOBEL PRIZE IN CHEMISTRY 2024

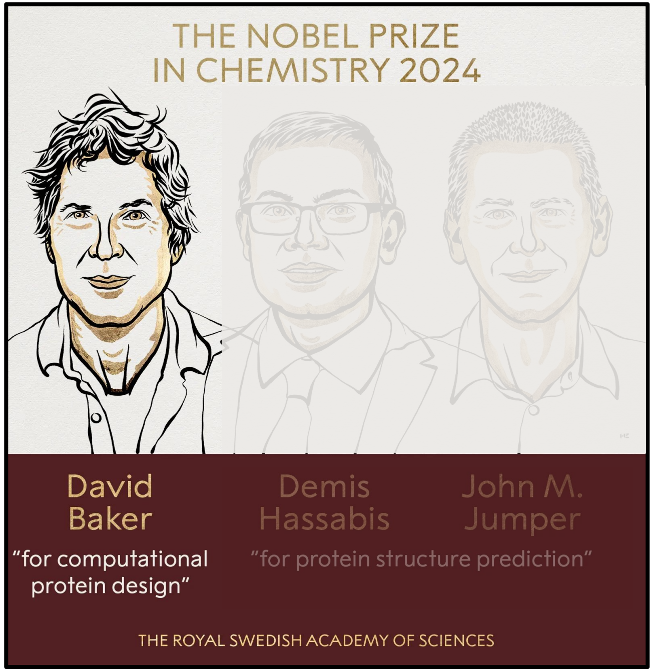
David Baker "for computational protein design"

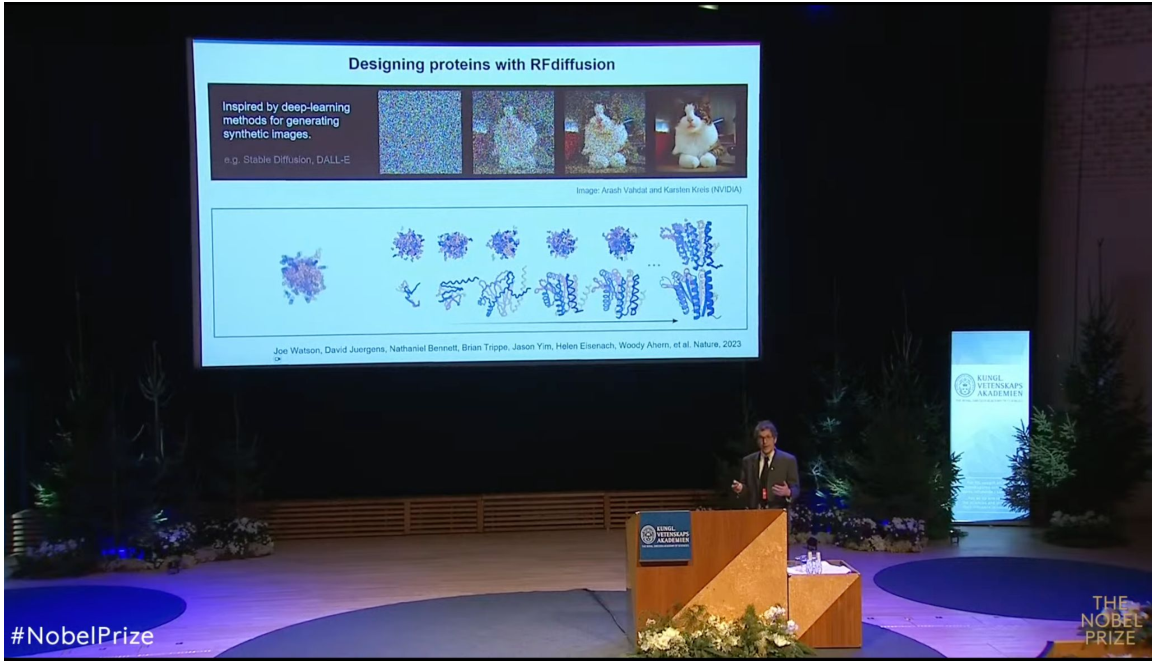
Demis Hassabis, John M. Jumper "for protein structure prediction"

THE ROYAL SWEDISH ACADEMY OF SCIENCES



The potential of deep learning for protein structure generation




Designing proteins with RFdiffusion

Inspired by deep-learning methods for generating synthetic images.

e.g. Stable Diffusion, DALL-E

Image: Arash Vahdat and Karsten Kreis (NVIDIA)

Joe Watson, David Juergens, Nathaniel Bennett, Brian Trippe, Jason Yim, Helen Eisenach, Woody Ahern, et al. Nature, 2023

What else might we need for drug discovery?

this talk

Co-generation
→ can we simultaneously generate sequence and structure?

Control
→ how can we specify complex and multi-objective constraints?

Immunogenicity & antigen expression
→ can we achieve organism specificity?

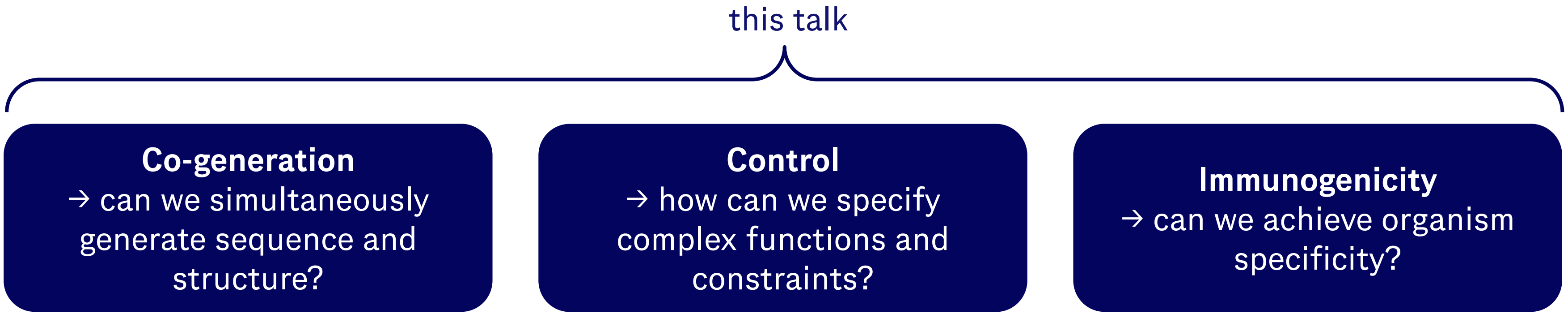
Deployment
→ can we speed up inference to improve “shots on goal”?

Biosecurity
→ how should we measure and prevent the potential for dual use?

Data curation
→ how should we collect data for model (pre)training in costly acquisition regimes?

other PhD works

What else might we need for drug discovery?



this talk

Co-generation
→ can we simultaneously generate sequence and structure?

Control
→ how can we specify complex functions and constraints?

Immunogenicity
→ can we achieve organism specificity?

Compressed protein representations
(Cell Patterns, to appear)

Latent diffusion for all-atom generation
(in submission)

What else might we need for drug discovery?

other PhD research: **deployment & model understanding**

Model-based optimization for protein engineering
(Kolli et al., 2022)

Dense passage retrieval for homology search
(Boger et al., 2023)

Guided diffusion with differentiable biophysical energies
(unpublished)

Effect of training data compositions on protein language model likelihoods
(Gordon et al., 2024)

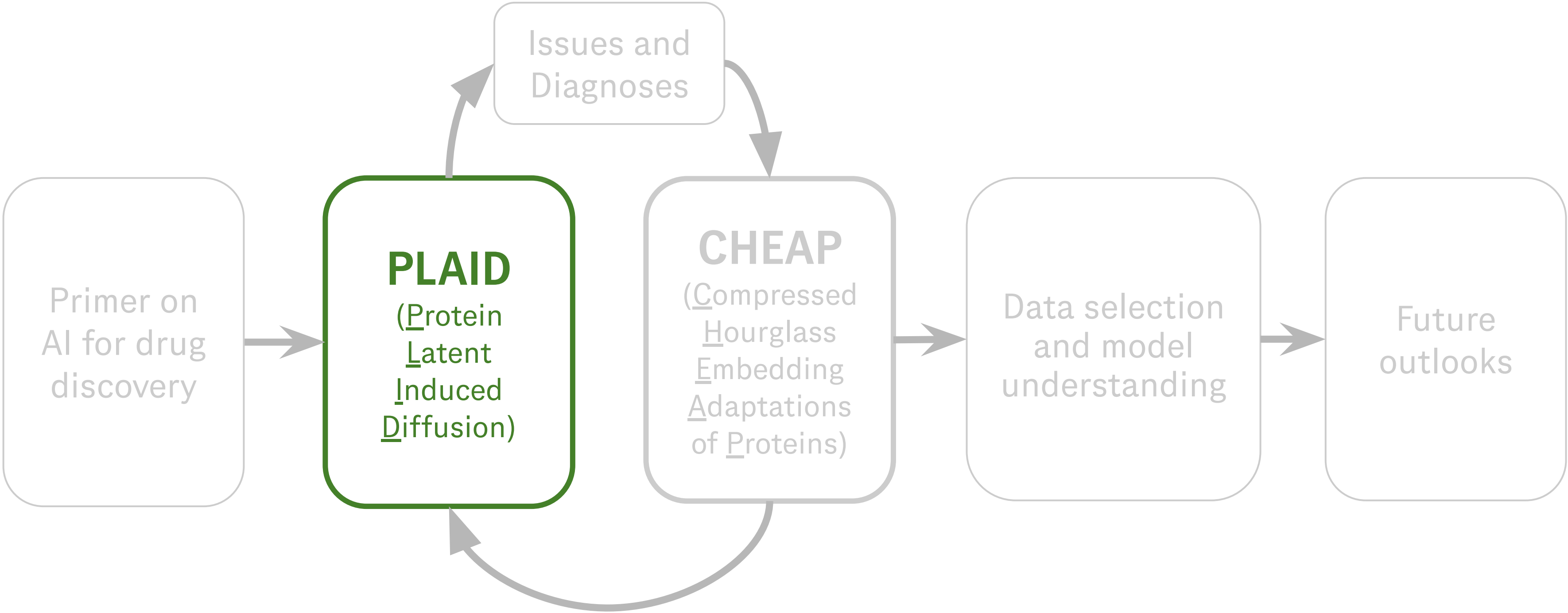
Evo2 biosecurity and inference optimization
(Brixi et al., 2025)

Compressed protein representations
(Cell Patterns, to appear)

Latent diffusion for all-atom generation
(in submission)

Biological data selection from an information theoretic perspective
(in progress)

Agenda



Primer on AI for drug discovery

PLAID

(Protein Latent Induced Diffusion)

Issues and Diagnoses

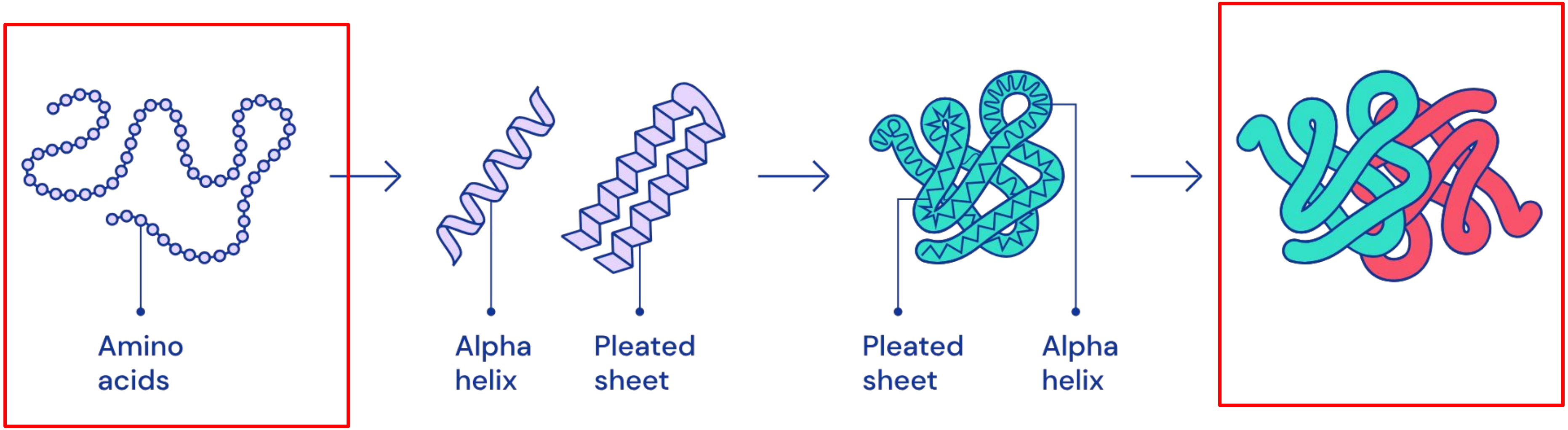
CHEAP

(Compressed Hourglass Embedding Adaptations of Proteins)

Data selection and model understanding

Future outlooks

What exactly is a protein?

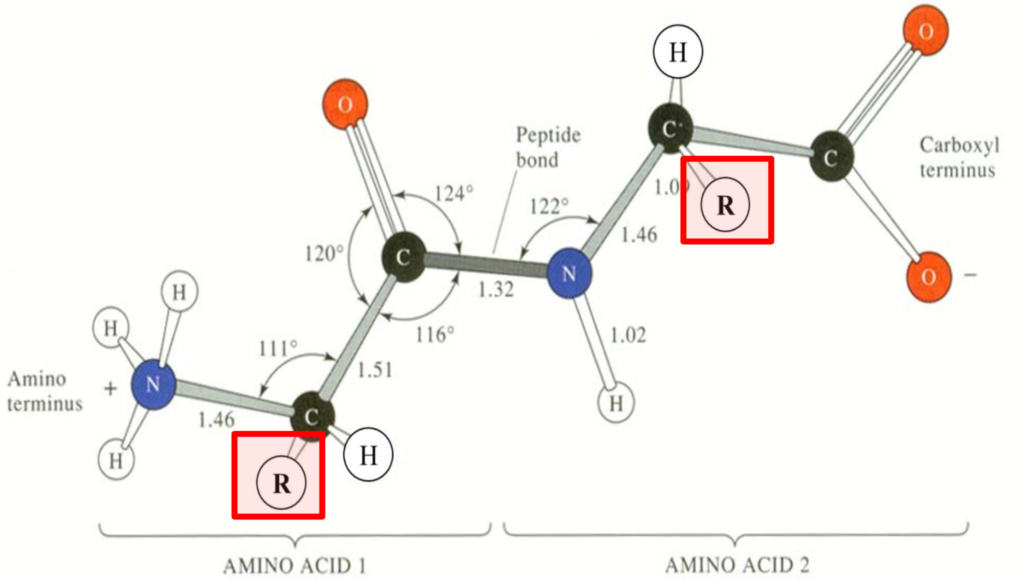


"text"
1D string

"image/video"
3D positions

Image source: AlphaFold1 blog post

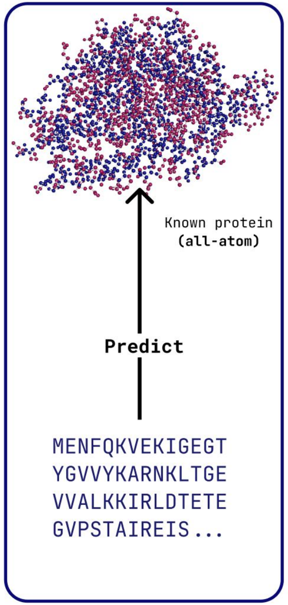
Backbone structure vs. all-atom structure



GBYR...

(order of t-shirts => protein sequence)

The co-generation problem

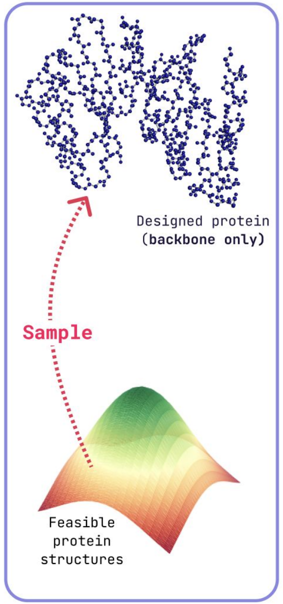


Known protein (all-atom)

Predict

MENFQKVEKIGEGT
YGVVYKARNKLTGE
VVALKKIRLDTETE
GVPSTAIREIS...

AlphaFold / ESMFold
Prediction only

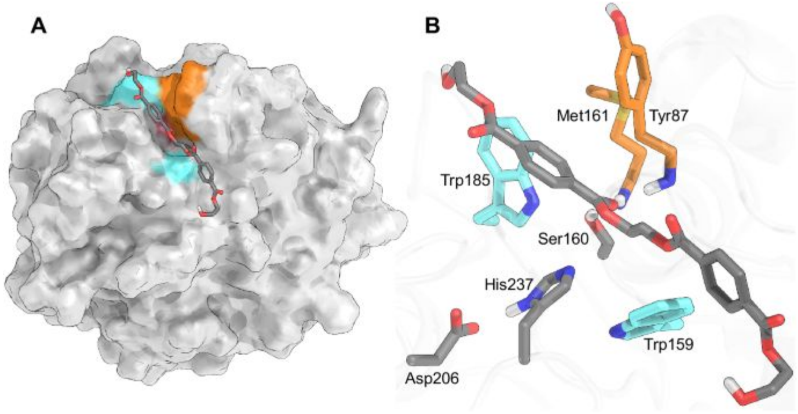


Designed protein (backbone only)

Sample

Feasible protein structures

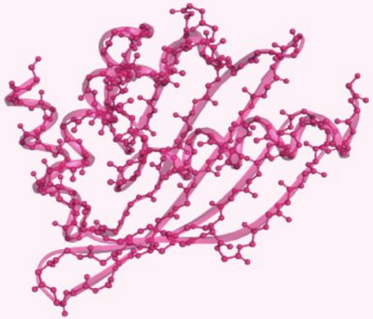
Previous generative methods
Backbone-only



sidechains are crucial for mediating function!

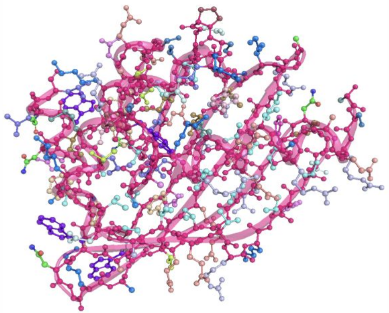
Sidechain atoms generation require knowing the sequence

BACKBONE ATOMS



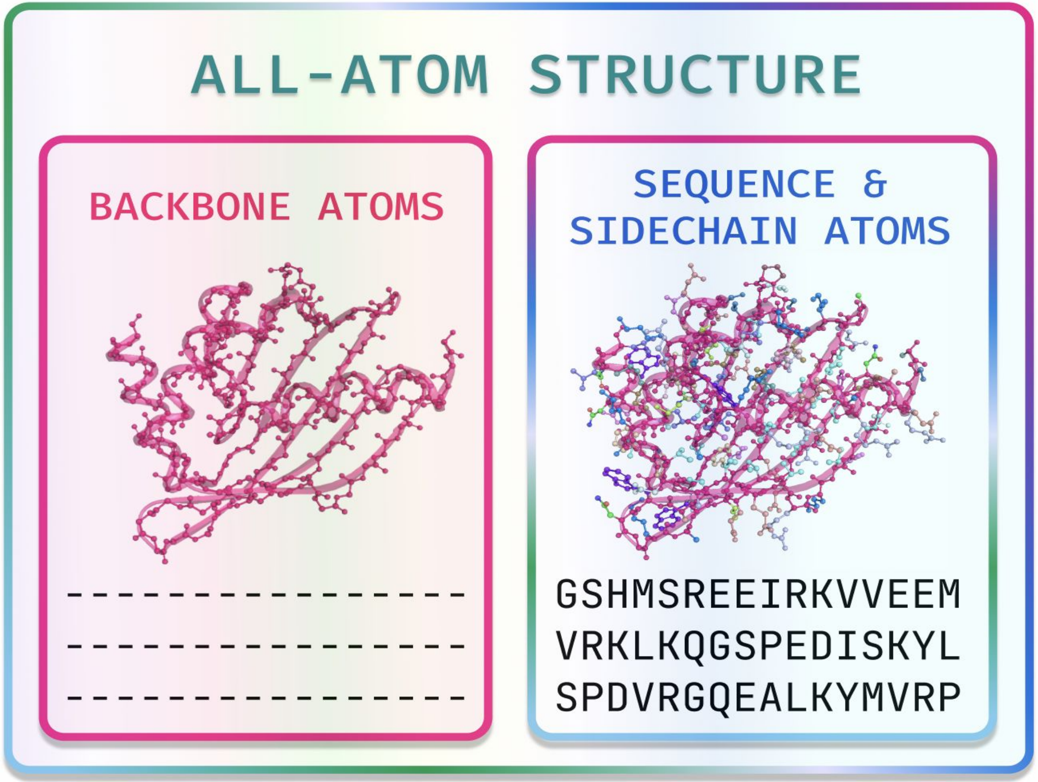
```
-----------------
-----------------
-----------------
```

SEQUENCE & SIDECHAIN ATOMS

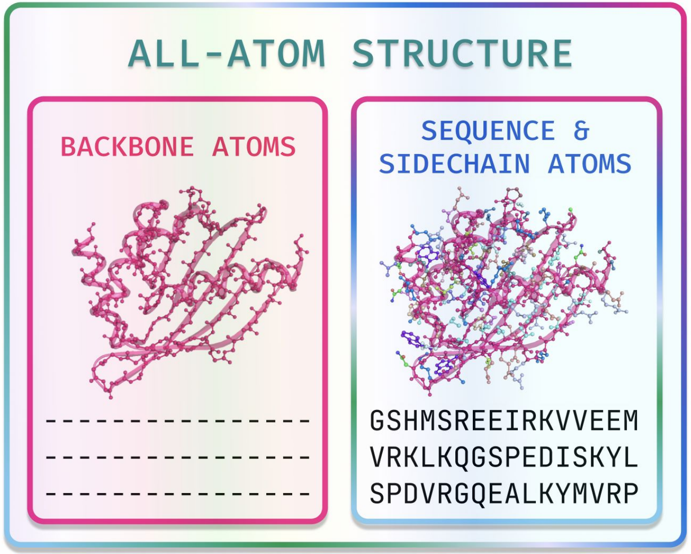


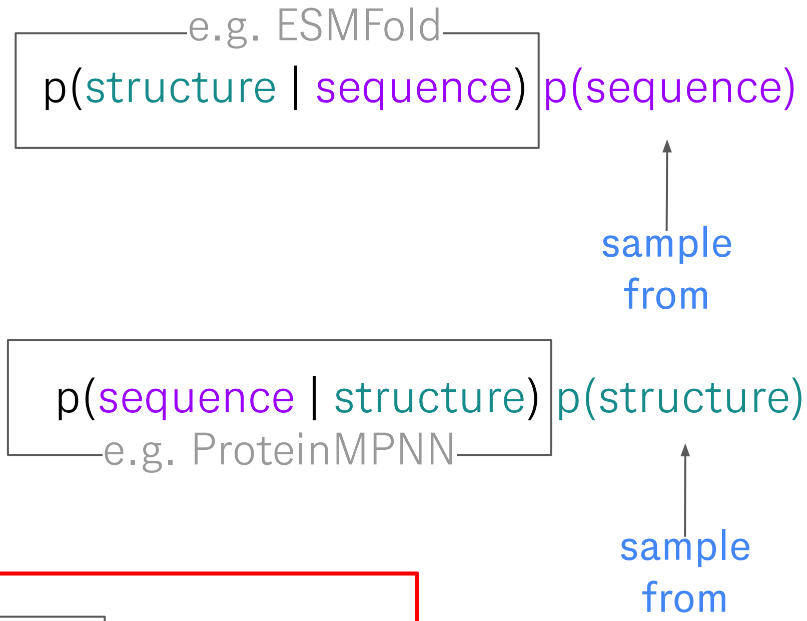
GSHMSREEIRKVVEEM
VRKLKQGSPEDISKYL
SPDVRGQEALKYMVRP

All-atom design as a multimodal generation problem



All-atom design as a multimodal generation problem




e.g. ESMFold

$p(\text{structure} \mid \text{sequence})\ p(\text{sequence})$

sample from

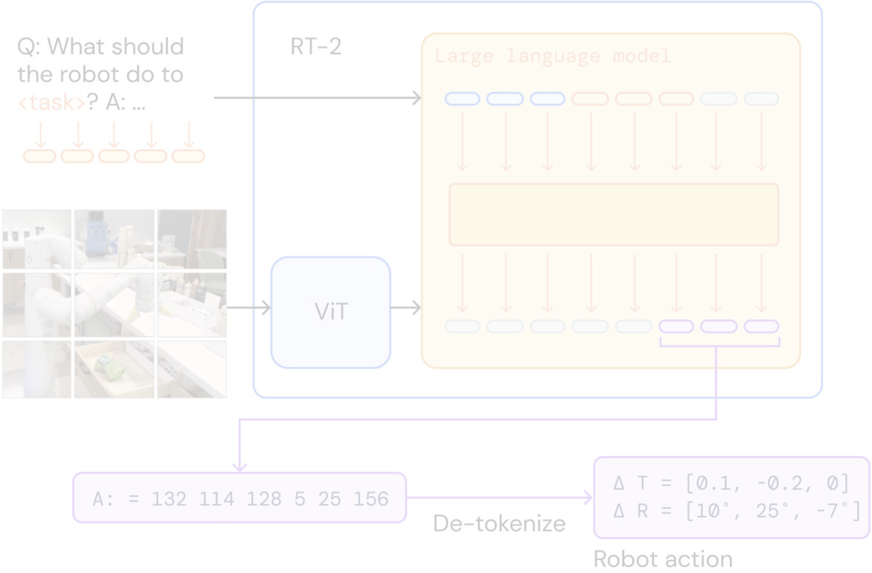
$p(\text{sequence} \mid \text{structure})\ p(\text{structure})$

e.g. ProteinMPNN

sample from

Goal: $p(\text{sequence}, \text{structure})$ ← sample from

Motivation: Can we repurpose priors from pretrained models?



RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control

Vision-language models trained on internet-scale datasets capture useful priors for decision making tasks.

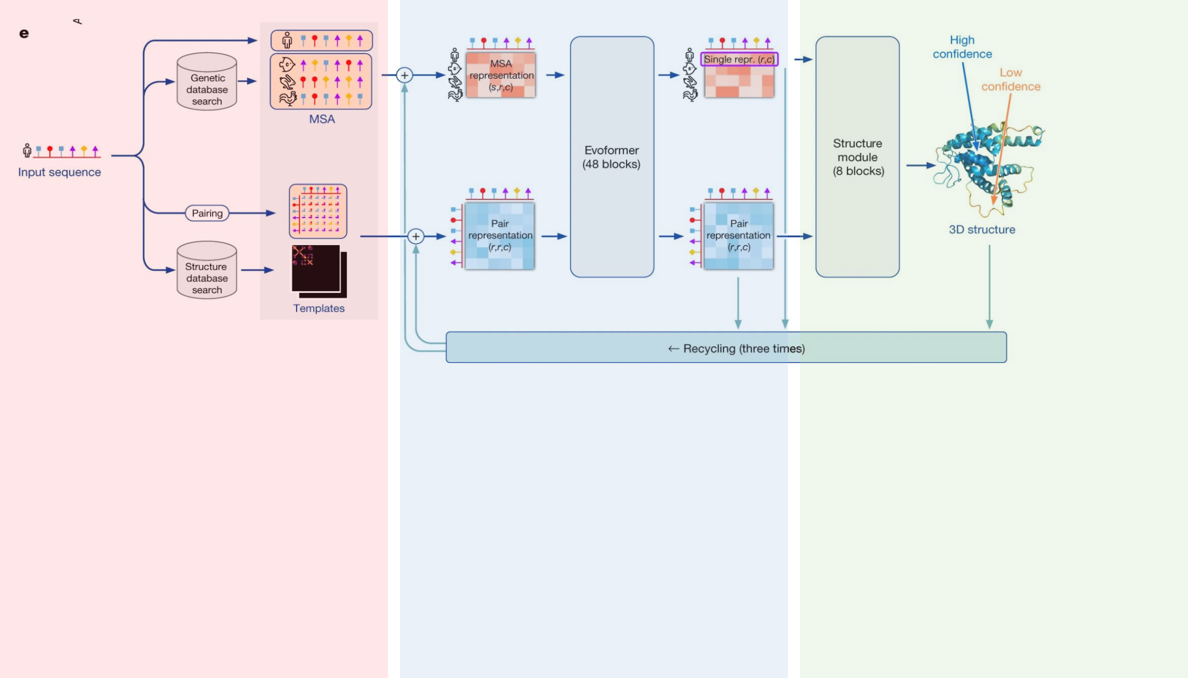
Can we apply this to biology?

Can we sample all-atom structure from the joint distribution p(sequence, structure) and use priors from pretrained protein folding models?

The base components: protein folding model architectures

AlphaFold2:

Uses an explicit retrieval step


e
Input sequence
Genetic database search
MSA
Pairing
Structure database search
Templates
MSA representation (s,r,c)
Pair representation (r,r,c)
Evoformer (48 blocks)
Single repr. (r,c)
Pair representation (r,r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (three times)

harness additional sequence-based priors

learn structural features from sequence latents

generate structures

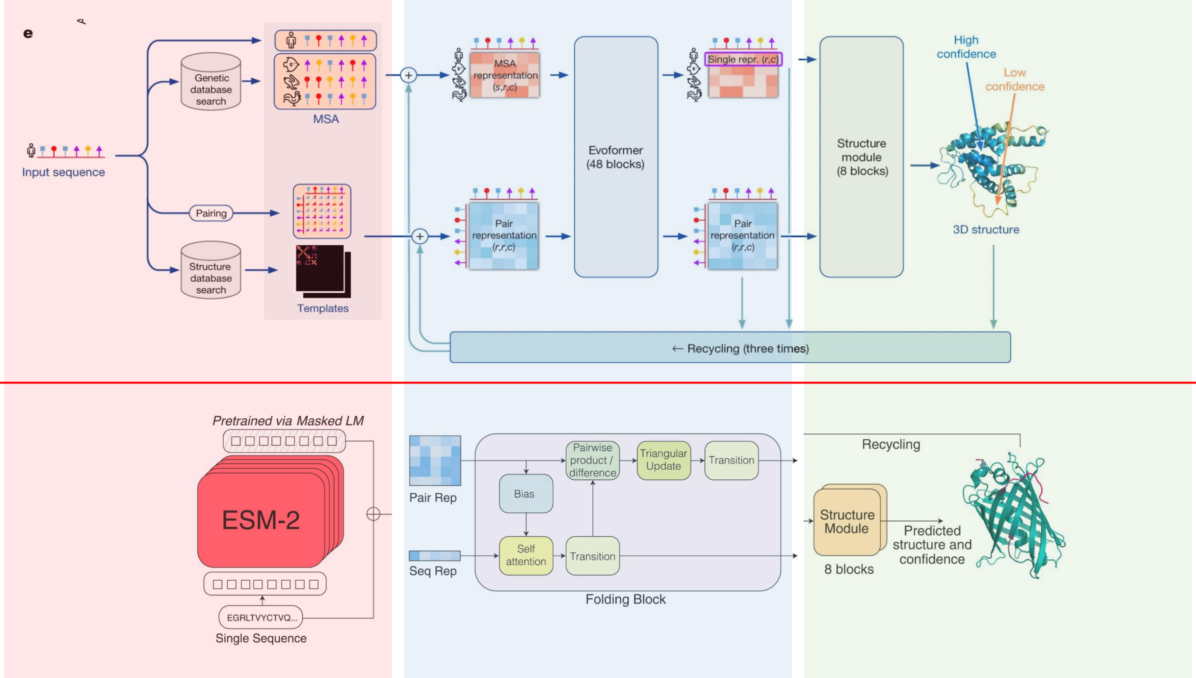
The base components: protein folding model architectures

AlphaFold2:

Uses an explicit retrieval step

ESMFold:

Replaces retrieval step with a **language model**



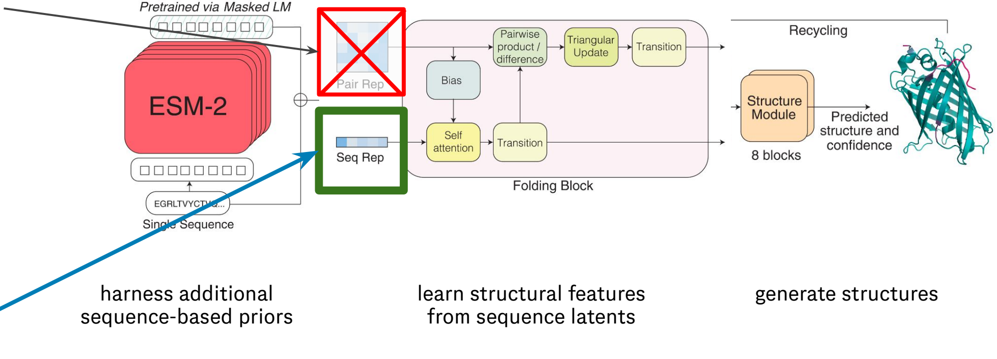
harness additional sequence-based priors

learn structural features from sequence latents

generate structures

Observation: at inference, the pairwise input is initialized as zeros…

→ Sequence representation contains all information about the structure!



harness additional sequence-based priors

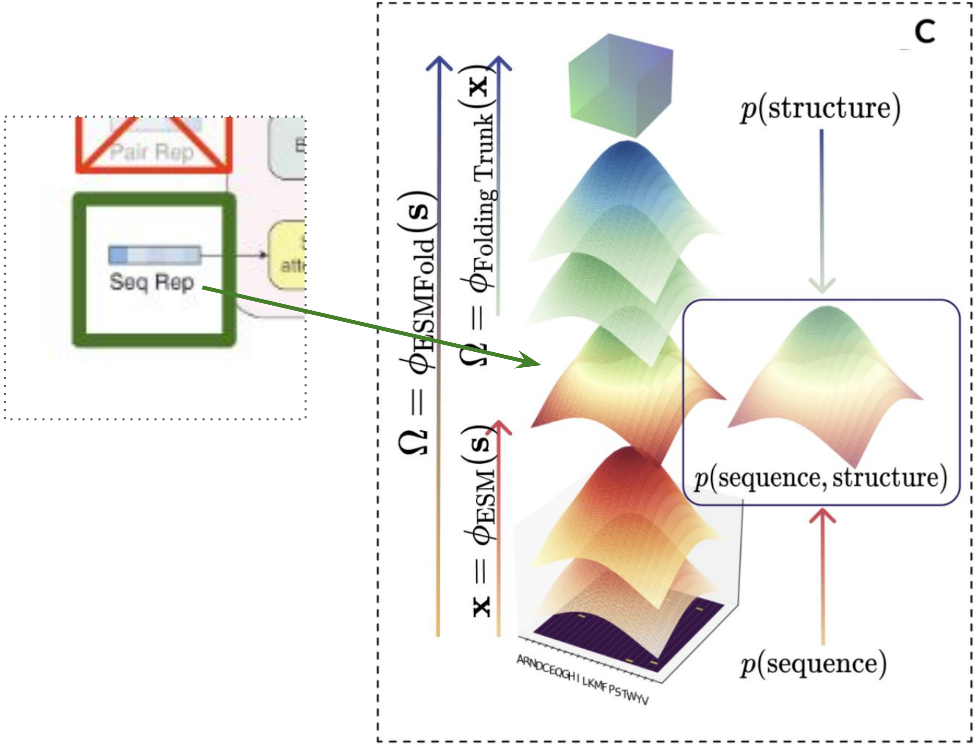
learn structural features from sequence latents

generate structures

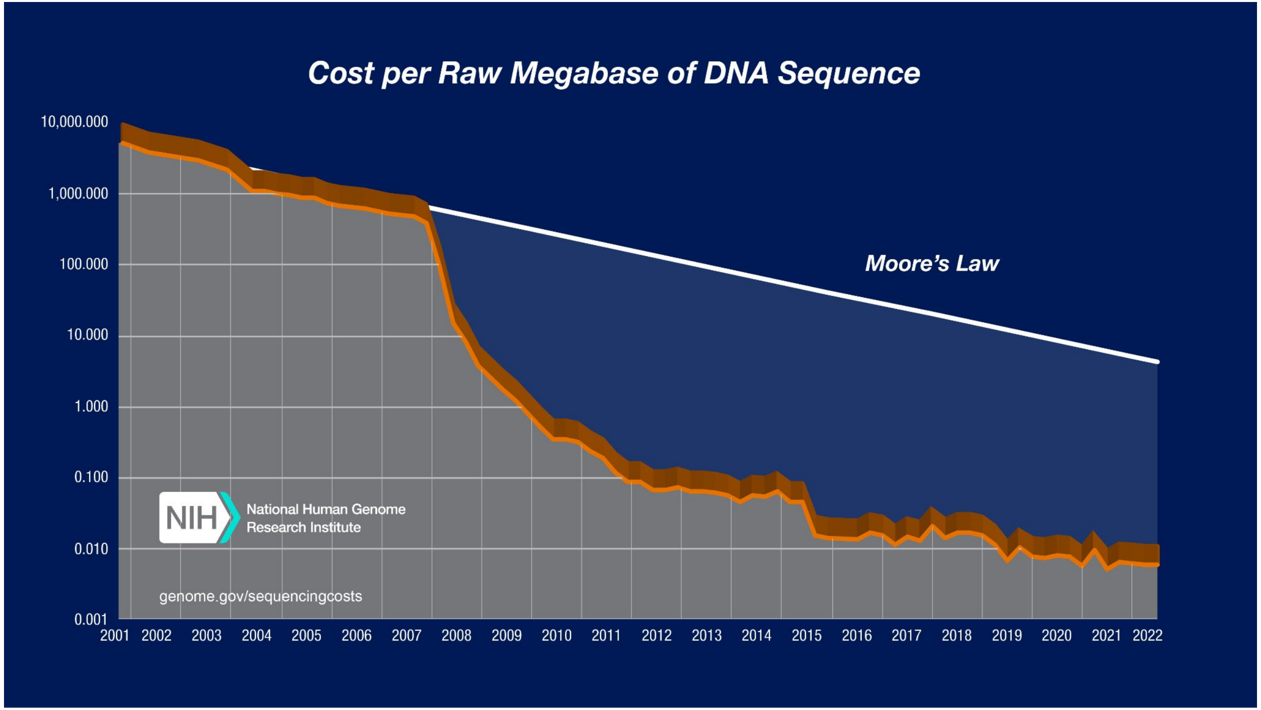
Observation: at inference, the pairwise input is initialized as zeros…

→ Sequence representation contains all information about the structure!

Generating this embedding would only **require the sequence during training (!)**

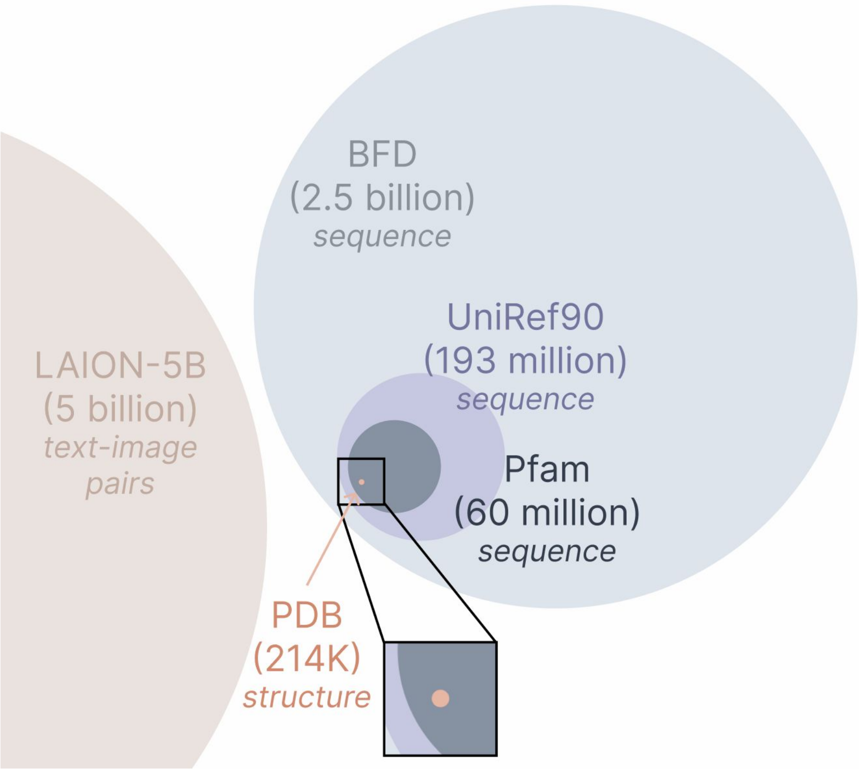

Pair Rep
Seq Rep
att
c
$\Omega = \phi_{\text{ESMFold}}(\mathbf{s})$
$\Omega = \phi_{\text{Folding Trunk}}(\mathbf{x})$
$\mathbf{x} = \phi_{\text{ESM}}(\mathbf{s})$
$p(\text{structure})$
$p(\text{sequence, structure})$
$p(\text{sequence})$
ARNDCEQGHILKMFPSTWYV

Sequence data is cheaper to collect than structure

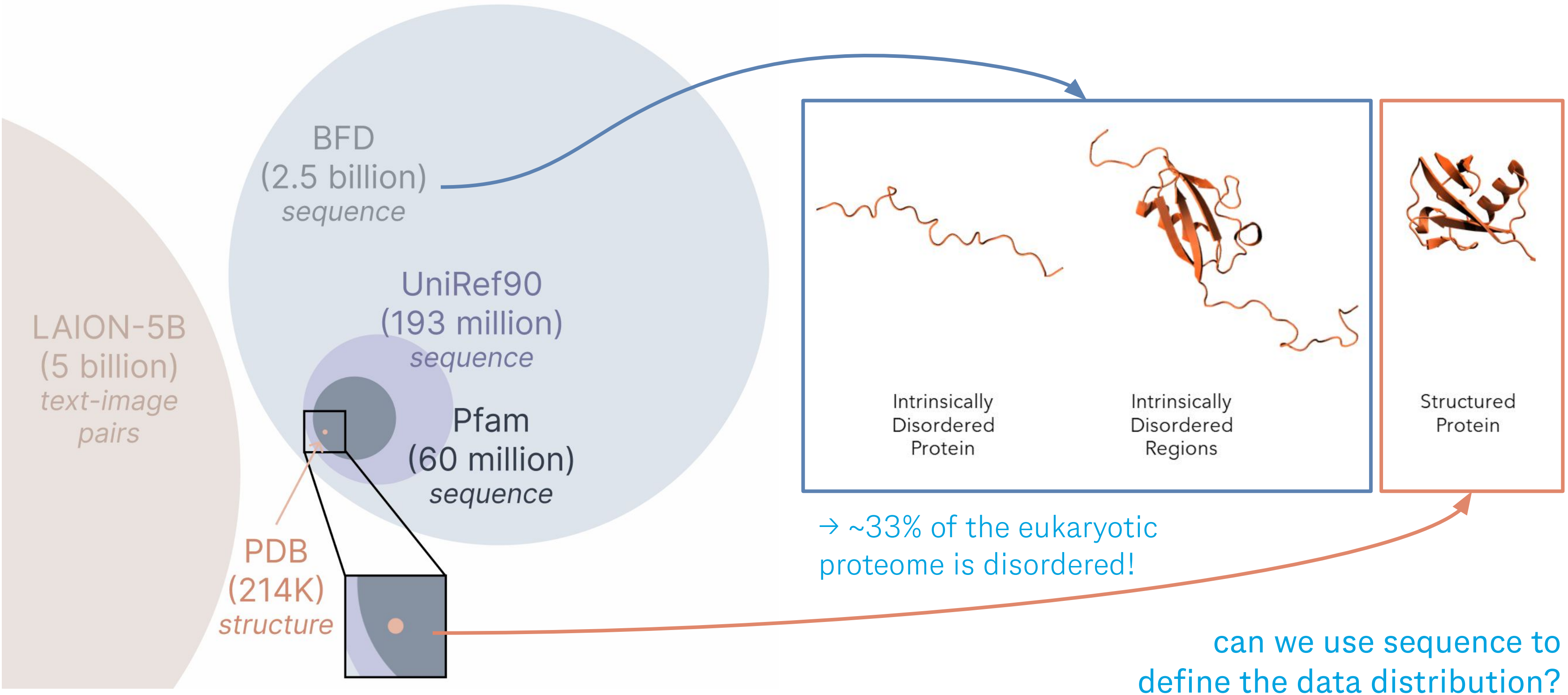


Source: https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

Sequence data is more abundant than structure



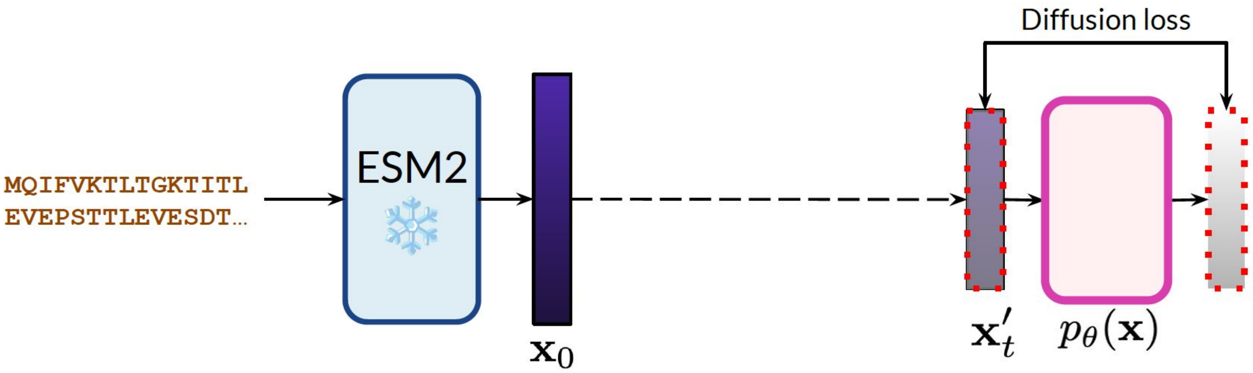
Sequence data has different coverage than structure



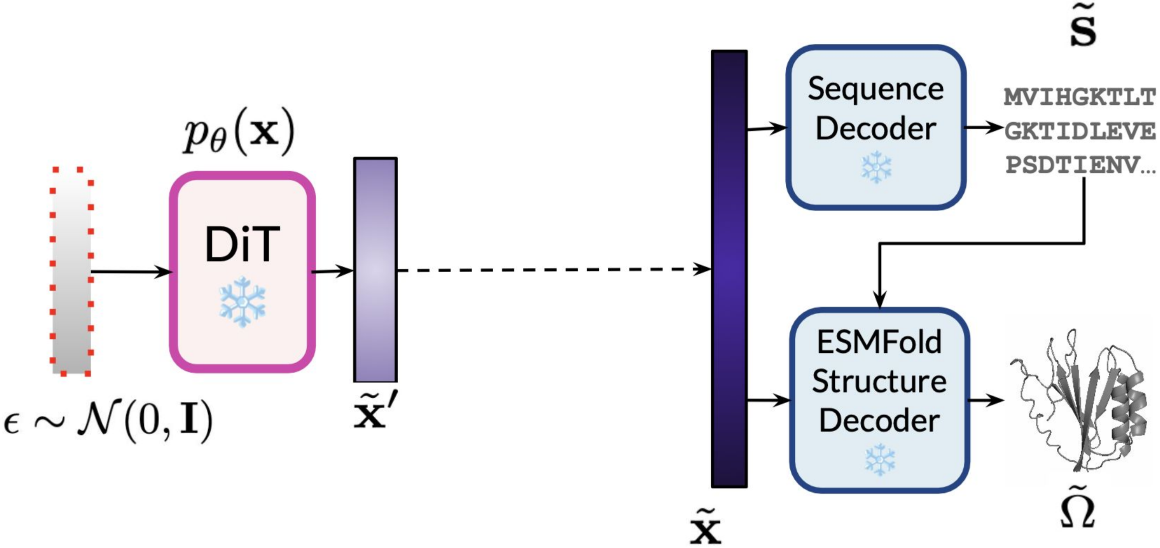
→ ~33% of the eukaryotic proteome is disordered!

can we use sequence to define the data distribution?

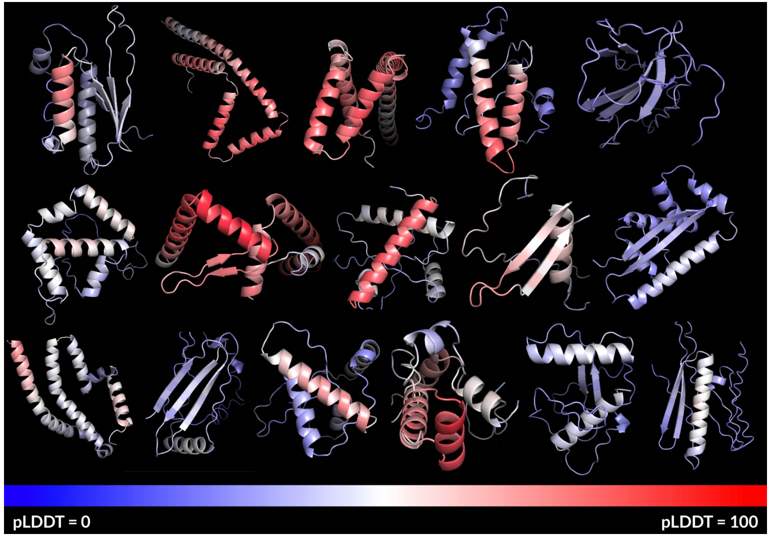
PLAID v0.5: Training a latent diffusion model

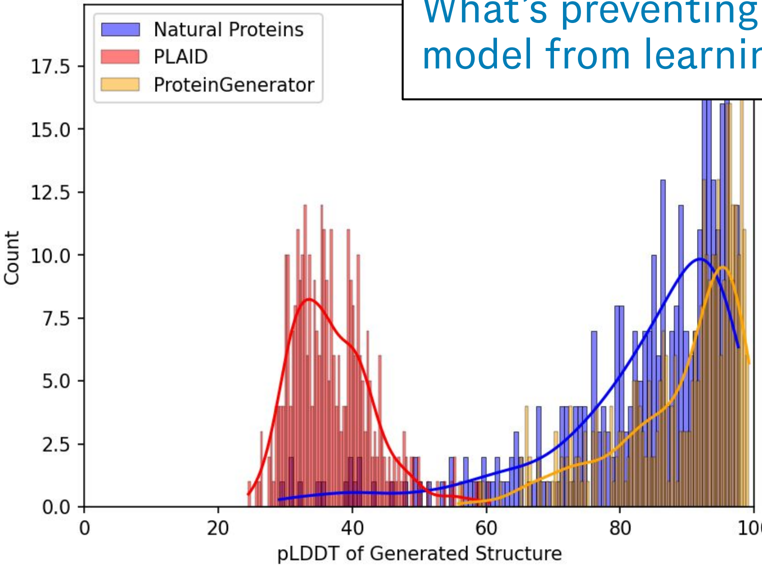


PLAID v0.5: Inference-time all-atom generation



PLAID v0.5: Early attempts

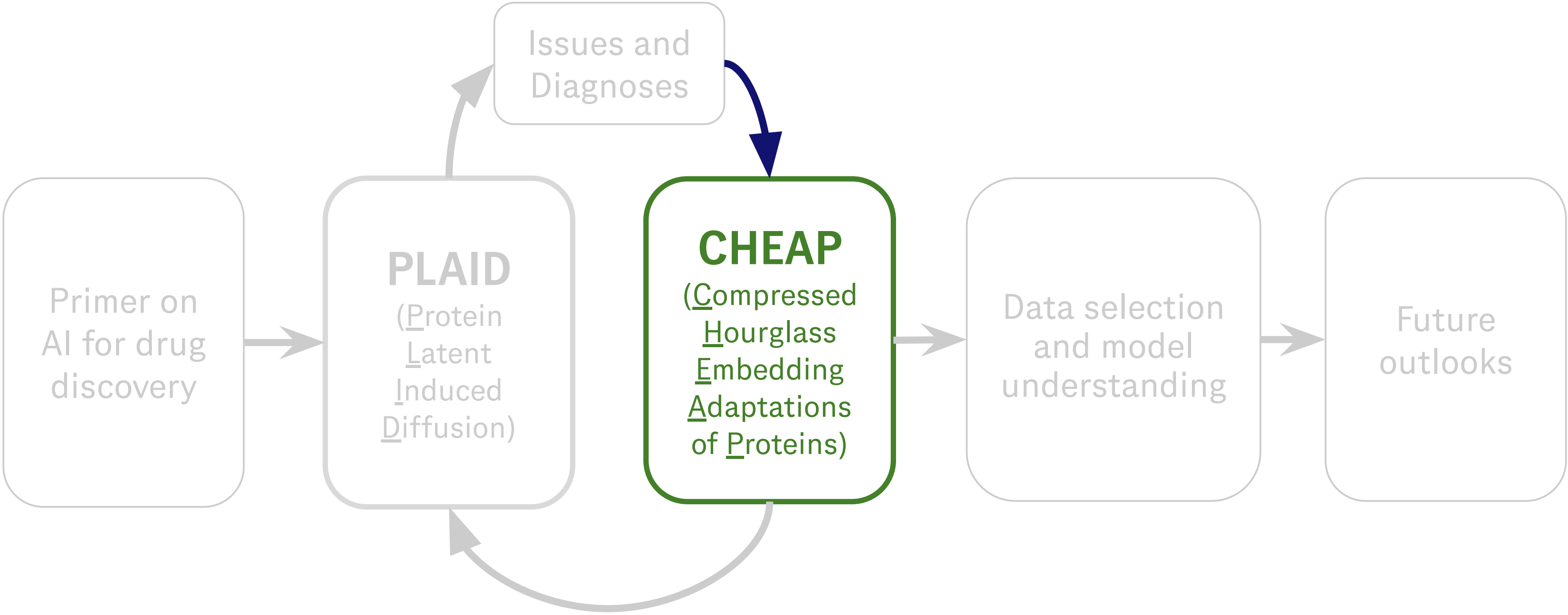




What's preventing the model from learning?

PLAID v0.5: Generating Protein Sequence and Structure Without Structural Training Data
Amy X. Lu, Kevin K. Yang, Pieter Abbeel
ICML 2024 Workshop on Machine Learning for Life and Material Sciences

Agenda



Primer on AI for drug discovery

PLAID (Protein Latent Induced Diffusion)

Issues and Diagnoses

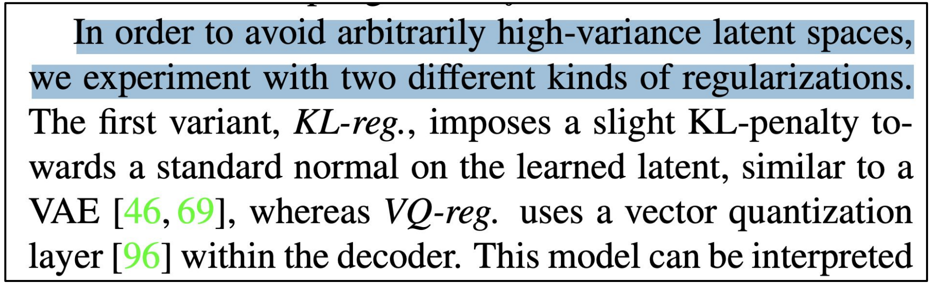
CHEAP (Compressed Hourglass Embedding Adaptations of Proteins)

Data selection and model understanding

Future outlooks

Issues and hypotheses

- Latent space requires regularization

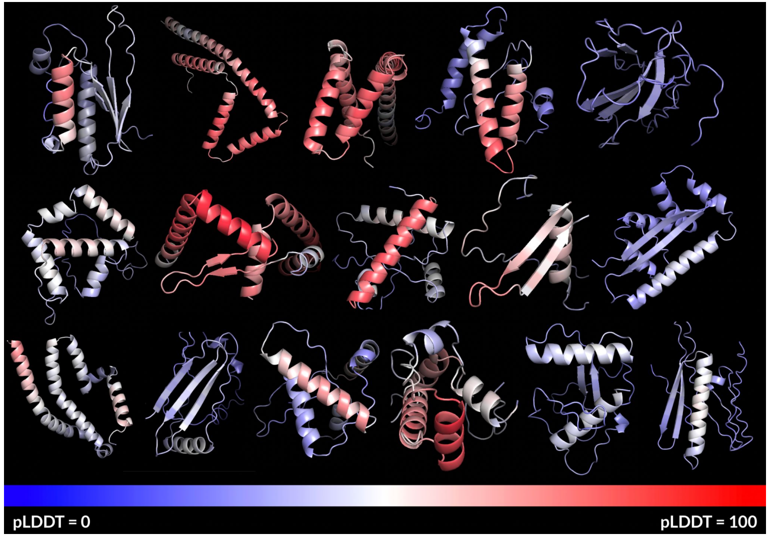


In order to avoid arbitrarily high-variance latent spaces, we experiment with two different kinds of regularizations. The first variant, *KL-reg.*, imposes a slight KL-penalty towards a standard normal on the learned latent, similar to a VAE [46, 69], whereas *VQ-reg.* uses a vector quantization layer [96] within the decoder. This model can be interpreted

Rombach et al. High-Resolution Image Synthesis with Latent Diffusion Models, CVPR 2022

Issues and hypotheses

- Latent space requires regularization
- Overcome $O(L^2)$ memory constraints and increase protein length to 512

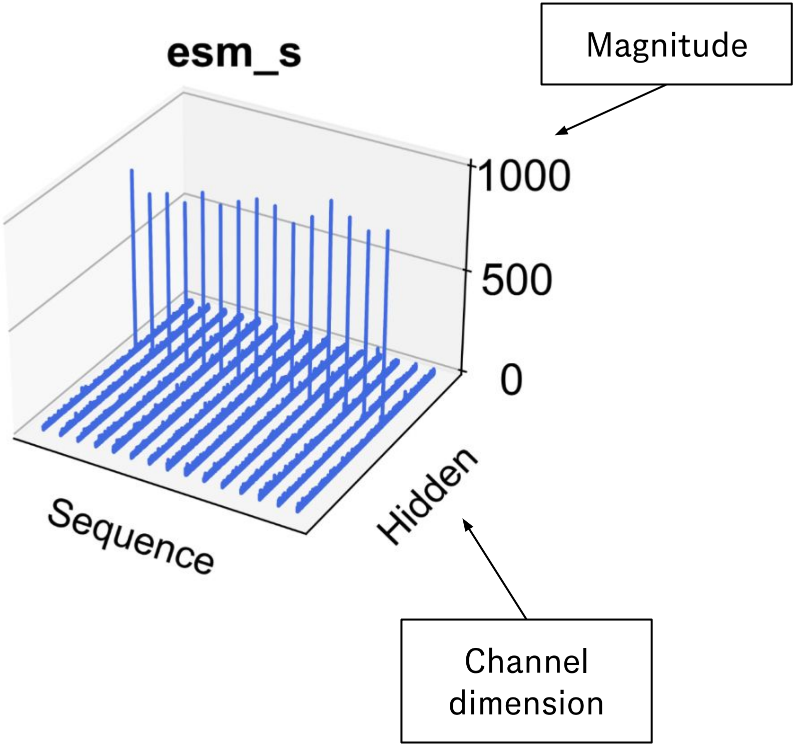


Issues and hypotheses

- Latent space requires regularization
- Overcome $O(L^2)$ memory constraints and increase protein length to 512
- Large latent space corresponds to **high-resolution** image generation
    - Rombach et al. latent space:
      HxWx4 = 64 x 64 x 4
    - Ours:
      Lx1024 = 512 x 1024

G. NCSN++ (Song et al., 2021) FFHQ-1024^2 Reference Samples

ESMFold latent space exhibits pathologically large values

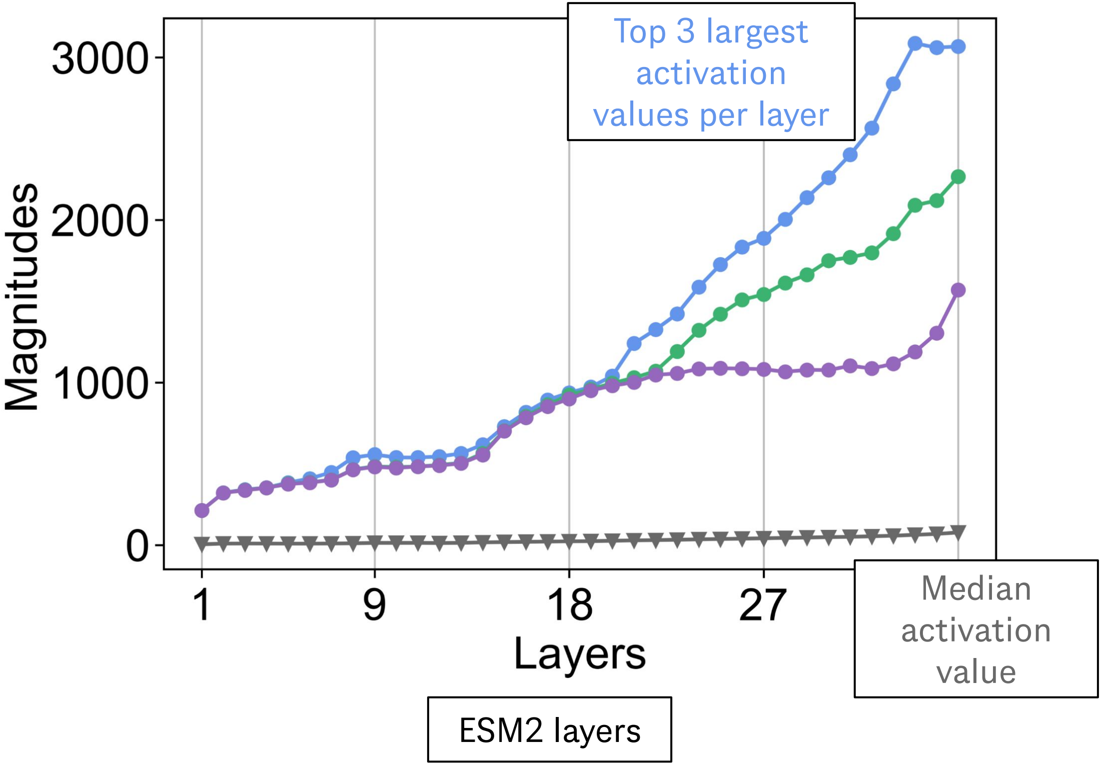


Magnitude

Channel dimension

Latent space will require regularization for diffusion to work.

~~ESMFold~~ ESM2 latent space exhibits pathologically large values

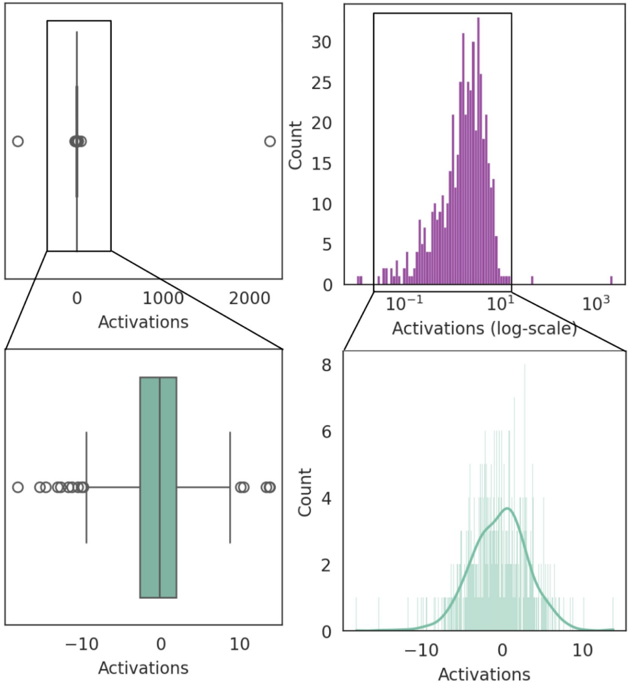


Top 3 largest activation values per layer

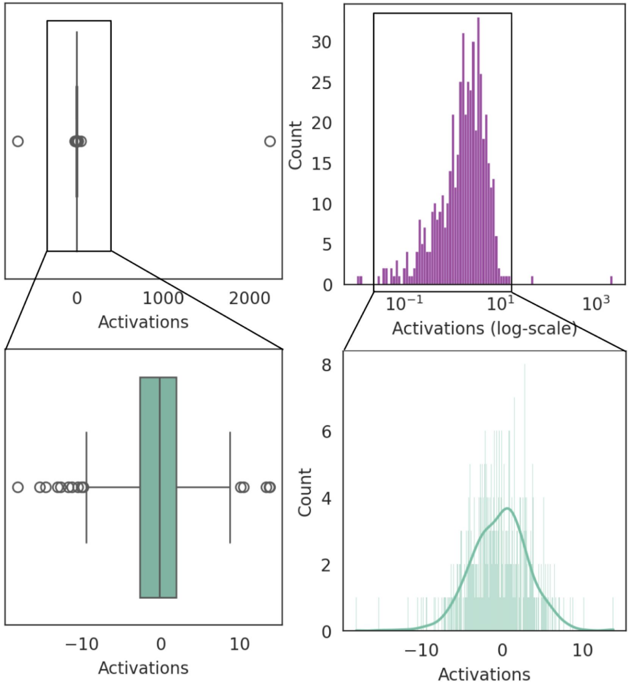
Median activation value

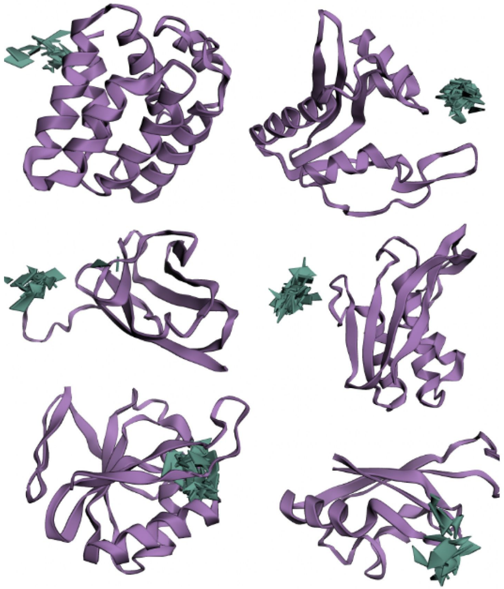
ESM2 layers

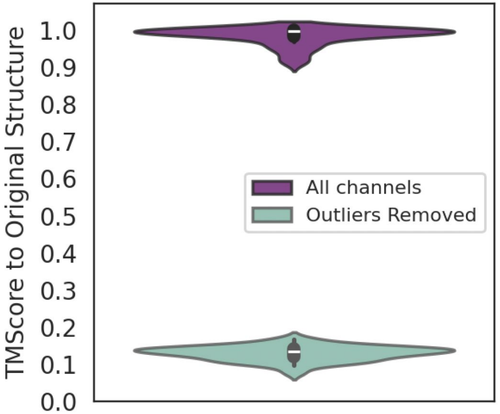
What if we just remove these wacky channels?

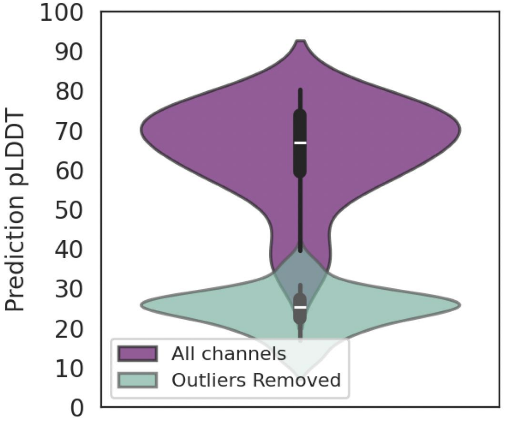


What if we just remove these wacky channels?

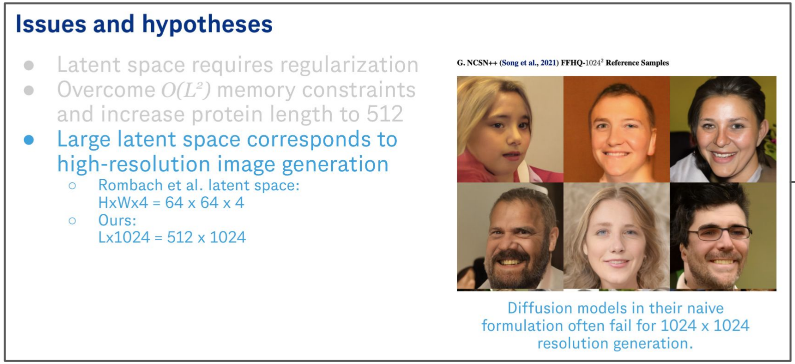








Addressing the hypotheses: embedding compression



Issues and hypotheses

- Latent space requires regularization
- Overcome $O(L^2)$ memory constraints and increase protein length to 512
- Large latent space corresponds to high-resolution image generation
  - Rombach et al. latent space: HxWx4 = 64 x 64 x 4
  - Ours: Lx1024 = 512 x 1024

G. NCSN++ (Song et al., 2021) FFHQ-1024^2 Reference Samples

Diffusion models in their naive formulation often fail for 1024 x 1024 resolution generation.

Since not all channels are necessary, can we compress the embedding?

Issues and hypotheses

- Latent space requires regularization
- Overcome $O(L^2)$ memory constraints and increase protein length to 512

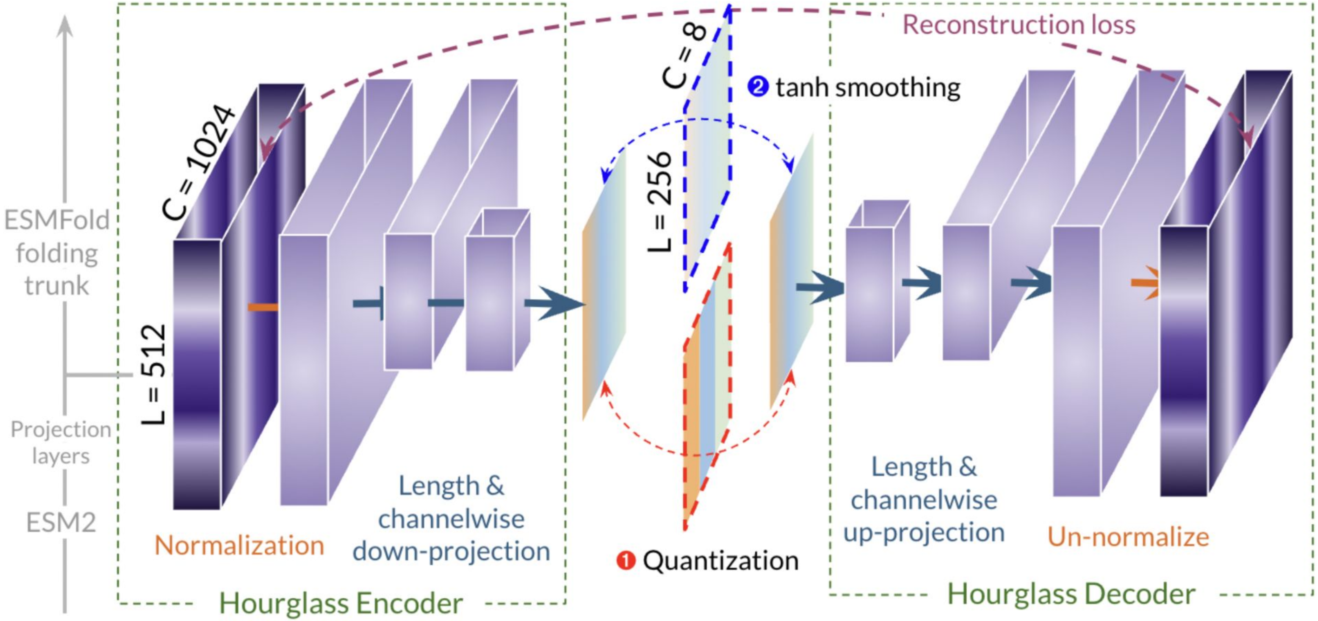
pLDDT = 0
pLDDT = 100

Can we also reduce the protein length?

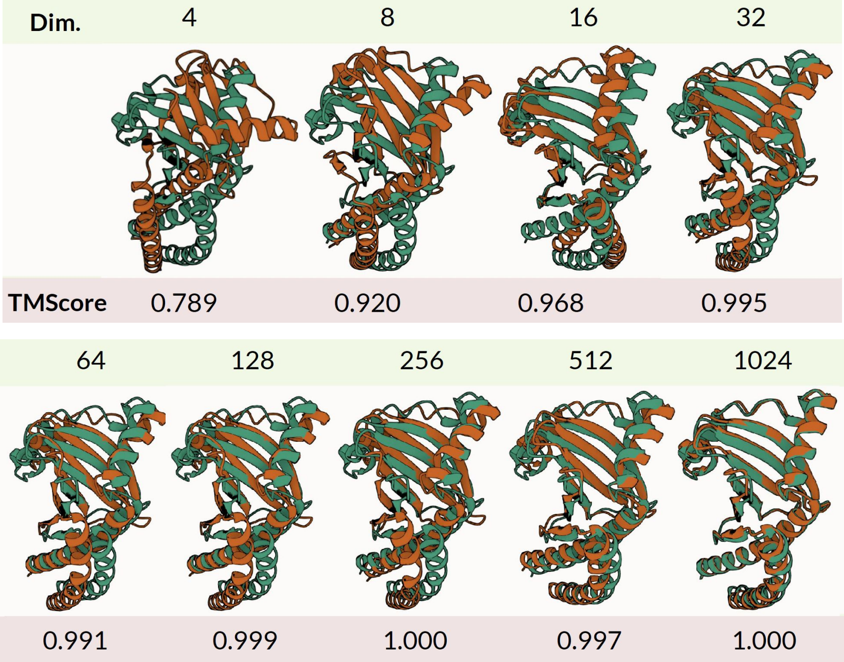
An autoencoder for protein embedding compression

Simple fix for massive activations: standardize each channel independently.

$$\mathbf{x}' = \frac{\mathbf{x} - \mathbf{x}_{\min}}{\mathbf{x}_{\max} - \mathbf{x}_{\min}}$$

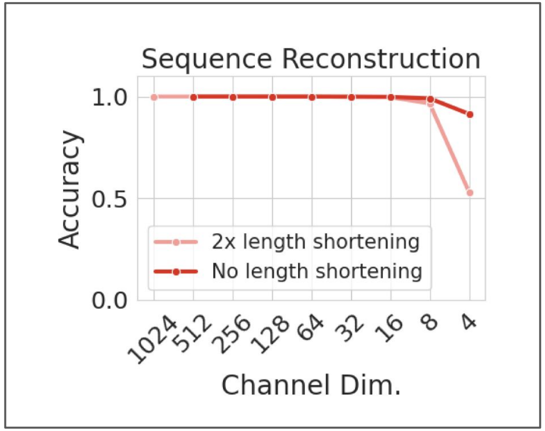


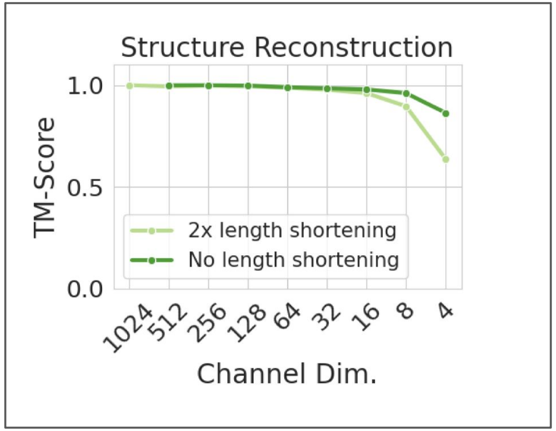
Turns out the latent space is highly compressible!



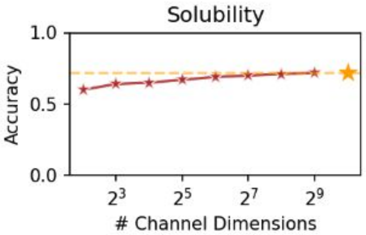
Dim.	4	8	16	32	64	128	256	512	1024
TMScore	0.789	0.920	0.968	0.995	0.991	0.999	1.000	0.997	1.000

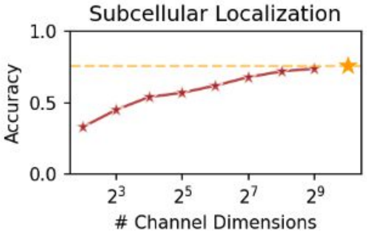
Turns out the latent space is highly compressible!





What about function information?

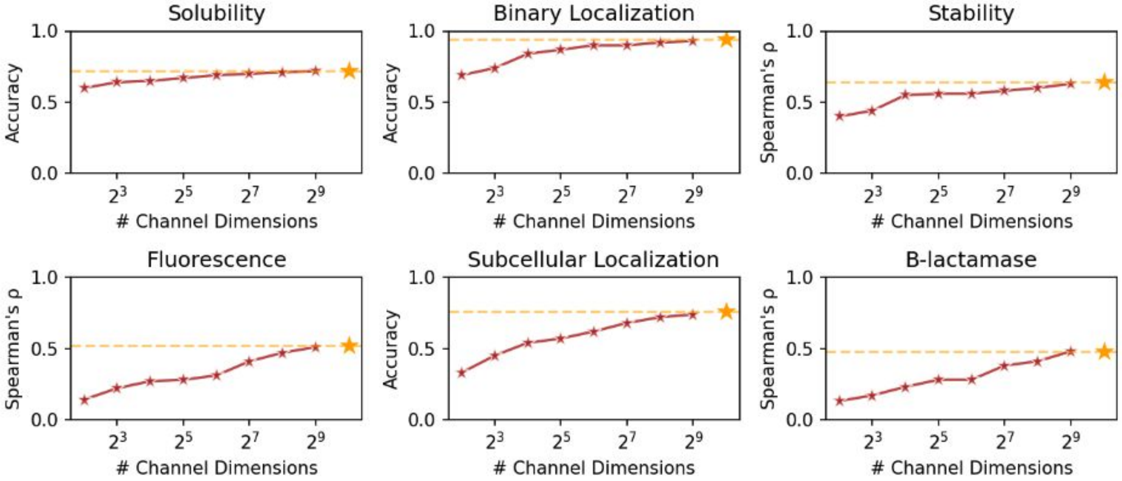




Performance degradation with compression is more gradual…

…for some functions.

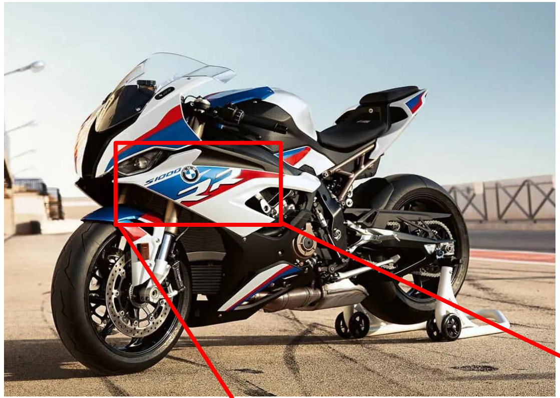
What about function information?

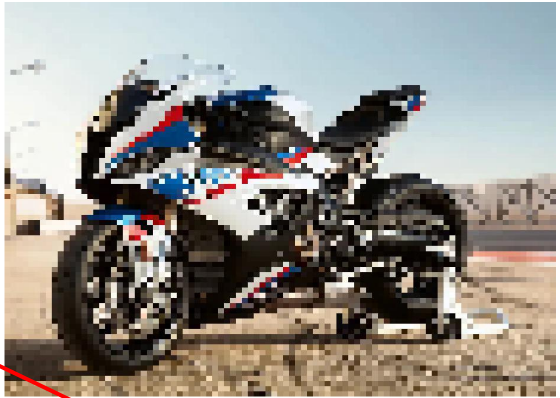


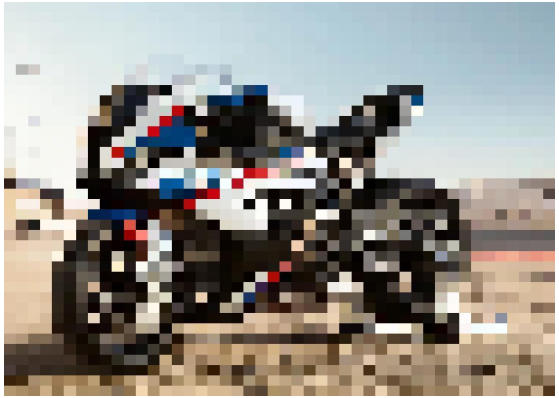
Performance degradation with compression is more gradual…

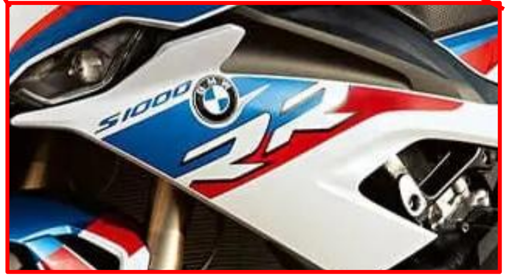
…for some functions.

Intuition: what is the speed of this motorcycle?



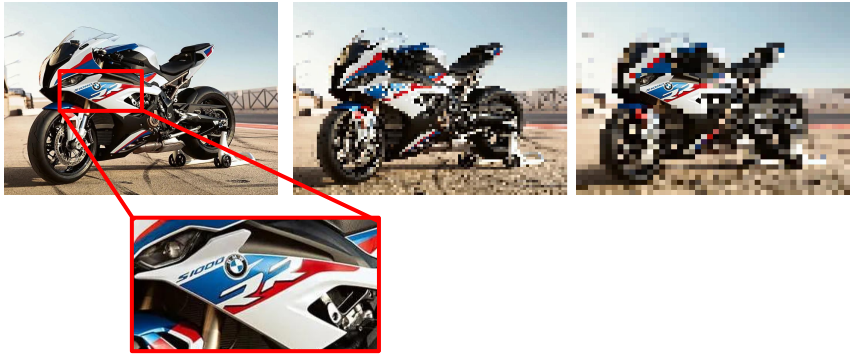




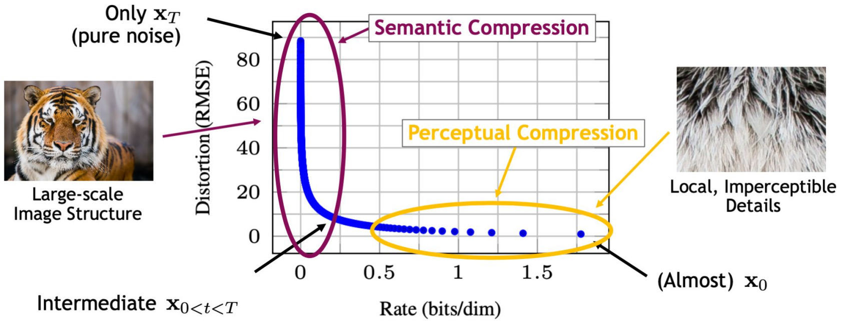


→ BMW S1000RR: 188 mph

Intuition: what is the speed of this motorcycle?

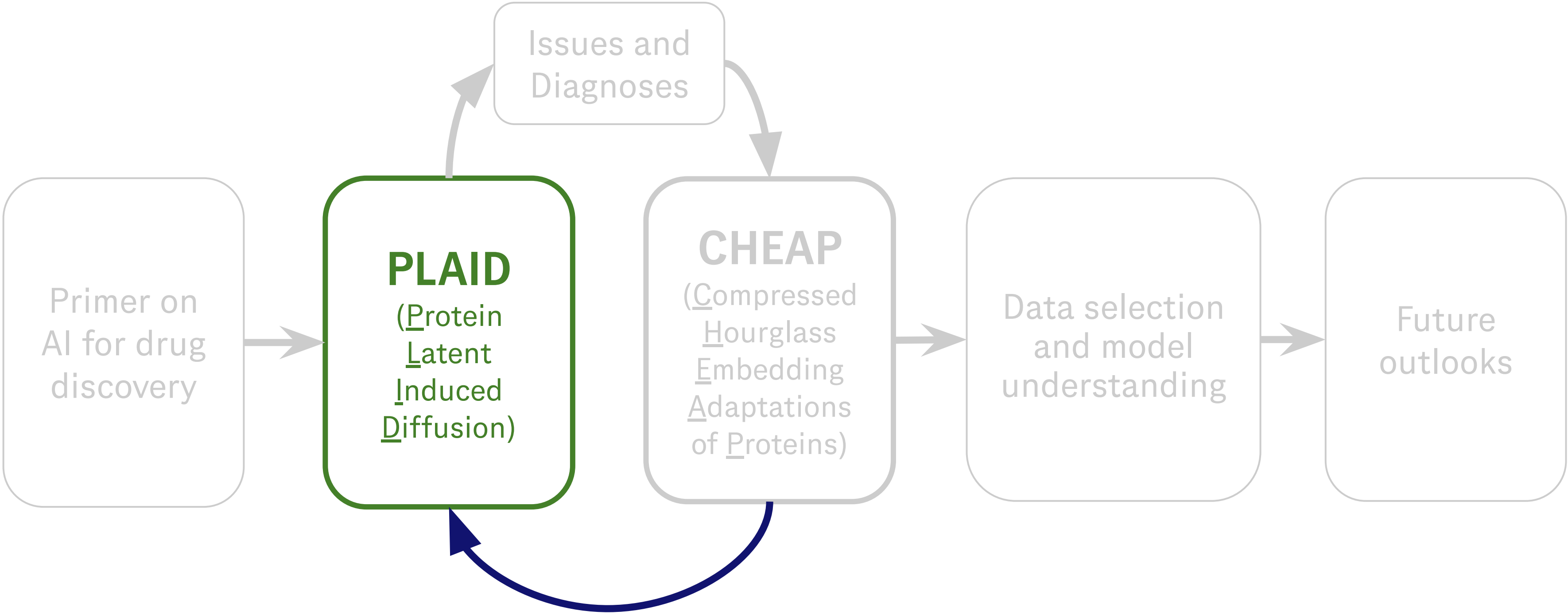


…what level of compression is optimal?



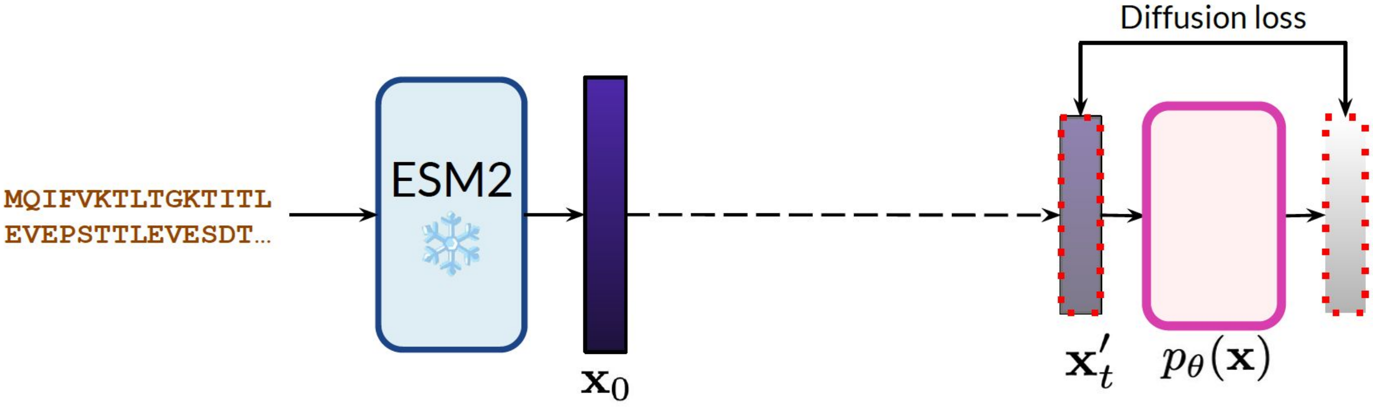
what constitutes semantic vs. perceptual compression for proteins? what level of detail do we need for drug discovery?

Agenda



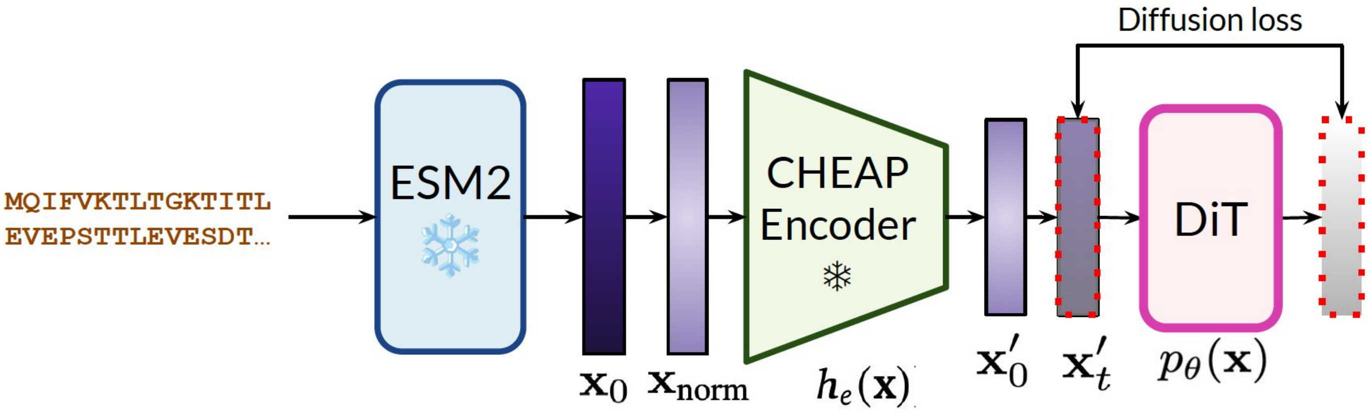
Primer on AI for drug discovery
PLAID (Protein Latent Induced Diffusion)
Issues and Diagnoses
CHEAP (Compressed Hourglass Embedding Adaptations of Proteins)
Data selection and model understanding
Future outlooks

Training the PLAID latent diffusion model…



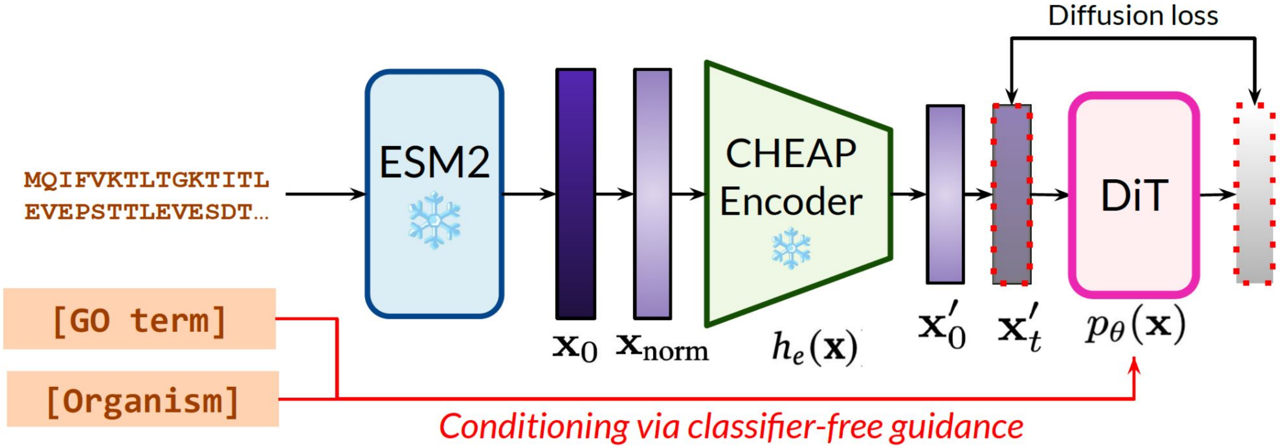
Dimensions:
L x 1024

…but add embedding compression with CHEAP



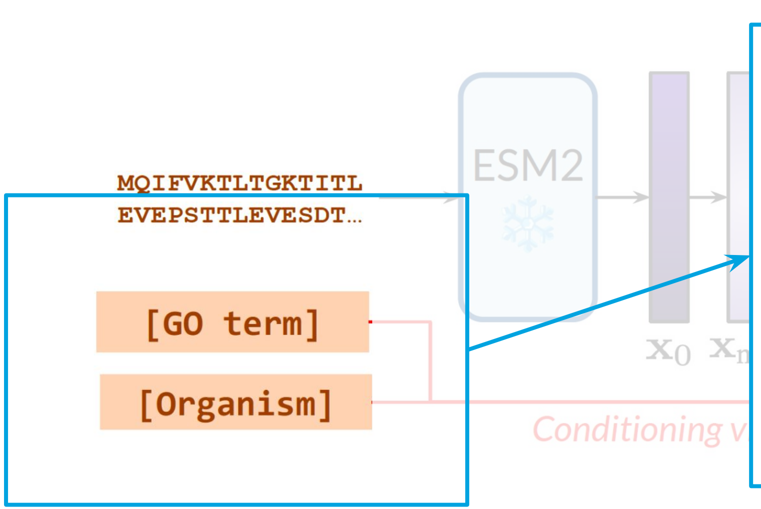
compress from **512x1024 -> 256x32**

Adding compositional function + taxonomic conditioning



Sequence databases have more sample-annotation pairs!

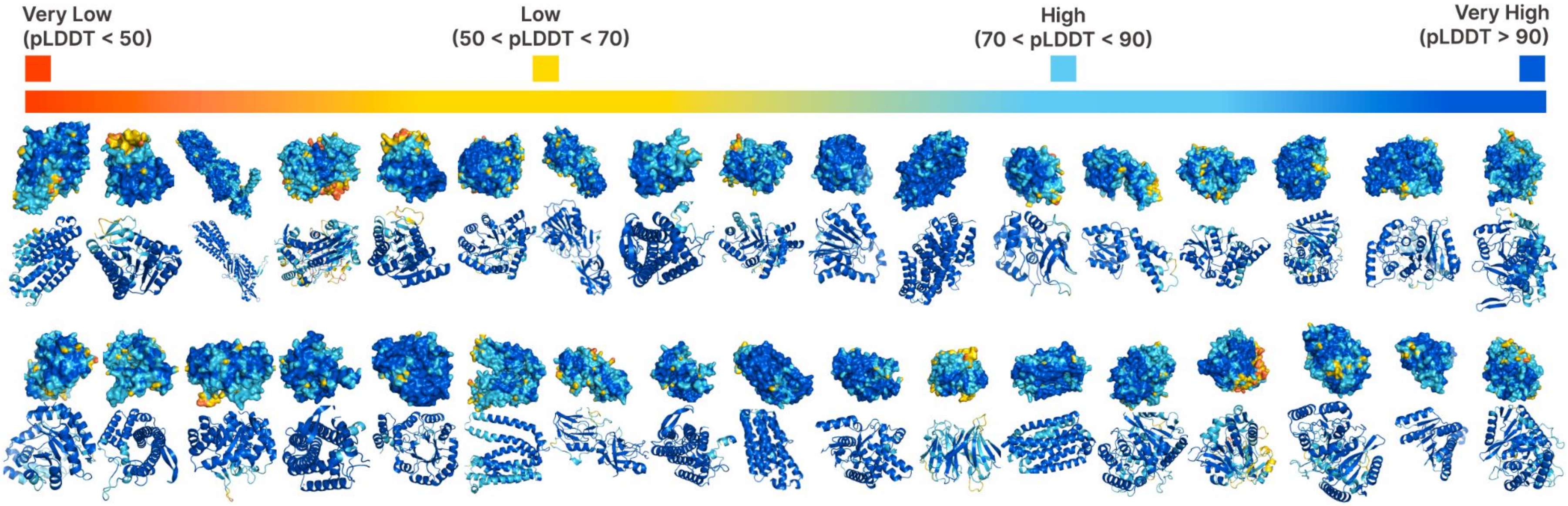
Adding compositional function + taxonomic conditioning



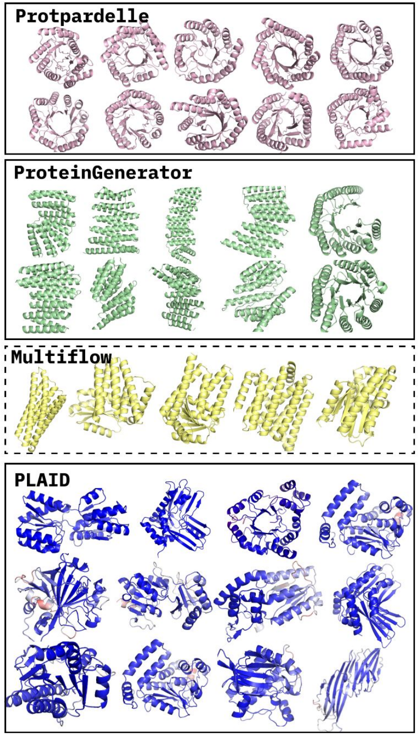
- towards text-controllable interface
  - ○ → **complex controllability** for drug discovery
- towards organism-specific conditioning
  - ○ → **towards controlling expression & humanization**

Sequence databases have more sample-annotation pairs!

PLAID unconditionally generates diverse all-atom structures



PLAID unconditionally generates diverse, high-quality folds



Protpardelle

```
>len600_samp97
AGGGGGGGGGGGGGGGGGGGGGGGGGLGLGLLLPPAGL...
>len600_samp98
PPPPGGAGGGGAAAALAGGSPGGPPGGGGGGGGGGGGG...
>len600_samp99
PPGPALPPSPGPGGVPPPPPLPPPPLPGGAPPAGGGLL...
```

ProteinGenerator

```
>len600_000097
GAAGLTAAAAVVGAAAAAGAAAAAALAAAAGAGAAAAA...
>len600_000098
AGAAGAAAAAAAAAAGAAAAGAGGGAGGAAAAAAAAAG...
>len600_000099
VAAAQAVQGAIAAAAALAATAALGLTAAGIAAPLLALV...
```

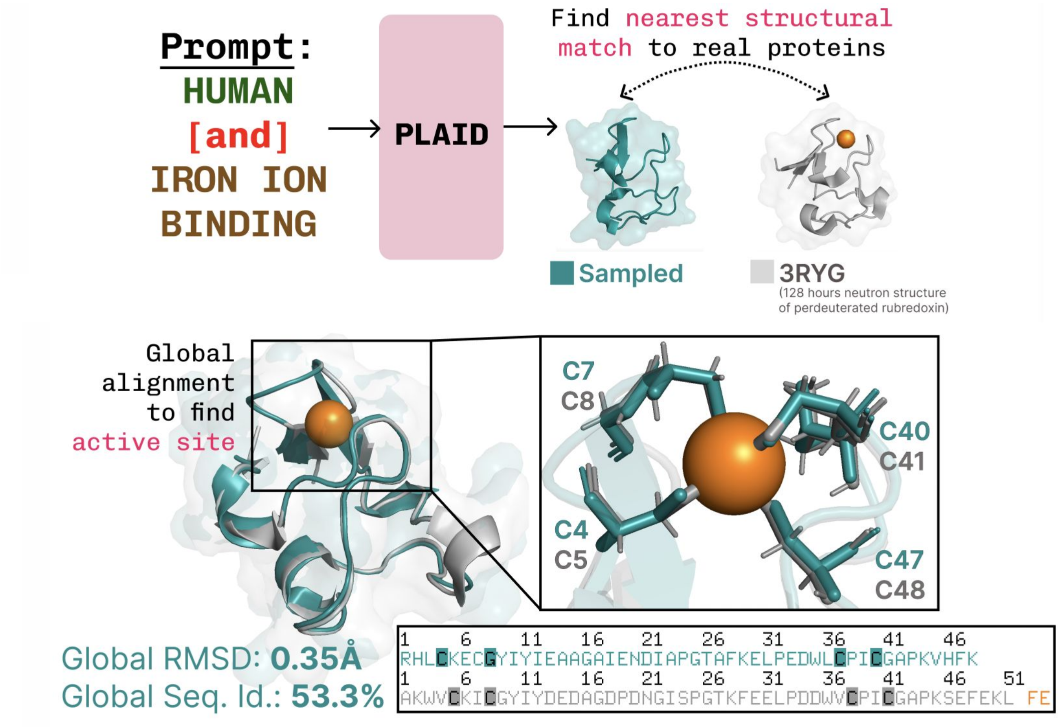
Multiflow

```
>len600_sample_97
LLGGLLGGLLGGAAGGAGAGAAAAGGGAVGVGVAGAVT...
>len600_sample_98
ADAATLTVGGGGTGGGGGAGGALGGAAAGGGGRVTLVV...
>len600_sample_99
AGGGAGLAGGAGGAGGAAAAAAAAAAAAAGAGGGAAAA...
```

PLAID

```
>len600_sample97
PDMGTVLGLAHSVGHLDFKTPDLSVADLETNLALLAAH...
>len600_sample98
FEMFDDKGGDLWERAASSGQLLIDVAYLANNGLRDGAT...
>len600_sample99
GNGGQARGTDDPLTHALQTLFQSAALDQSLQGDPENAV...
```

Function-prompted generations learn active site sidechains

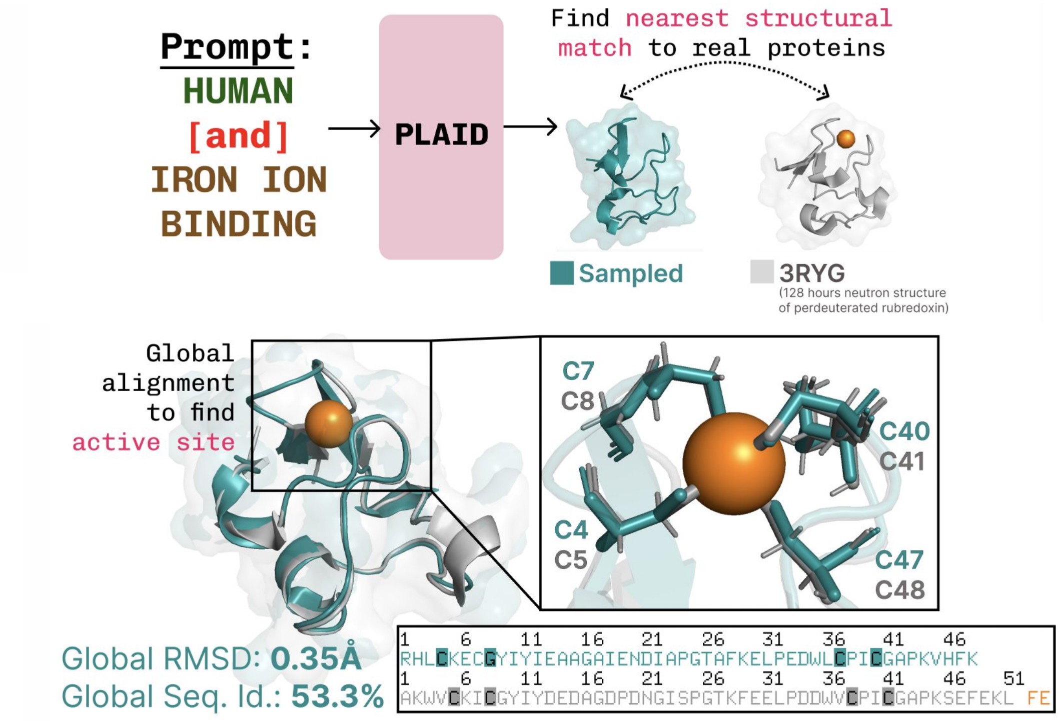

Prompt:
HUMAN
[and]
IRON ION
BINDING
PLAID
Find nearest structural match to real proteins
Sampled
3RYG
(128 hours neutron structure of perdeuterated rubredoxin)
Global alignment to find active site
C7
C8
C40
C41
C4
C5
C47
C48

Global RMSD: **0.35Å**
Global Seq. Id.: **53.3%**

```
1    6    11   16   21   26   31   36   41   46
RHLCKECGYIYIEAAGAIENDIAPGTAFKELPEDWLCPICGAPKVHFK
1    6    11   16   21   26   31   36   41   46   51
AKWVCKICGYIYDEDAGDPDNGISPGTKFEELPDDWVCPICGAPKSEFEKL FE
```

PLAID not only learns that cysteines coordinate the iron ion, but also the sidechain positioning…

Function-prompted generations learn active site sidechains

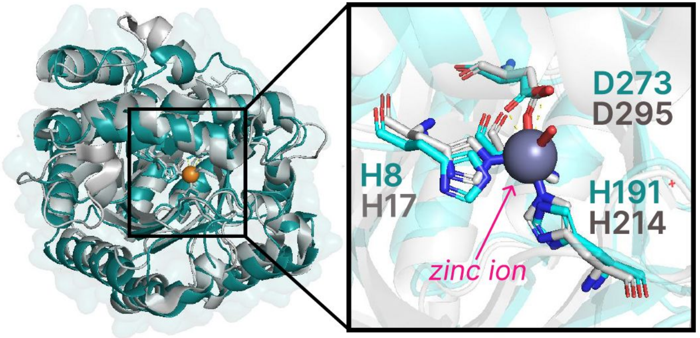

Prompt:
HUMAN
[and]
IRON ION
BINDING
PLAID
Find nearest structural match to real proteins
Sampled
3RYG
(128 hours neutron structure of perdeuterated rubredoxin)
Global alignment to find active site
C7
C8
C40
C41
C4
C5
C47
C48
Global RMSD: 0.35Å
Global Seq. Id.: 53.3%
1 6 11 16 21 26 31 36 41 46
RHLCKECGYIYIEAAGAIENDIAPGTAFKELPEDWLCPICGAPKVHFK
1 6 11 16 21 26 31 36 41 46 51
AKWVCKICGYIYDEDAGDPDNGISPGTKFEELPDDWVCPICGAPKSEFEKL FE

PLAID not only learns that cysteines coordinate the iron ion, but also the sidechain positioning…

Function-prompted generations learn active site sidechains

Prompt:

HUMAN [and] DEAMINASE ACTIVITY



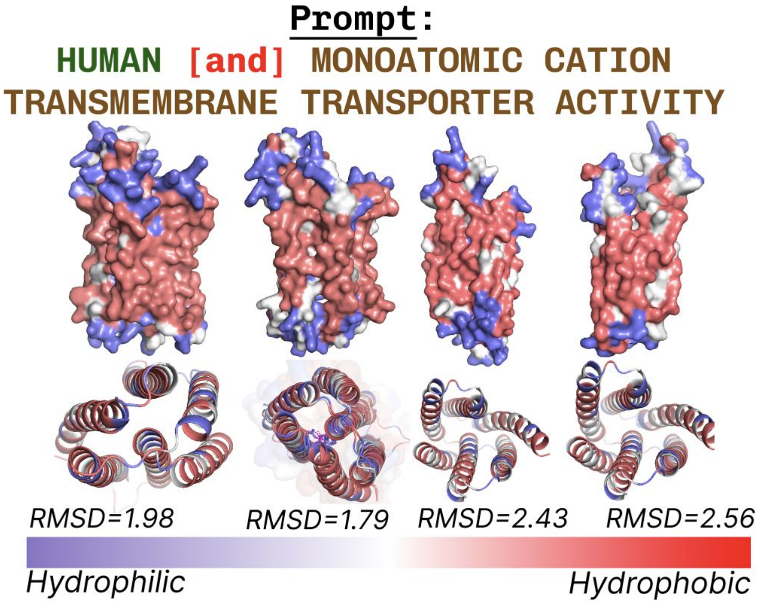
RMSD: **2.25Å**
Seq. Id.: **24.3%**

🟩 Sampled

⬜ 7RTG (Crystal Structure of the Human Adenosine Deaminase 1)

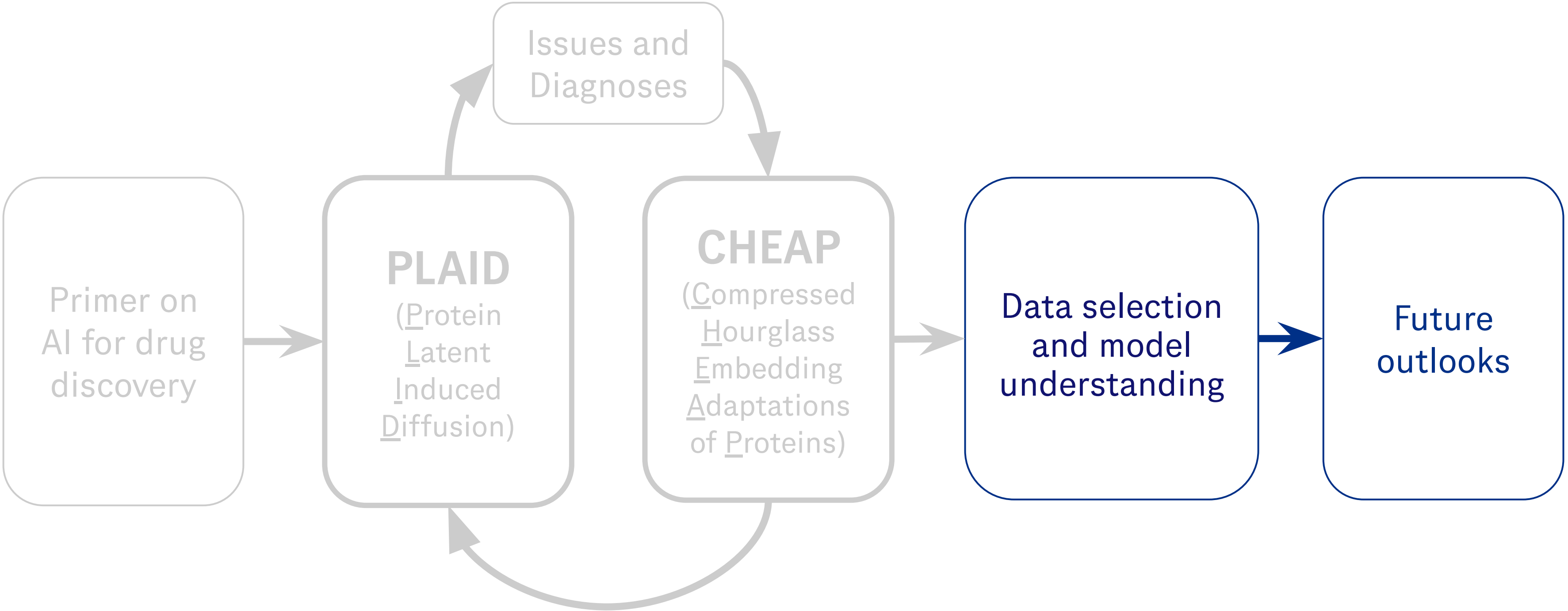
...despite these key residues not being adjacent in the sequence.

Transmembrane proteins exhibit expected hydrophobicity patterns



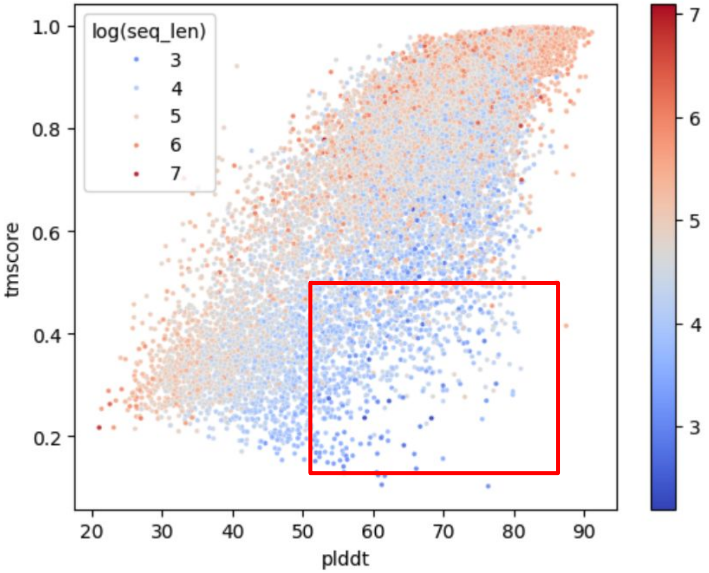
Hydrophobic residues are found at the core, as expected.

Agenda


Issues and Diagnoses
Primer on AI for drug discovery
PLAID
(Protein Latent Induced Diffusion)
CHEAP
(Compressed Hourglass Embedding Adaptations of Proteins)
Data selection and model understanding
Future outlooks

From proof-of-concept to deployment in AI for drug discovery

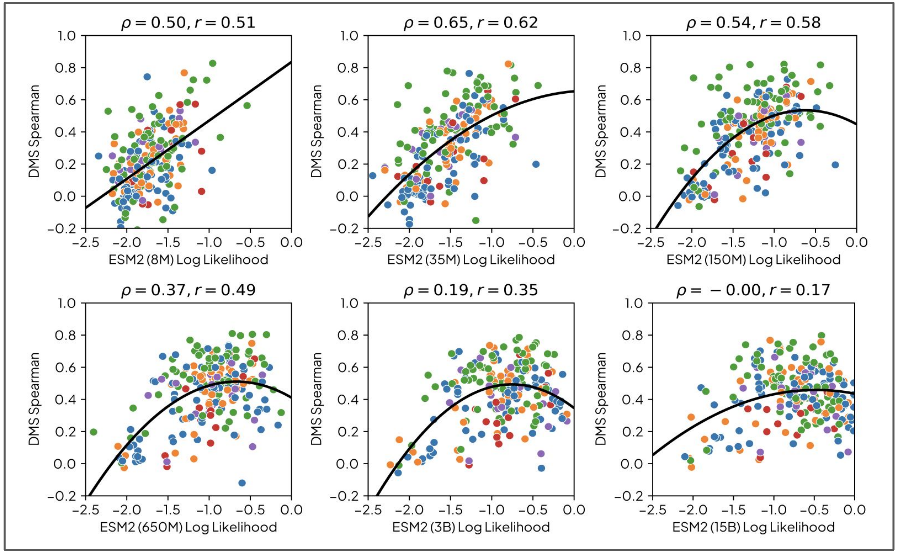
- Is the data learning a "biological world model", or artifacts of the training data?



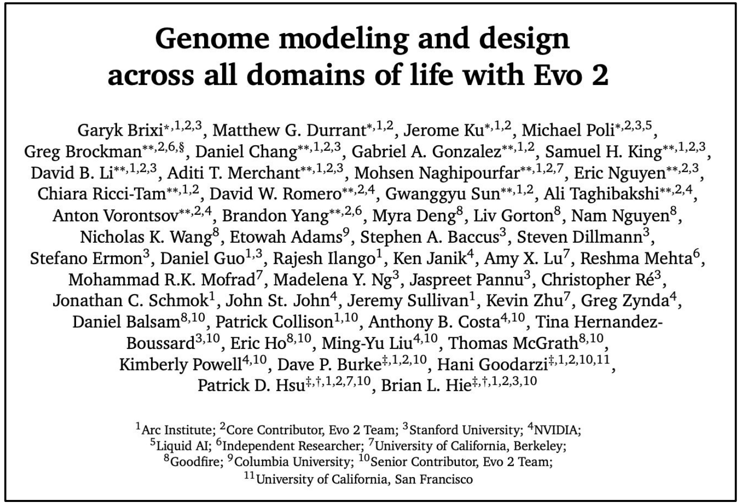
Length determines overconfident predictions, but we often use pLDDT for generative model evaluation.

From proof-of-concept to deployment in AI for drug discovery

- Is the data learning a “biological world model”, or artifacts of the training data?



From proof-of-concept to deployment in AI for drug discovery

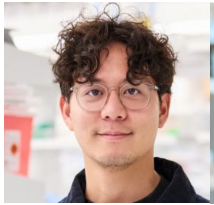


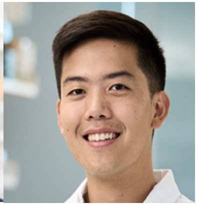
Genome modeling and design across all domains of life with Evo 2

Garyk Brixi[*,1,2,3], Matthew G. Durrant[*,1,2], Jerome Ku[*,1,2], Michael Poli[*,2,3,5], Greg Brockman[**,2,6,§], Daniel Chang[**,1,2,3], Gabriel A. Gonzalez[**,1,2], Samuel H. King[**,1,2,3], David B. Li[**,1,2,3], Aditi T. Merchant[**,1,2,3], Mohsen Naghipourfar[**,1,2,7], Eric Nguyen[**,2,3], Chiara Ricci-Tam[**,1,2], David W. Romero[**,2,4], Gwanggyu Sun[**,1,2], Ali Taghibakshi[**,2,4], Anton Vorontsov[**,2,4], Brandon Yang[**,2,6], Myra Deng[8], Liv Gorton[8], Nam Nguyen[8], Nicholas K. Wang[8], Etowah Adams[9], Stephen A. Baccus[3], Steven Dillmann[3], Stefano Ermon[3], Daniel Guo[1,3], Rajesh Ilango[1], Ken Janik[4], Amy X. Lu[7], Reshma Mehta[6], Mohammad R.K. Mofrad[7], Madelena Y. Ng[3], Jaspreet Pannu[3], Christopher Ré[3], Jonathan C. Schmok[1], John St. John[4], Jeremy Sullivan[1], Kevin Zhu[7], Greg Zynda[4], Daniel Balsam[8,10], Patrick Collison[1,10], Anthony B. Costa[4,10], Tina Hernandez-Boussard[3,10], Eric Ho[8,10], Ming-Yu Liu[4,10], Thomas McGrath[8,10], Kimberly Powell[4,10], Dave P. Burke[‡,1,2,10], Hani Goodarzi[‡,1,2,10,11], Patrick D. Hsu[‡,†,1,2,7,10], Brian L. Hie[‡,†,1,2,3,10]

[1]Arc Institute; [2]Core Contributor, Evo 2 Team; [3]Stanford University; [4]NVIDIA;
[5]Liquid AI; [6]Independent Researcher; [7]University of California, Berkeley;
[8]Goodfire; [9]Columbia University; [10]Senior Contributor, Evo 2 Team;
[11]University of California, San Francisco

...

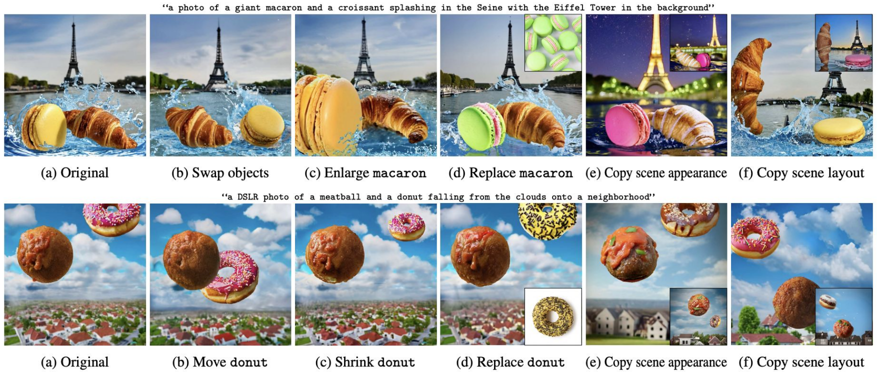




Medium-term directions…

Latent diffusion for drug design

- Leveraging **“fuzziness”** in # of atoms and binding position
- Alleviating computational challenges for large complexes with **compression**
- Semantic control in latent space



“a photo of a giant macaron and a croissant splashing in the Seine with the Eiffel Tower in the background”

(a) Original (b) Swap objects (c) Enlarge macaron (d) Replace macaron (e) Copy scene appearance (f) Copy scene layout

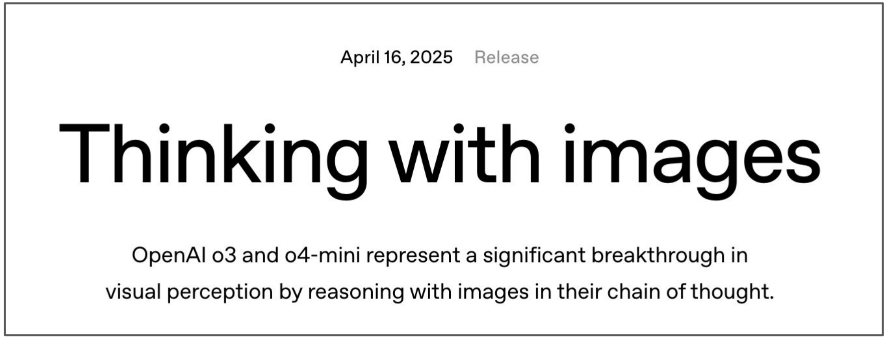
“a DSLR photo of a meatball and a donut falling from the clouds onto a neighborhood”

(a) Original (b) Move donut (c) Shrink donut (d) Replace donut (e) Copy scene appearance (f) Copy scene layout

Diffusion self-guidance for controllable image generation. Epstein et al., 2023

Medium-term directions…

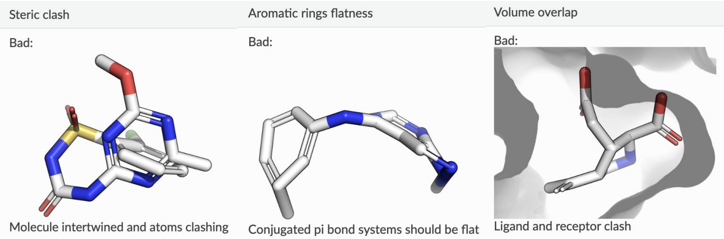
Multimodal biophysical reasoning / chain-of-thought “scratchpad”



April 16, 2025 Release

Thinking with images

OpenAI o3 and o4-mini represent a significant breakthrough in visual perception by reasoning with images in their chain of thought.



Steric clash

Bad:

Molecule intertwined and atoms clashing

Aromatic rings flatness

Bad:

Conjugated pi bond systems should be flat

Volume overlap

Bad:

Ligand and receptor clash

Example of how biochemical implausibilities can be reasoned through language.

Image source: PoseBusters documentation

Long term goals…

- How can we move to a "target-agnostic" paradigm in drug discovery using advances in task-agnostic AI systems?
    - AI for biology as reasoning about the molecular-level world

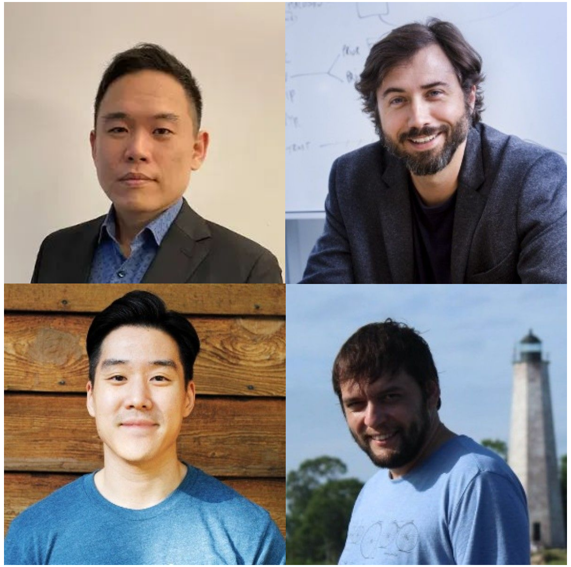
- How can we extrapolate / "reason" beyond human intelligence?
    - How does data availability & simulation fidelity affect how this is done?

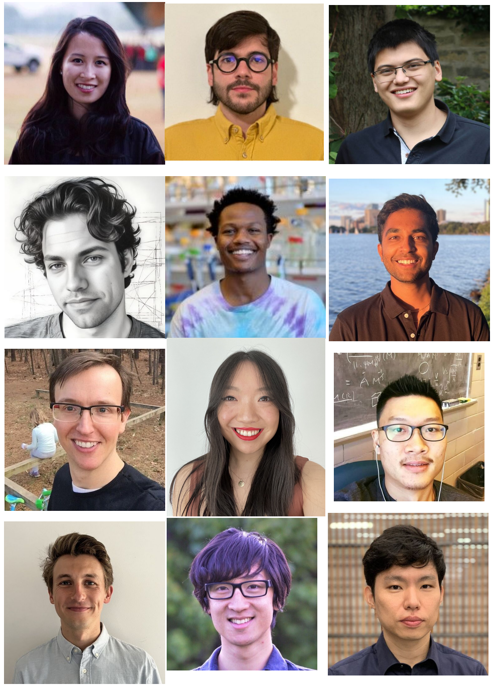
- How can we work *with* rather than *against* Moravec's Paradox, using scientific applications as a testbed?

Acknowledgements

Acknowledgements: Pieter!

Acknowledgements: Prescient Design / Genentech





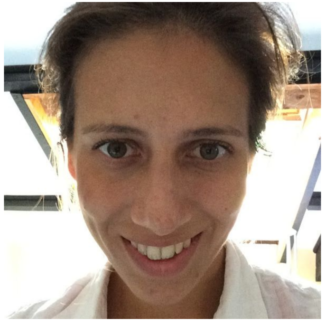
Prescient
Design
A Genentech Accelerator

and very many more Prescient team members!

Acknowledgements: MSR & Google Brain

Microsoft Research

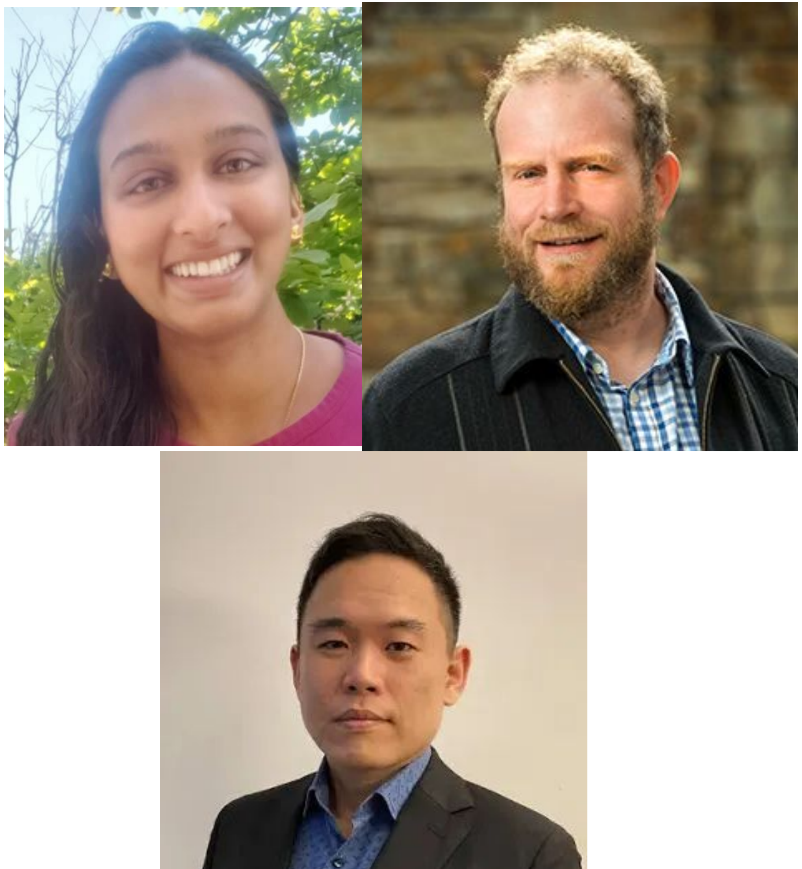
Google Research

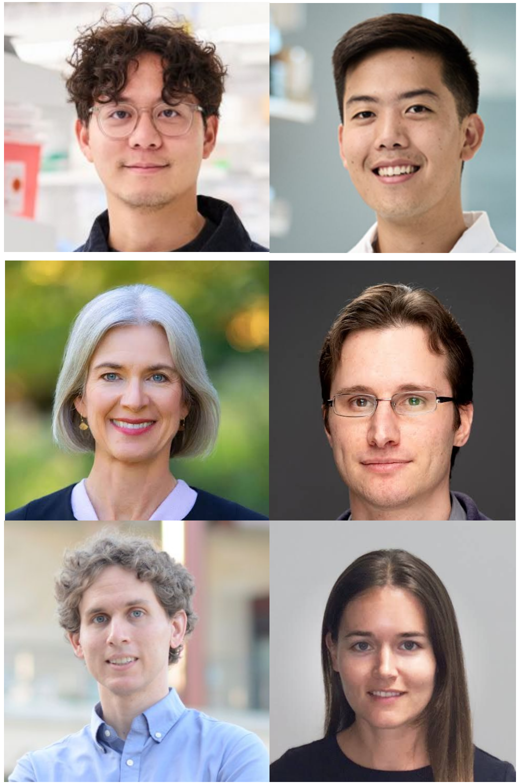


Acknowledgements: collaborators

and very many more!

Acknowledgements: committee & other faculty





Arc Institute

Innovative Genomics Institute

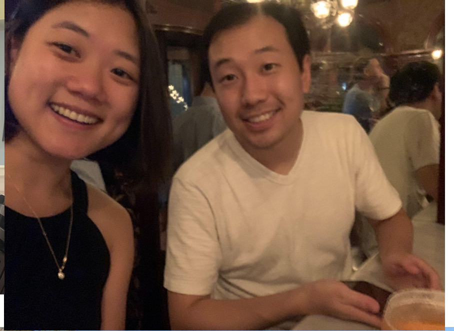
Acknowledgements: RLL labmates

Acknowledgements: admin

Acknowledgements: friends!! berkeley!!

Acknowledgements: friends!! berkeley!!

Acknowledgements: friends!! toronto/sf!!



Acknowledgements: family <3

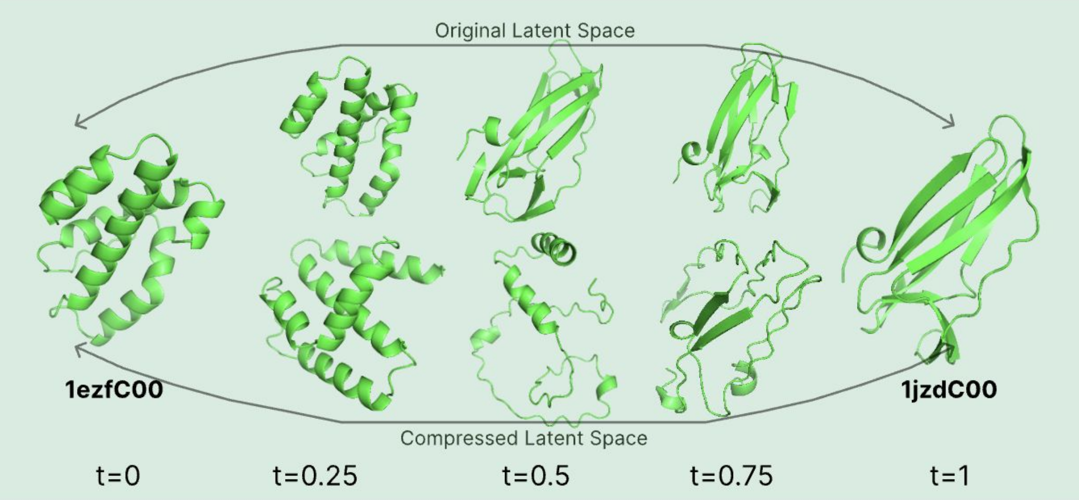
Acknowledgements: family <3

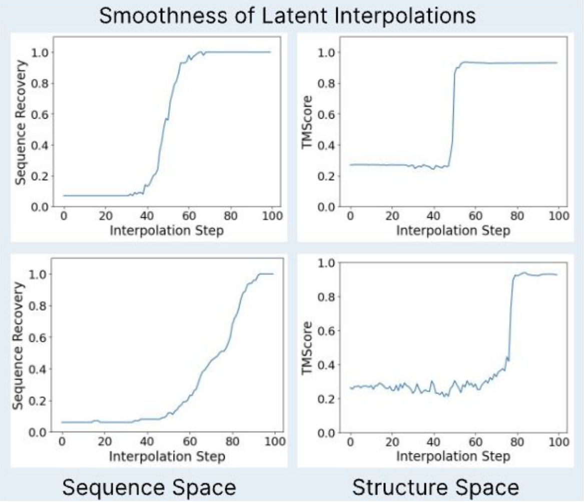
Acknowledgements: family <3

Acknowledgements

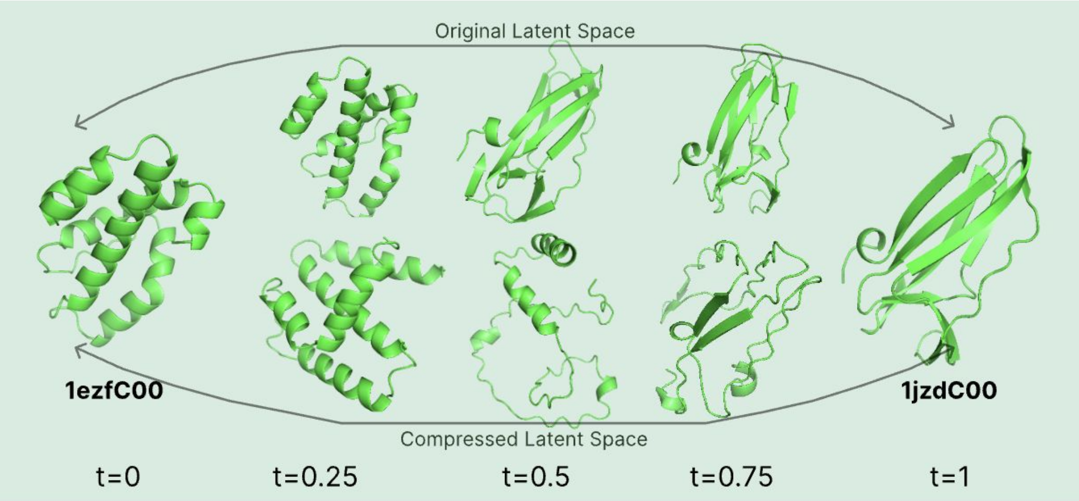
the end!!!! 🫶

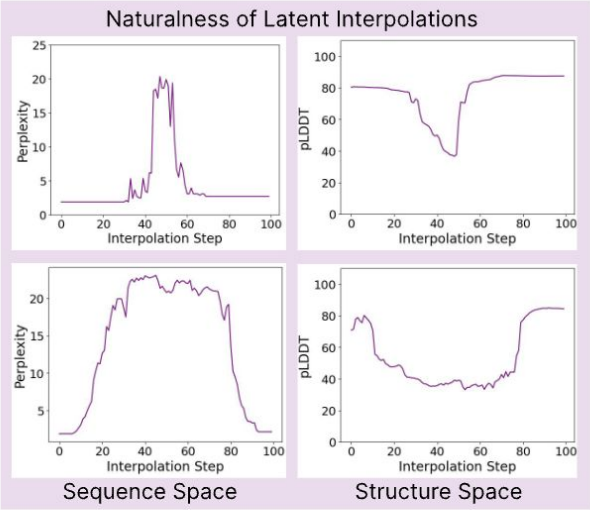
Linear interpolation in the latent space





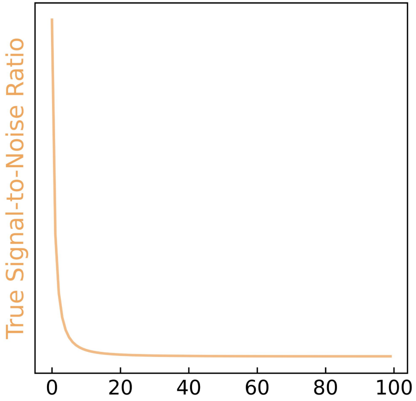
Linear interpolation in the latent space

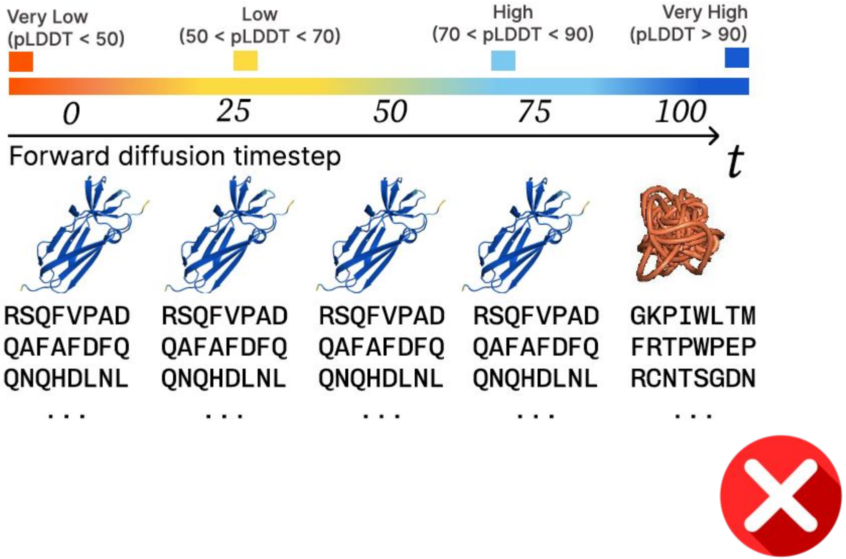




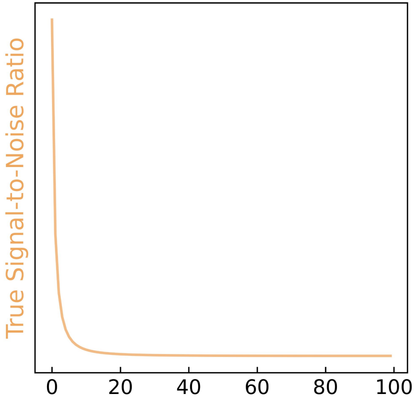
Protein language model latent spaces are less rugged than true fitness landscapes!

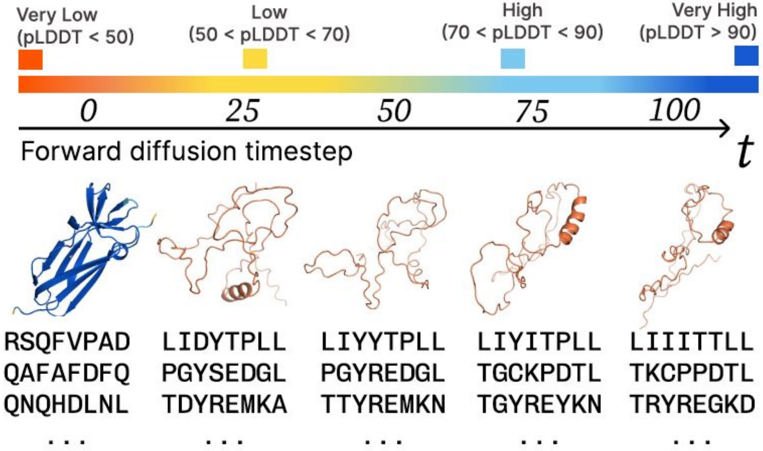
Noising the original latent space does not affect the structure…





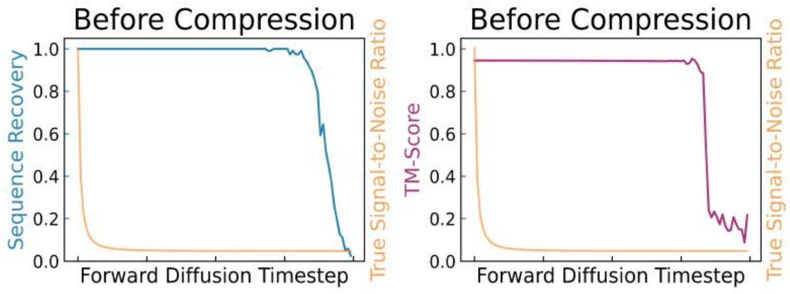
…noising the compressed latent space does map to corrupted structures



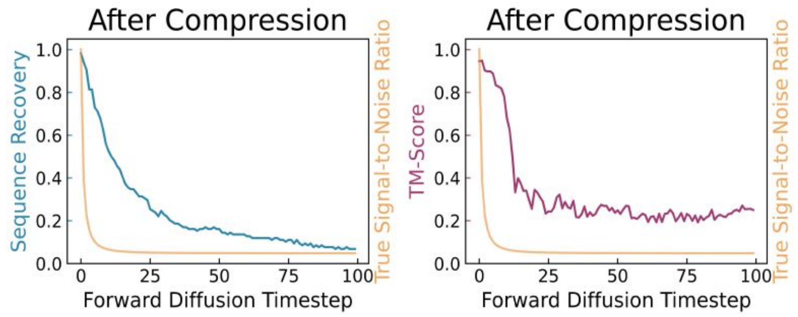


✓

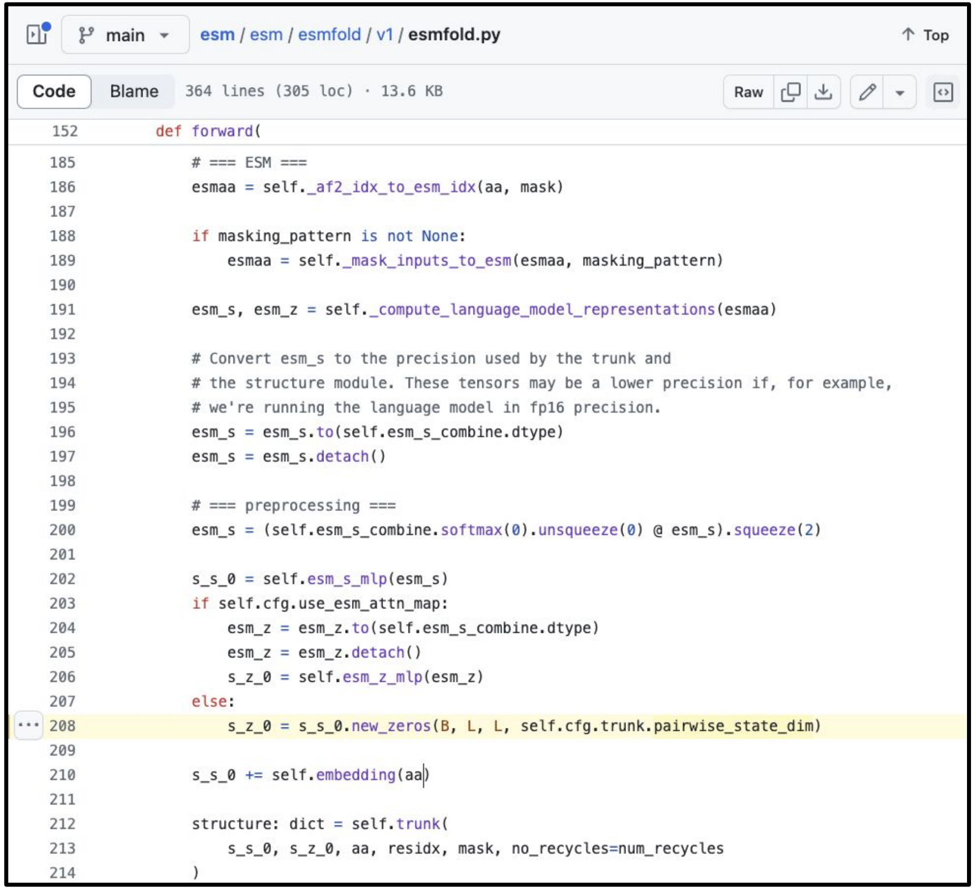
…noising the compressed latent space does map to corrupted structures



✗



✓

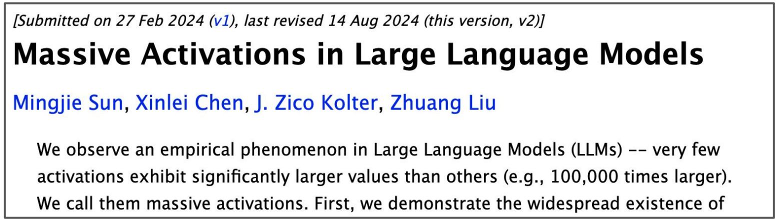

main
esm / esm / esmfold / v1 / esmfold.py
Top
Code
Blame
364 lines (305 loc) · 13.6 KB
Raw

```
    def forward(
        # === ESM ===
        esmaa = self._af2_idx_to_esm_idx(aa, mask)

        if masking_pattern is not None:
            esmaa = self._mask_inputs_to_esm(esmaa, masking_pattern)

        esm_s, esm_z = self._compute_language_model_representations(esmaa)

        # Convert esm_s to the precision used by the trunk and
        # the structure module. These tensors may be a lower precision if, for example,
        # we're running the language model in fp16 precision.
        esm_s = esm_s.to(self.esm_s_combine.dtype)
        esm_s = esm_s.detach()

        # === preprocessing ===
        esm_s = (self.esm_s_combine.softmax(0).unsqueeze(0) @ esm_s).squeeze(2)

        s_s_0 = self.esm_s_mlp(esm_s)
        if self.cfg.use_esm_attn_map:
            esm_z = esm_z.to(self.esm_s_combine.dtype)
            esm_z = esm_z.detach()
            s_z_0 = self.esm_z_mlp(esm_z)
        else:
            s_z_0 = s_s_0.new_zeros(B, L, L, self.cfg.trunk.pairwise_state_dim)

        s_s_0 += self.embedding(aa)

        structure: dict = self.trunk(
            s_s_0, s_z_0, aa, residx, mask, no_recycles=num_recycles
        )
```

Observation: at inference, the pairwise input is initialized as zeros…

~~ESMFold~~ ~~ESM2~~ Large transformers latent space exhibits pathologically large values

→ a pervasive issue across LLMs, ViTs, etc.

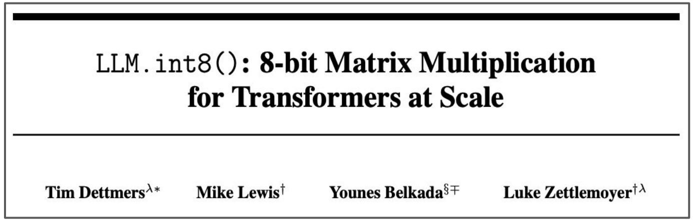


[Submitted on 27 Feb 2024 (v1), last revised 14 Aug 2024 (this version, v2)]

Massive Activations in Large Language Models

Mingjie Sun, Xinlei Chen, J. Zico Kolter, Zhuang Liu

We observe an empirical phenomenon in Large Language Models (LLMs) -- very few activations exhibit significantly larger values than others (e.g., 100,000 times larger). We call them massive activations. First, we demonstrate the widespread existence of



LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale

Tim Dettmers[λ*] Mike Lewis[†] Younes Belkada[§∓] Luke Zettlemoyer[†λ]

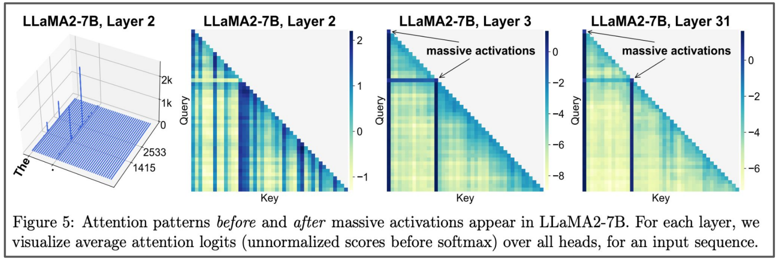
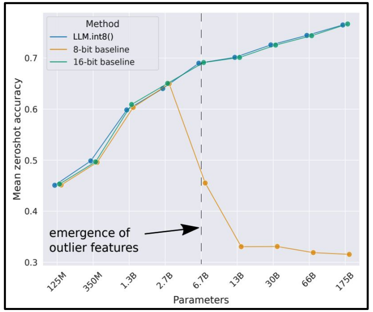
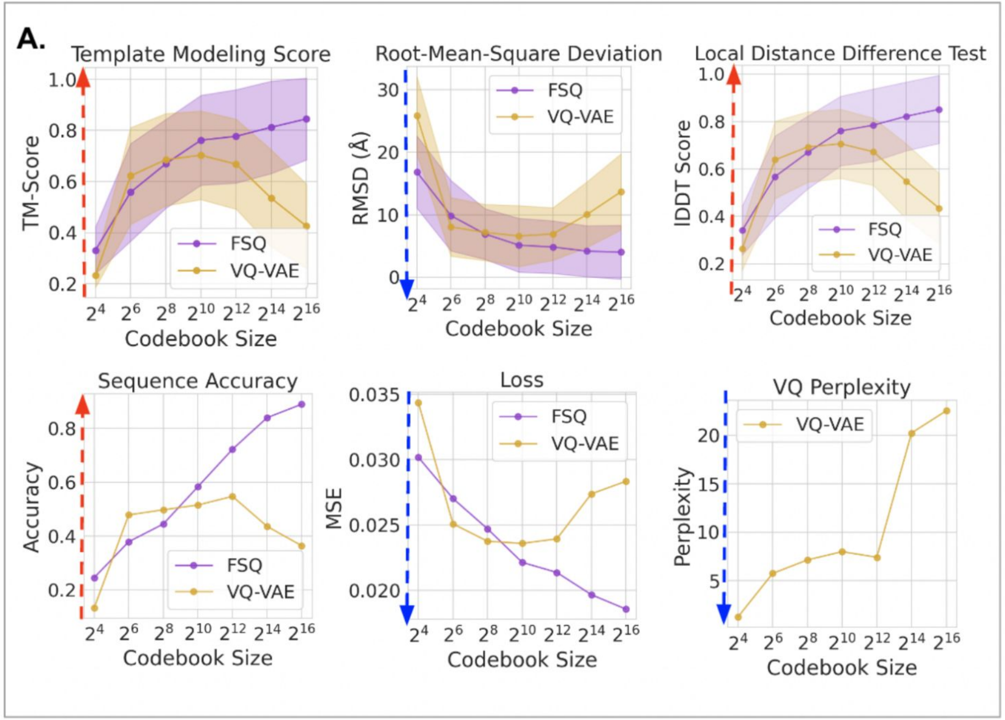


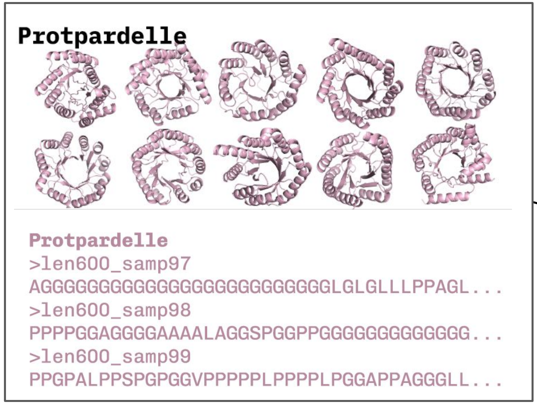
Figure 5: Attention patterns *before* and *after* massive activations appear in LLaMA2-7B. For each layer, we visualize average attention logits (unnormalized scores before softmax) over all heads, for an input sequence.



All-atom structural tokenizer, obtained from sequence alone

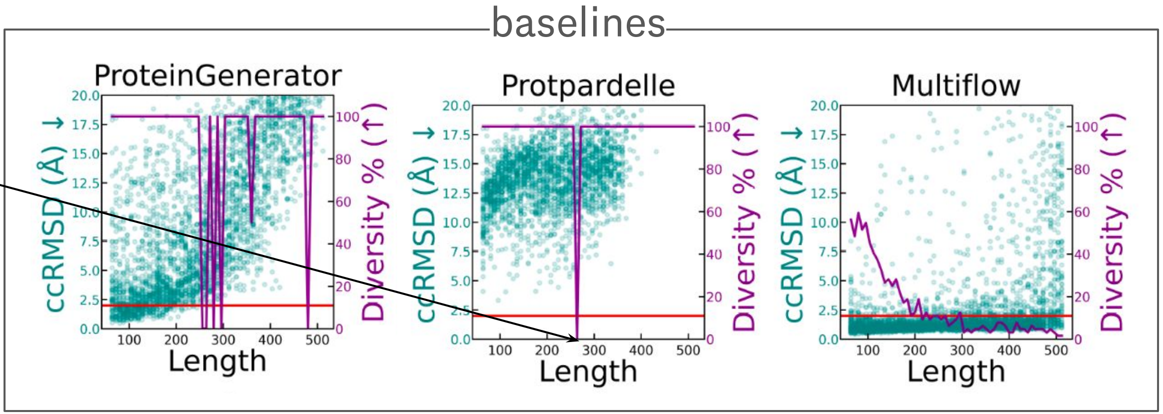


PLAID unconditionally generates diverse, high-quality folds



Protpardelle

Protpardelle
>len600_samp97
AGGGGGGGGGGGGGGGGGGGGGGGGGLGLGLLLPPAGL...
>len600_samp98
PPPPGGAGGGGAAAALAGGSPGGPPGGGGGGGGGGGGG...
>len600_samp99
PPGPALPPSPGPGGVPPPPPLPPPPLPGGAPPAGGGLL...

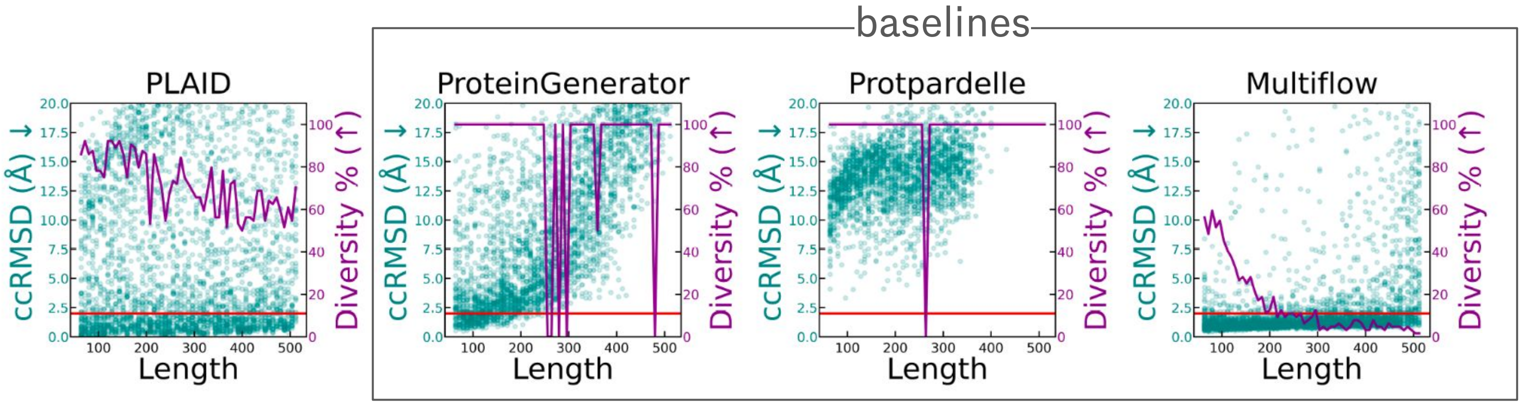


purple: diversity (↑)
(# of foldseek clusters / # of samples)

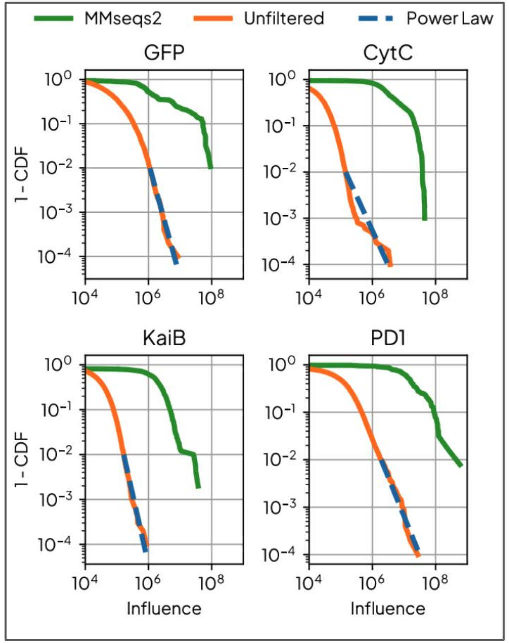
teal: quality (↓)
(ccRMSD between generated structure and predicted structure of generated sequence)

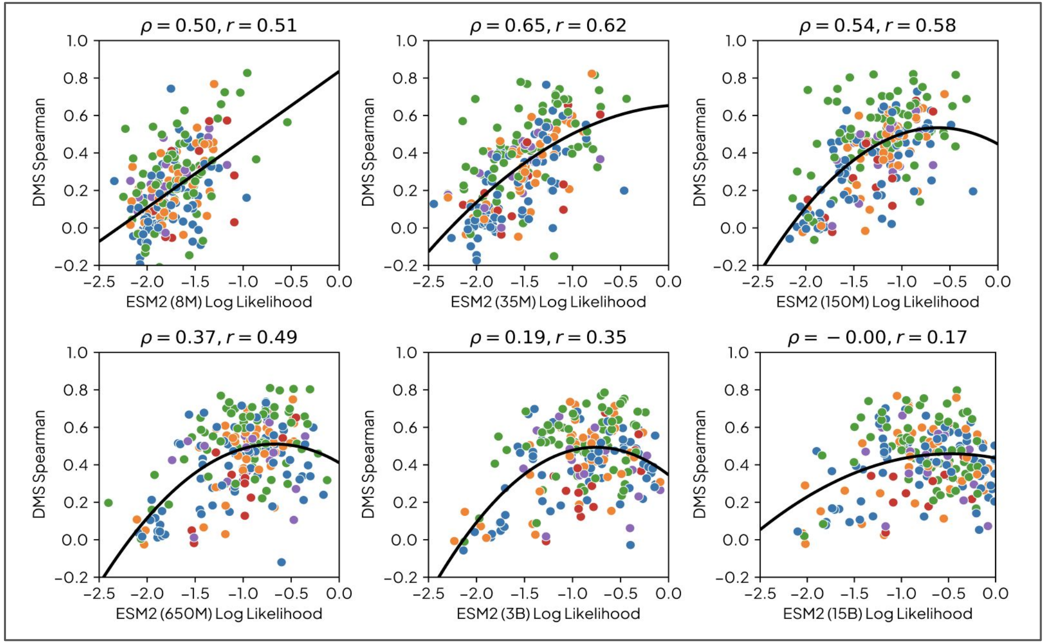
PLAID unconditionally generates diverse, high-quality folds

PLAID better balances diversity and quality, especially at longer sequence lengths.



From proof-of-concept to deployment in AI for drug discovery





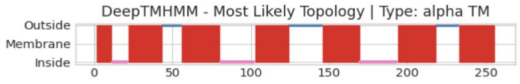
Transmembrane proteins exhibit expected numbers of helices

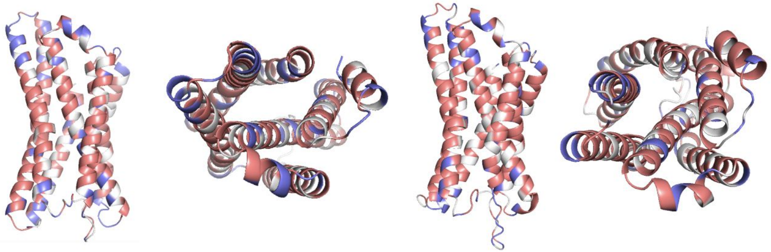
Prompt:

HUMAN [and]

G PROTEIN-COUPLED RECEPTOR ACTIVITY

```
GNVLVLIMMILKQREVKSMPNV
WVFNLALSDLLFLLSTPLLVVK
MSDTSWNLGLSPCKITTFLLFL
NLYSSVFFLACLSLDRYLTVRQ
        VRSN...
```





GPCRs have the expected 7-transmembrane topology, both when analyzing the sequence and structure.