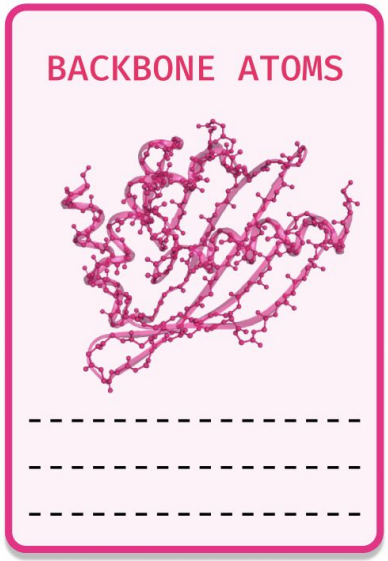
Genentech
A Member of the Roche Group

Generating All-Atom Protein Structure from Sequence-Only Training Data

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

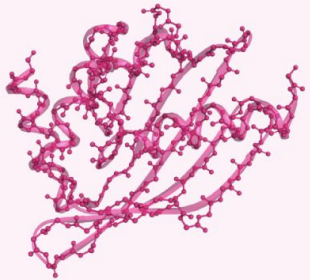
NeurIPS 2024 Workshop on Machine Learning for Structural Biology (MLSB)
Paper: bit.ly/plaid-proteins

Problem: **Existing protein generation methods are often backbone-only**



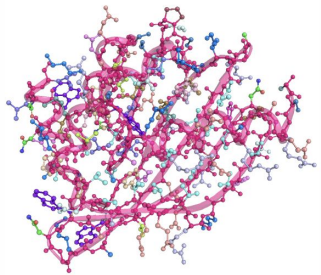
Problem: **Sidechain atom generation requires knowing the sequence**

BACKBONE ATOMS



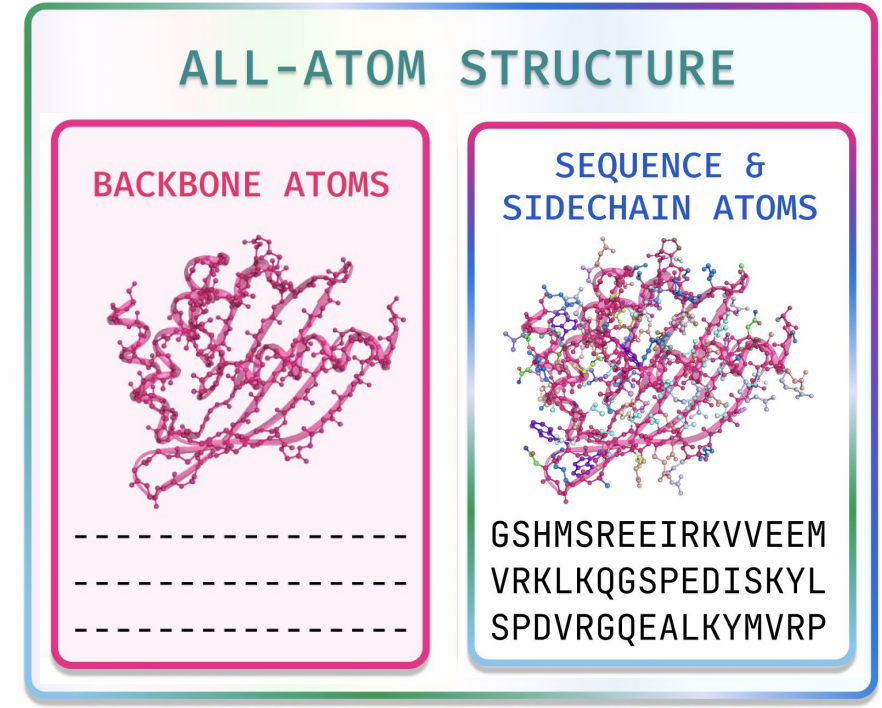
```
----------------
----------------
----------------
```

SEQUENCE & SIDECHAIN ATOMS

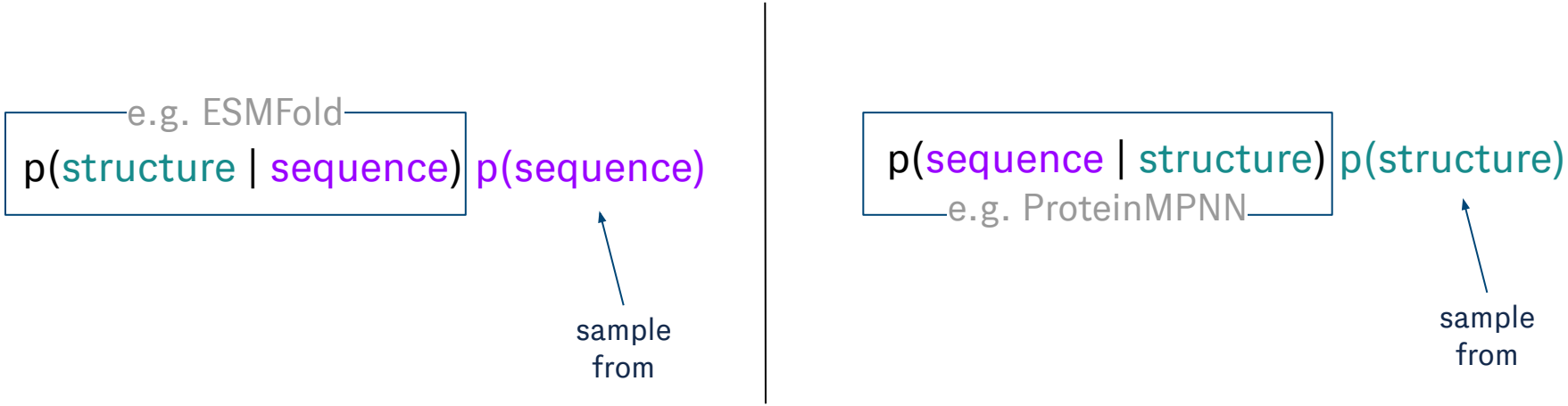


```
GSHMSREEIRKVVEEM
VRKLKQGSPEDISKYL
SPDVRGQEALKYMVRP
```

Problem: **All-atom generation requires multimodal generation**



Problem: **Existing all-atom generation often sample from the marginal rather than joint distribution**



Problem: **Existing all-atom generation often sample from the marginal rather than joint distribution**

e.g. ESMFold

p(structure | sequence) p(sequence)

sample from

p(sequence | structure) p(structure)

e.g. ProteinMPNN

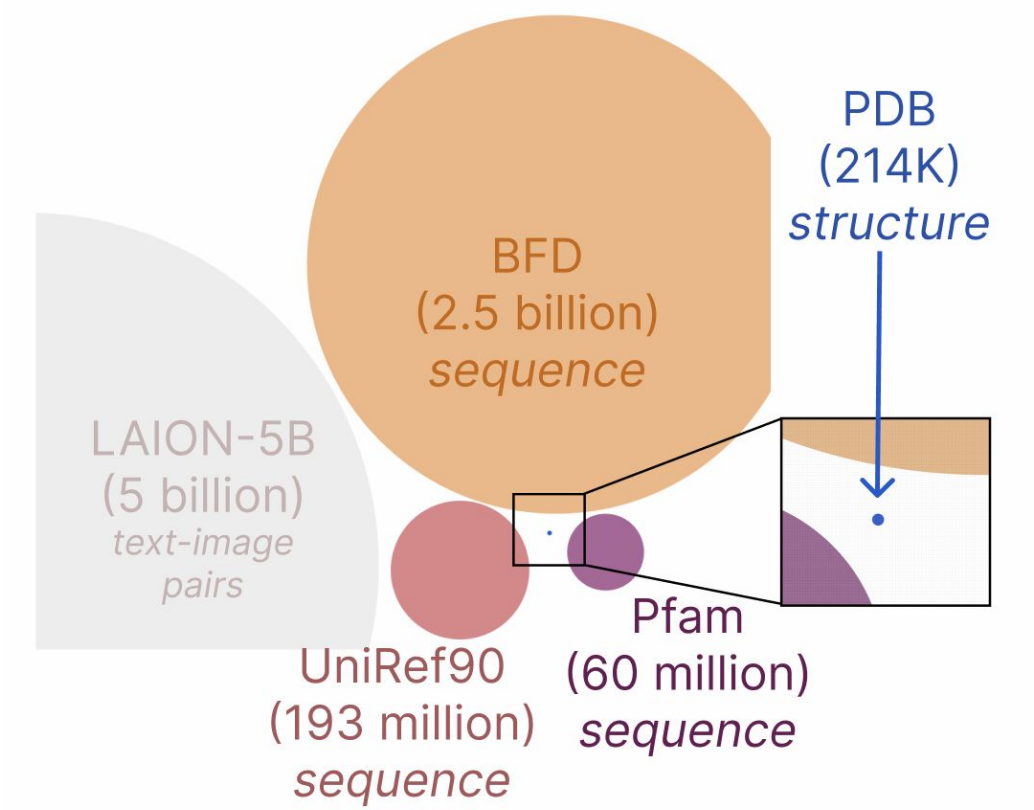
sample from

Goal: p(sequence, structure)

sample from

Genentech
A Member of the Roche Group

Problem: **Structure data is less abundant and annotated than sequences**

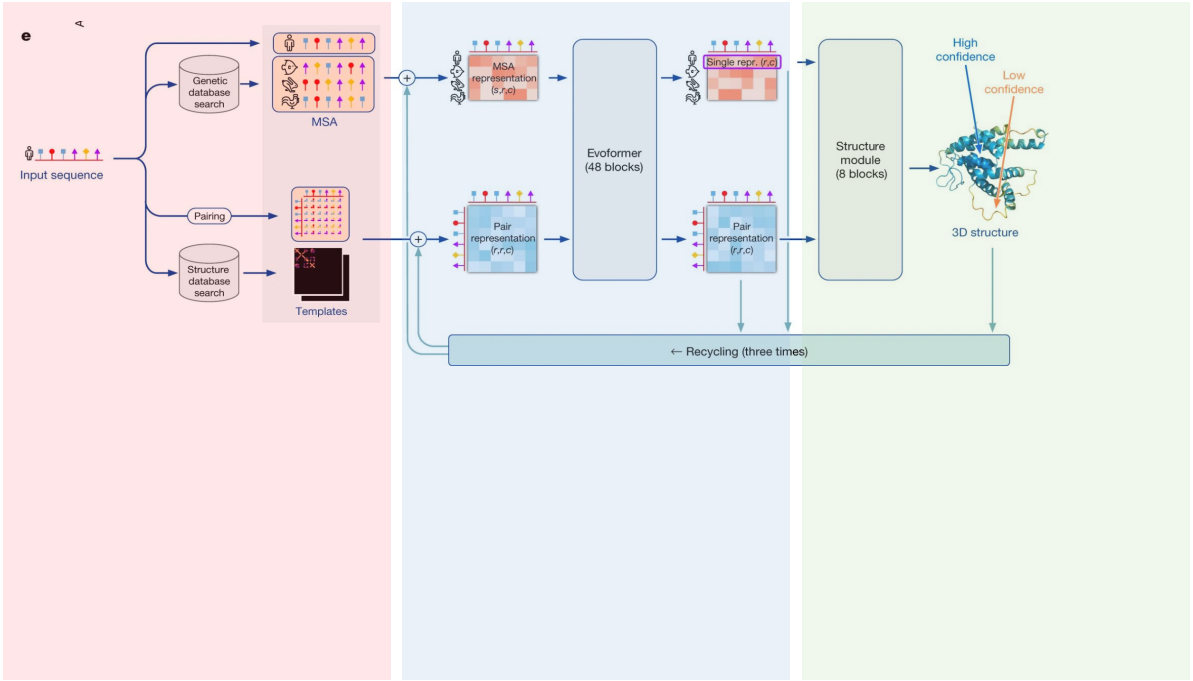


How can we sample from the joint distribution p(sequence, structure) for all-atom generation?

Refresher: **AlphaFold2 for sequence-to-structure prediction**

AlphaFold2:

Uses an explicit retrieval step



harness additional sequence-based priors

learn structural features from sequence latents

generate structures

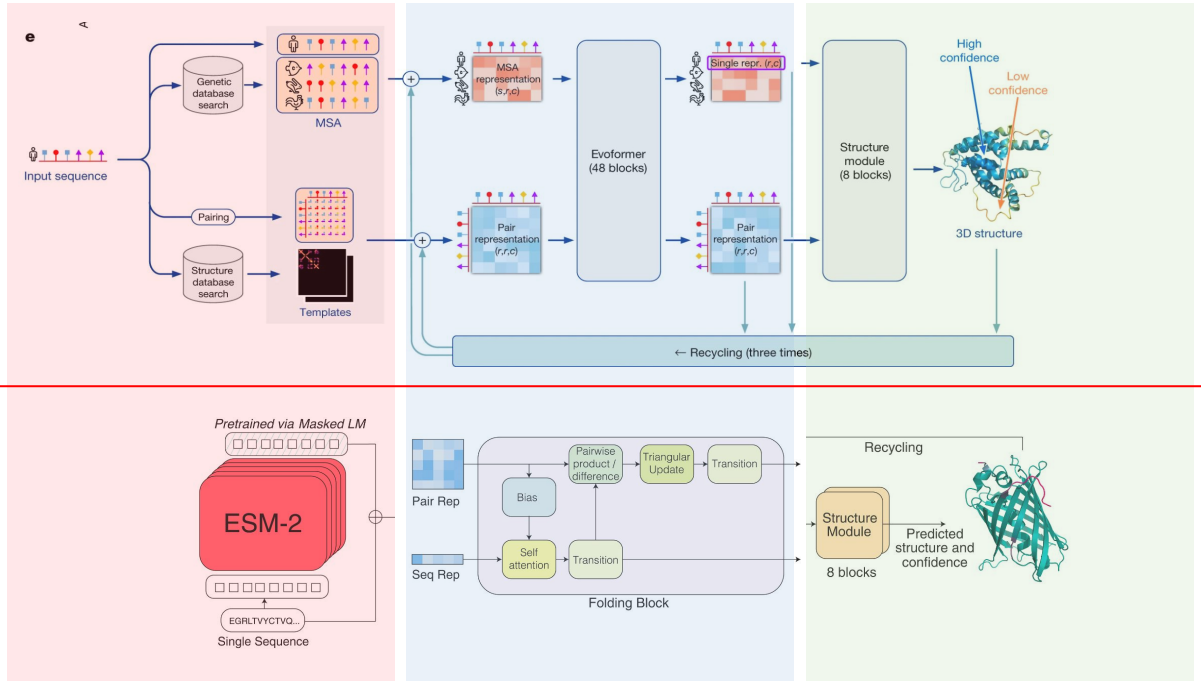
Refresher: ESMFold for sequence-to-structure prediction

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

Genentech
A Member of the Roche Group

main esm / esm / esmfold / v1 / esmfold.py ↑ Top

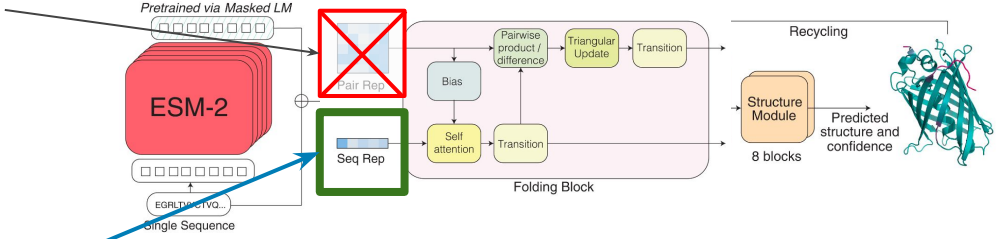
Code Blame 364 lines (305 loc) · 13.6 KB Raw

```
152     def forward(
185         # === ESM ===
186         esmaa = self._af2_idx_to_esm_idx(aa, mask)
187
188         if masking_pattern is not None:
189             esmaa = self._mask_inputs_to_esm(esmaa, masking_pattern)
190
191         esm_s, esm_z = self._compute_language_model_representations(esmaa)
192
193         # Convert esm_s to the precision used by the trunk and
194         # the structure module. These tensors may be a lower precision if, for example,
195         # we're running the language model in fp16 precision.
196         esm_s = esm_s.to(self.esm_s_combine.dtype)
197         esm_s = esm_s.detach()
198
199         # === preprocessing ===
200         esm_s = (self.esm_s_combine.softmax(0).unsqueeze(0) @ esm_s).squeeze(2)
201
202         s_s_0 = self.esm_s_mlp(esm_s)
203         if self.cfg.use_esm_attn_map:
204             esm_z = esm_z.to(self.esm_s_combine.dtype)
205             esm_z = esm_z.detach()
206             s_z_0 = self.esm_z_mlp(esm_z)
207         else:
208             s_z_0 = s_s_0.new_zeros(B, L, L, self.cfg.trunk.pairwise_state_dim)
209
210         s_s_0 += self.embedding(aa)
211
212         structure: dict = self.trunk(
213             s_s_0, s_z_0, aa, residx, mask, no_recycles=num_recycles
214         )
```

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

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

→ The sequence embedding contains all the structural information, but only needs sequence data to obtain during training!



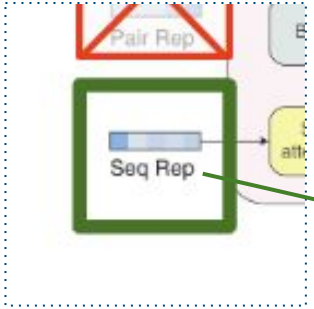
harness additional sequence-based priors

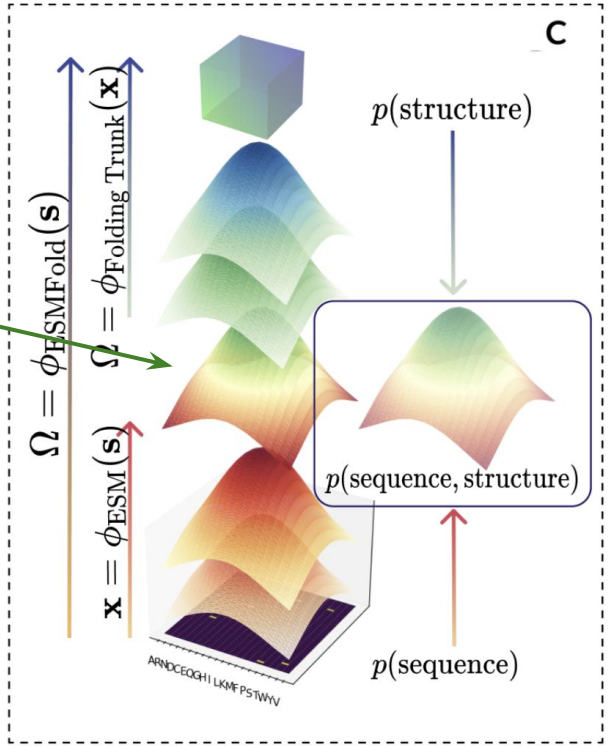
learn structural features from sequence latents

generate structures

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

→ The sequence embedding contains all the structural information, but only needs sequence data to obtain during training!

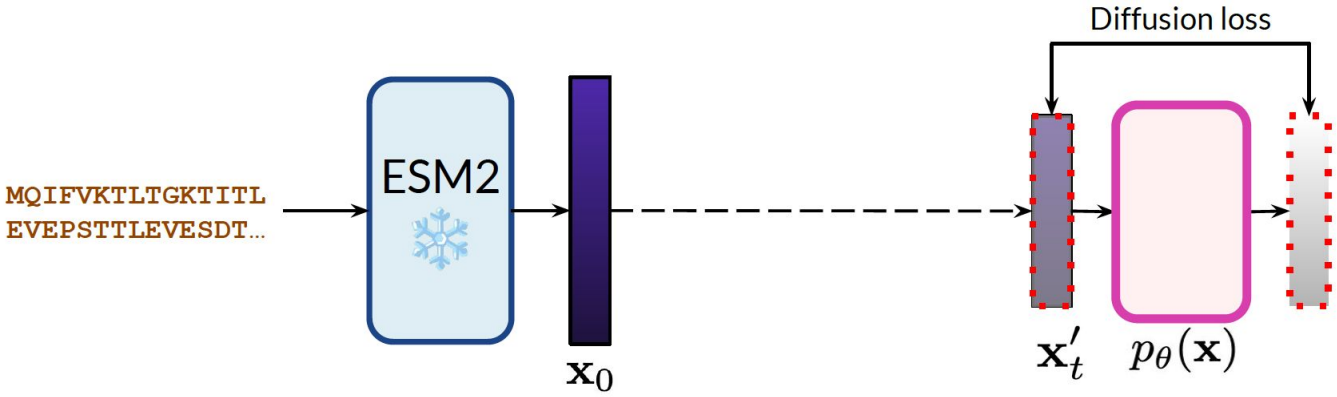




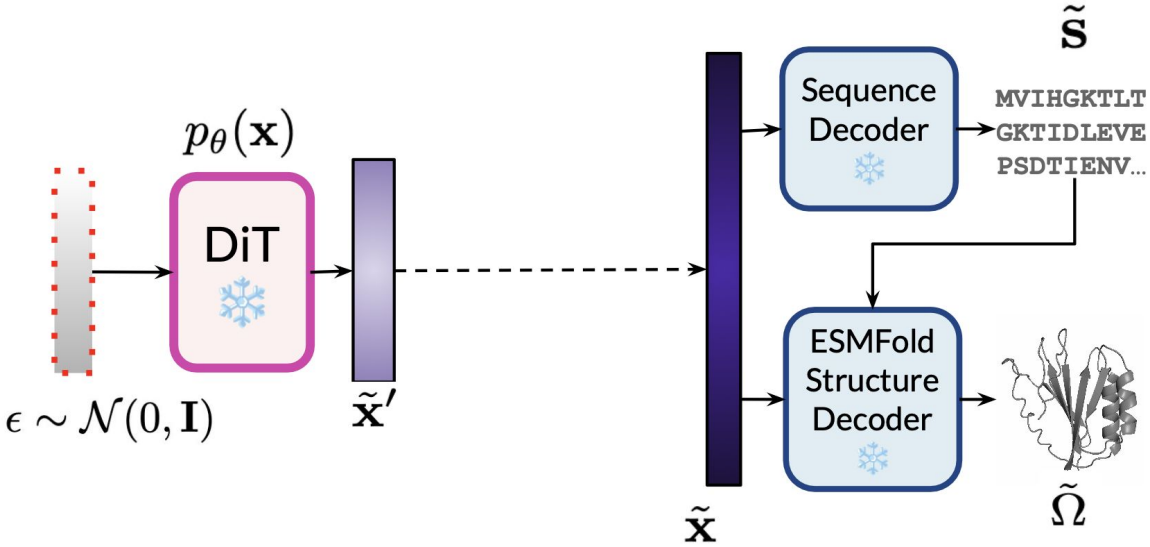
This latent space jointly represents sequence and structure, derived solely from sequence input.

Genentech
A Member of the Roche Group

PLAID: **Training the diffusion model**



PLAID: **Inference-time generation**



PLAID v0.5: **Our early attempt at diffusing in this latent space…**

Commits on Jan 25, 2023

Rename project

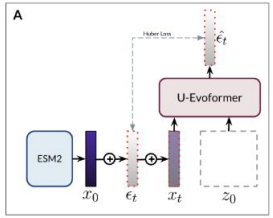
amyxlu committed on Jan 25, 2023

Add exploratory notebooks and ESMFold as denoiser

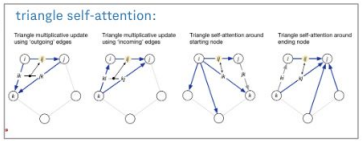
amyxlu committed on Jan 25, 2023

training: standard diffusion training from ESM2 outputs

inference: generate new sequence representations, decode to sequence and structure



architecture: stacked triangle self-attention blocks, as introduced in AlphaFold, to capture pairwise interactions



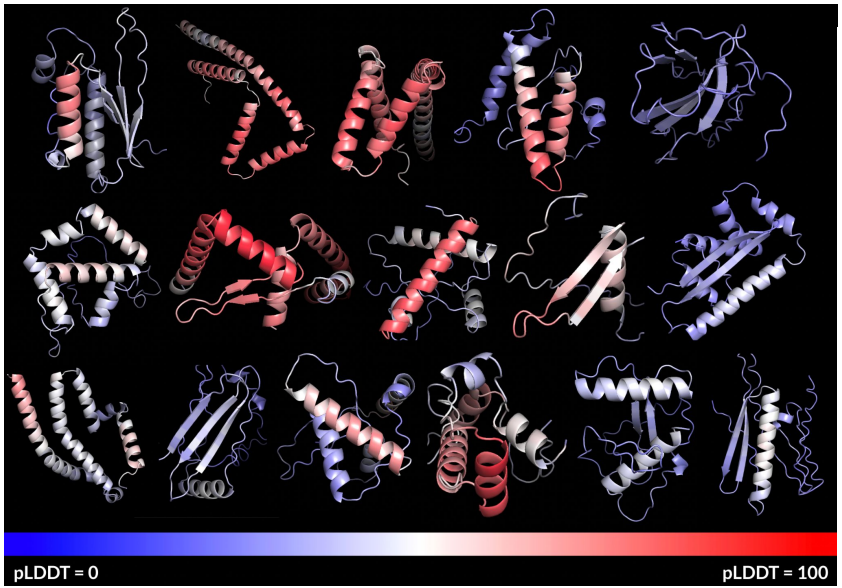
PLAID v0.5: Generating Protein Sequence and Structure Without Structural Training Data

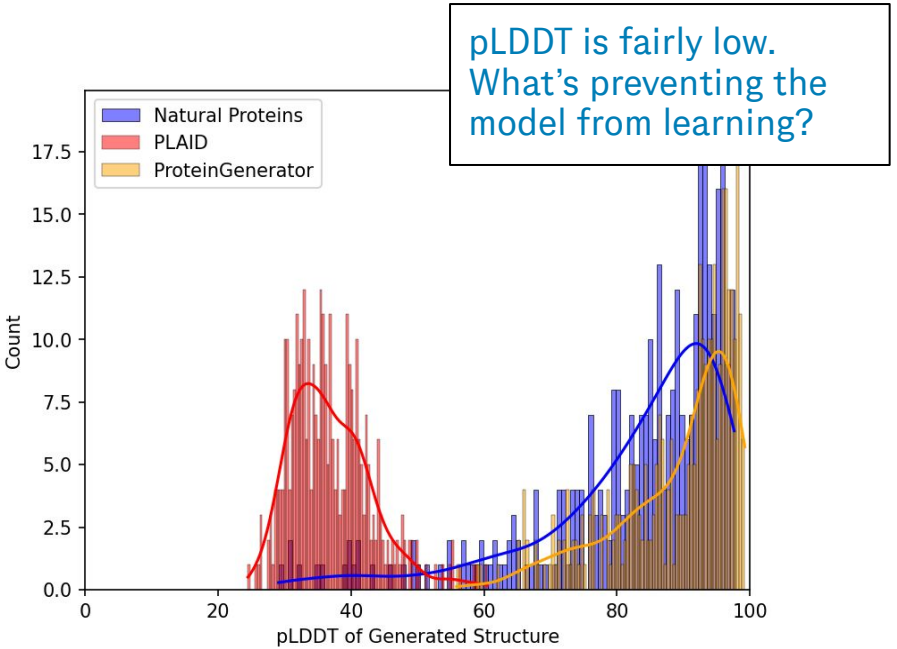
Amy X. Lu, Kevin K. Yang, Pieter Abbeel

ICML 2024 Workshop on Machine Learning for Life and Material Sciences

Genentech
A Member of the Roche Group

PLAID v0.5: **Our early attempt at diffusing in this latent space…**



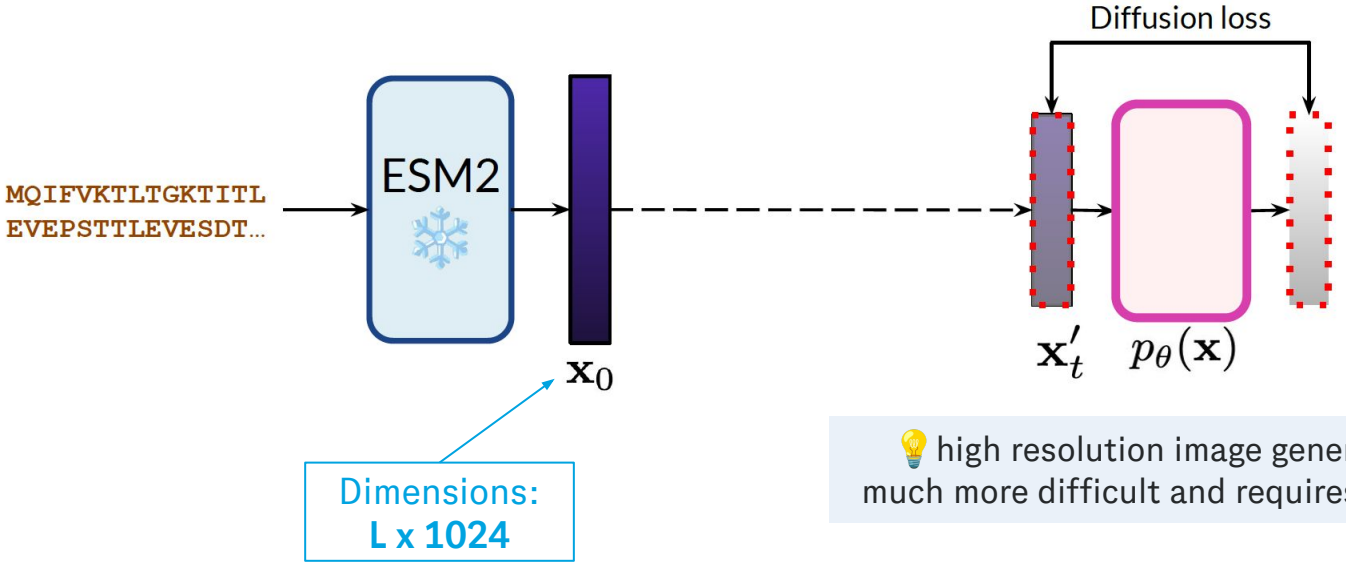


pLDDT is fairly low. What's preventing the model from learning?

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

Genentech
A Member of the Roche Group

Adding embedding compression with CHEAP...

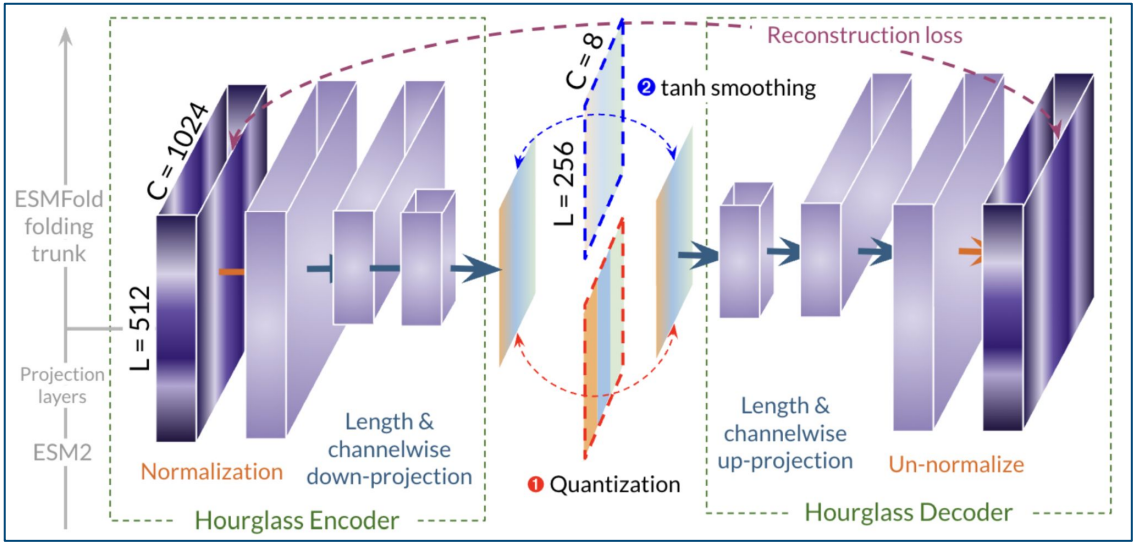


💡 high resolution image generation is often much more difficult and requires compression...

Tokenized and Continuous Embedding Compressions of Protein Sequence and Structure
Amy X. Lu, Wilson Yan, Kevin K. Yang, Vladimir Gligorijevic, Kyunghyun Cho, Pieter Abbeel, Richard Bonneau, Nathan Frey
bioRxiv
bit.ly/cheap-proteins

Genentech
A Member of the Roche Group

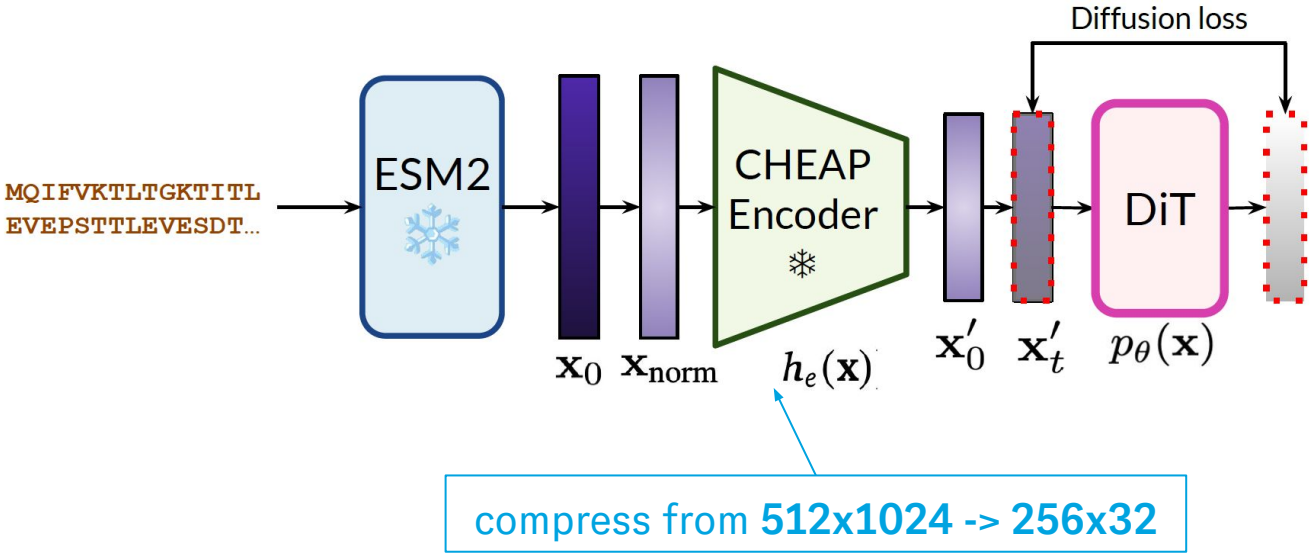
Adding embedding compression with CHEAP...



Tokenized and Continuous Embedding Compressions of Protein Sequence and Structure
Amy X. Lu, Wilson Yan, Kevin K. Yang, Vladimir Gligorijevic, Kyunghyun Cho, Pieter Abbeel, Richard Bonneau, Nathan Frey
bioRxiv
bit.ly/cheap-proteins

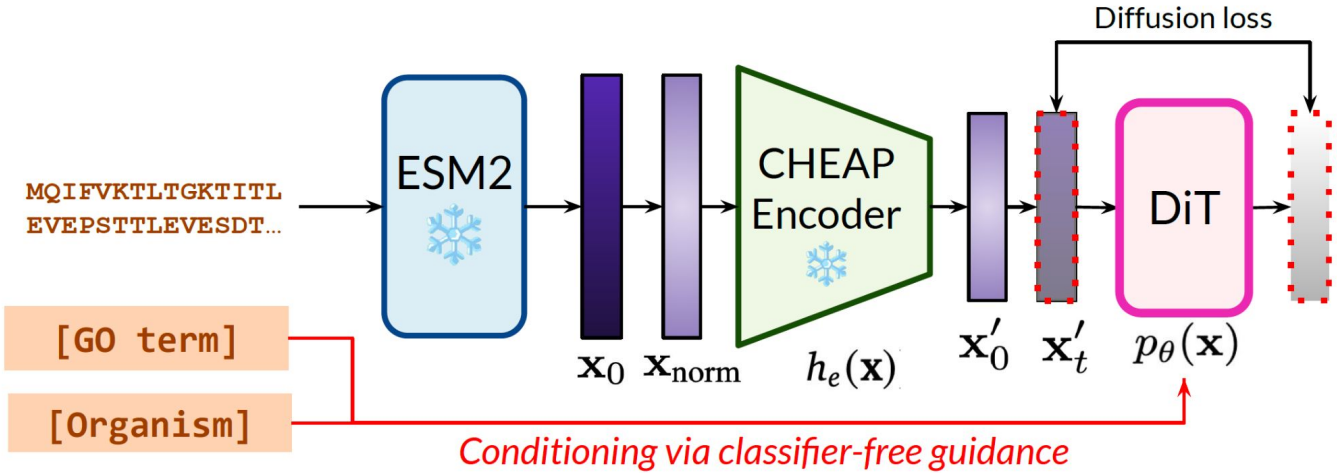
Genentech
A Member of the Roche Group

Adding embedding compression with CHEAP...



Tokenized and Continuous Embedding Compressions of Protein Sequence and Structure
Amy X. Lu, Wilson Yan, Kevin K. Yang, Vladimir Gligorijevic, Kyunghyun Cho, Pieter Abbeel, Richard Bonneau, Nathan Frey
bioRxiv
bit.ly/cheap-proteins

Adding compositional **function** + **taxonomic** conditioning

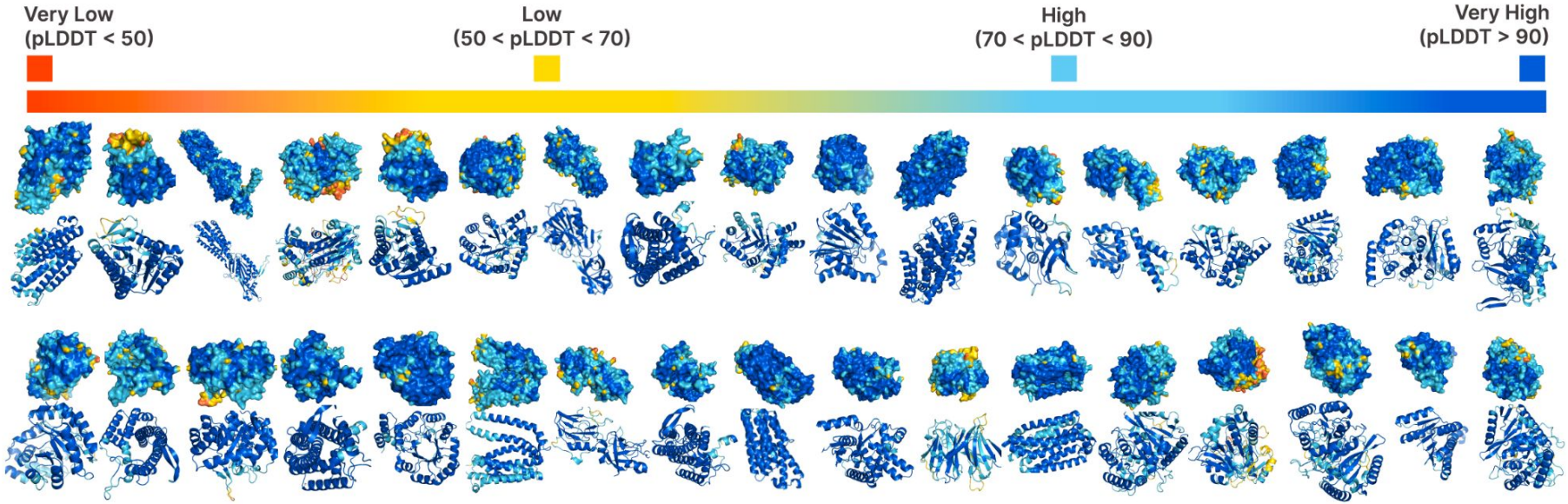


Since sequence databases have more annotations, we can also better control generation!

Genentech
A Member of the Roche Group

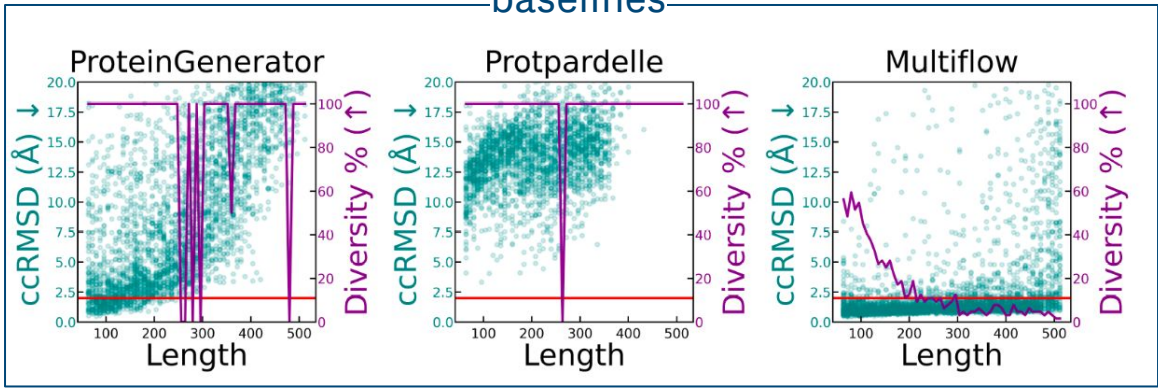
Results

PLAID unconditionally generates diverse, high-quality folds



PLAID unconditionally generates diverse, high-quality folds

baselines



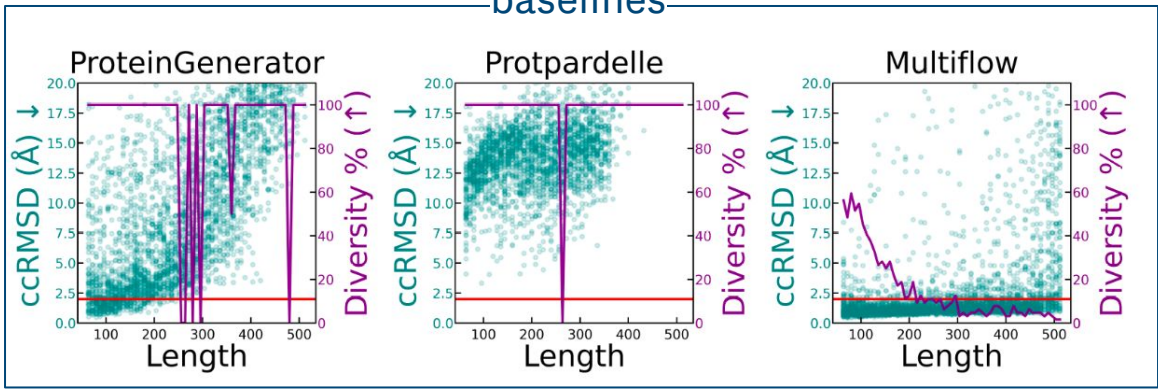
teal: quality (↓)
(RMSD between generated structure and predicted structure of generated sequence)

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

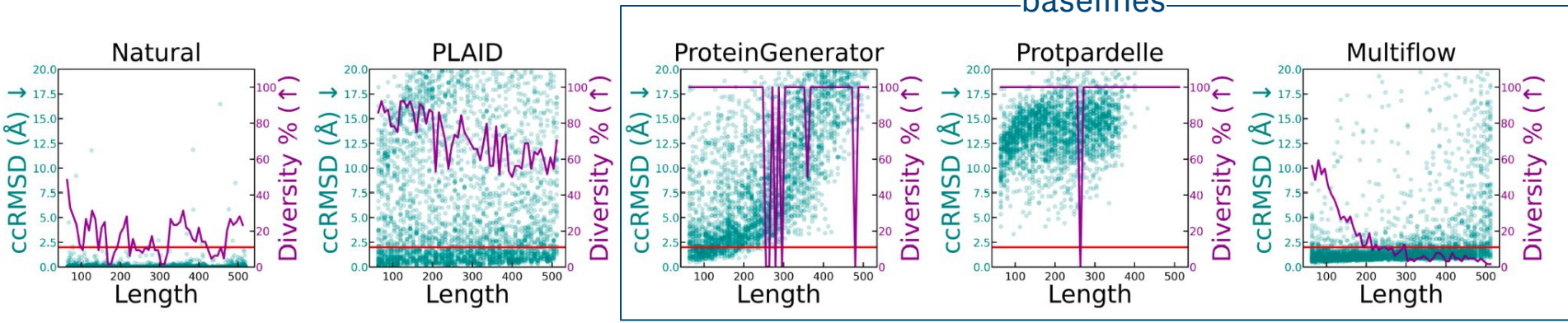
Genentech
A Member of the Roche Group

PLAID unconditionally generates diverse, high-quality folds

Existing methods struggle to generate designable (ccRMSD <2Å) structures at longer sequence lengths while maintaining diversity.



PLAID unconditionally generates diverse, high-quality folds

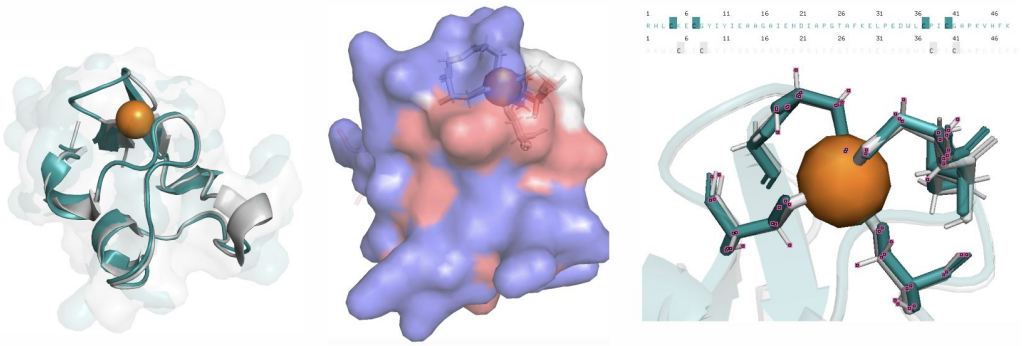


PLAID better balances diversity and quality, especially at longer sequence lengths.

Function-prompted generations learn active site sidechains

Prompt:

HUMAN [and] IRON ION BINDING



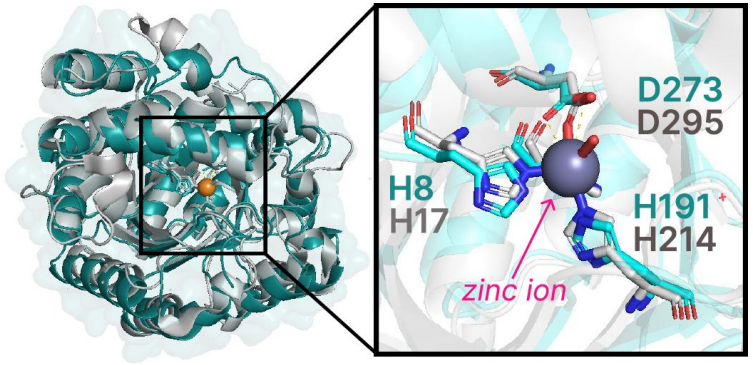
RMSD: **0.35Å**
Seq. Id.: **53.3%**

Sampled
3RYG (128 hours neutron structure of perdeuterated rubredoxin)

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



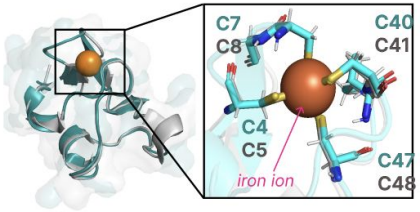
RMSD: **2.25Å**
Seq. Id.: **24.3%**

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

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

Function-prompted generations learn active site sidechains

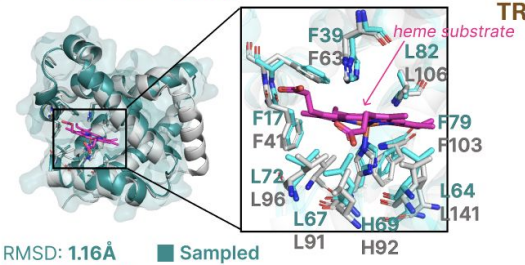
Prompt:
HUMAN [and] IRON ION BINDING



RMSD: **0.35Å**
Seq. Id.: **53.3%**

■ Sampled
3RYG (128 hours neutron structure of perdeuterated rubredoxin)

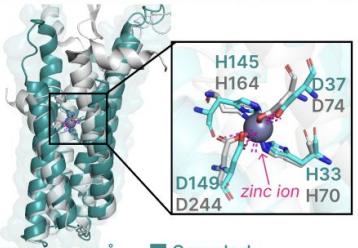
Prompt:
HUMAN [and] HEME BINDING



RMSD: **1.16Å**
Seq. Id.: **53.3%**

■ Sampled
1y4v (T-To-T(High) quaternary transitions in human hemoglobin)

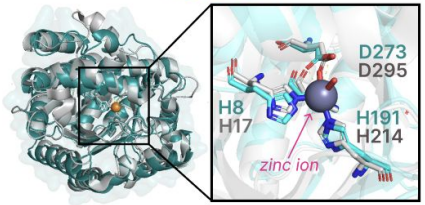
Prompt:
HUMAN [and] MONOATOMIC CATION TRANSMEMBRANE TRANSPORTER ACTIVITY



RMSD: **0.78Å**
Seq. Id.: **53.6%**

■ Sampled
8J80 (hZnT7-Fab complex in zinc state 1)

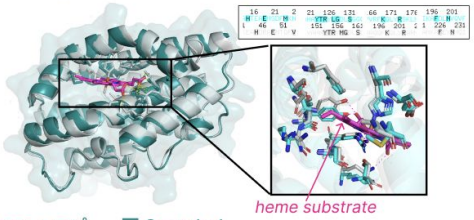
Prompt:
HUMAN [and] DEAMINASE ACTIVITY



RMSD: **2.25Å**
Seq. Id.: **24.3%**

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

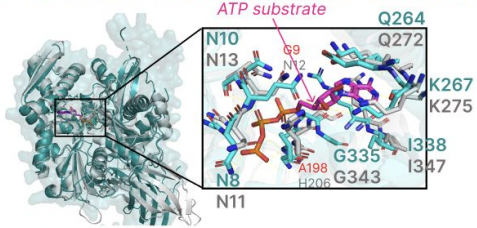
Prompt:
HUMAN [and] HEME OXYGENASE (DECYCLIZING) ACTIVITY



RMSD: **0.78Å**
Seq. Id.: **53.6%**

■ Sampled
2QPP (Crystal structure of human heme oxygenase-2 C127A (HO-2) with bound heme)

Prompt:
HUMAN [and] ATP-DEPENDENT PROTEIN FOLDING CHAPERONE

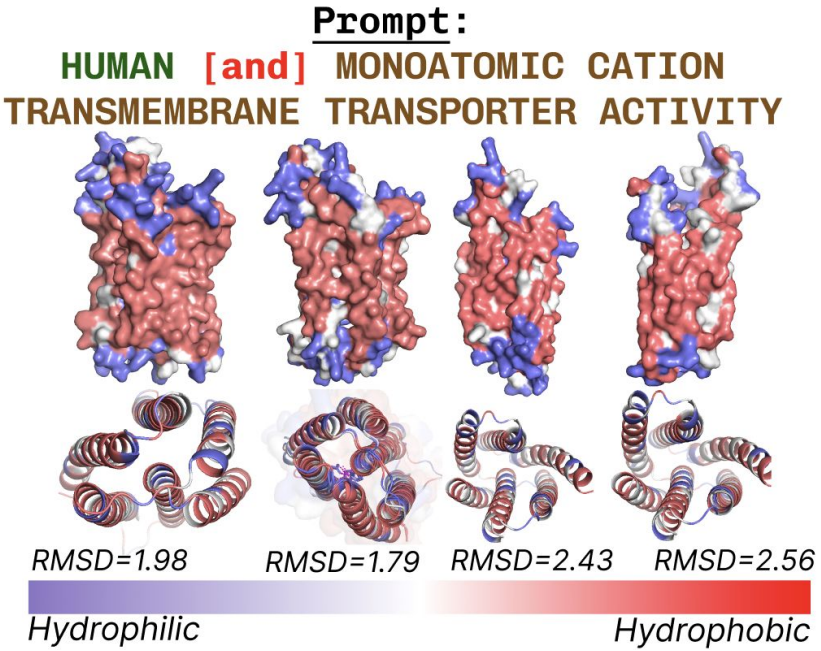


RMSD: **1.73Å**
Seq. Id.: **29.2%**

■ Sampled
2QXL (Crystal Structure Analysis of Sse1, a yeast Hsp110)

Genentech
A Member of the Roche Group

Transmembrane proteins exhibit expected hydrophobicity patterns



Hydrophobic residues are found at the core, as expected.

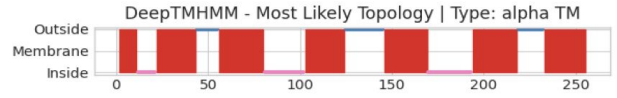
Transmembrane proteins exhibit expected numbers of helices

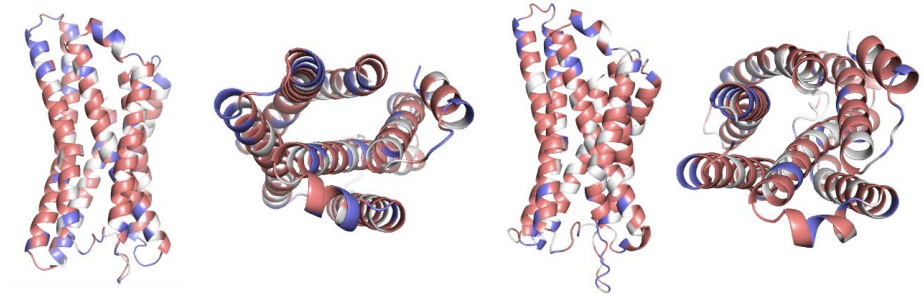
Prompt:

HUMAN [and]

G PROTEIN-COUPLED RECEPTOR ACTIVITY

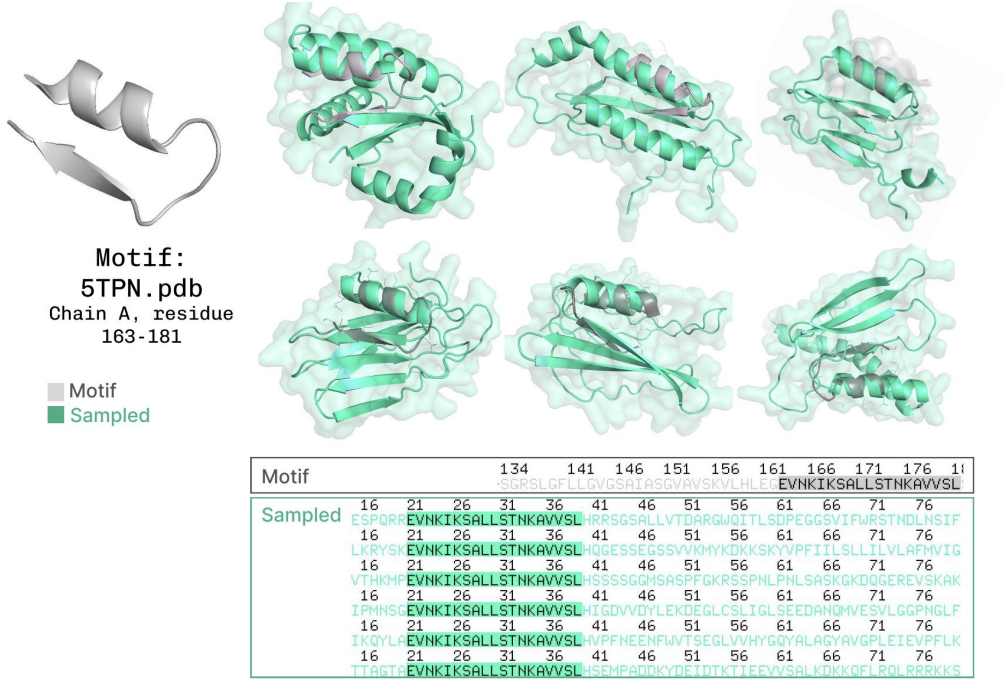
GNVLVLIMMILKQREVKSMPNV
WVFNLALSDLLFLLSTPLLVVK
MSDTSWNLGLSPCKITTFLLFL
NLYSSVFFLACLSLDRYLTVRQ
VRSN...

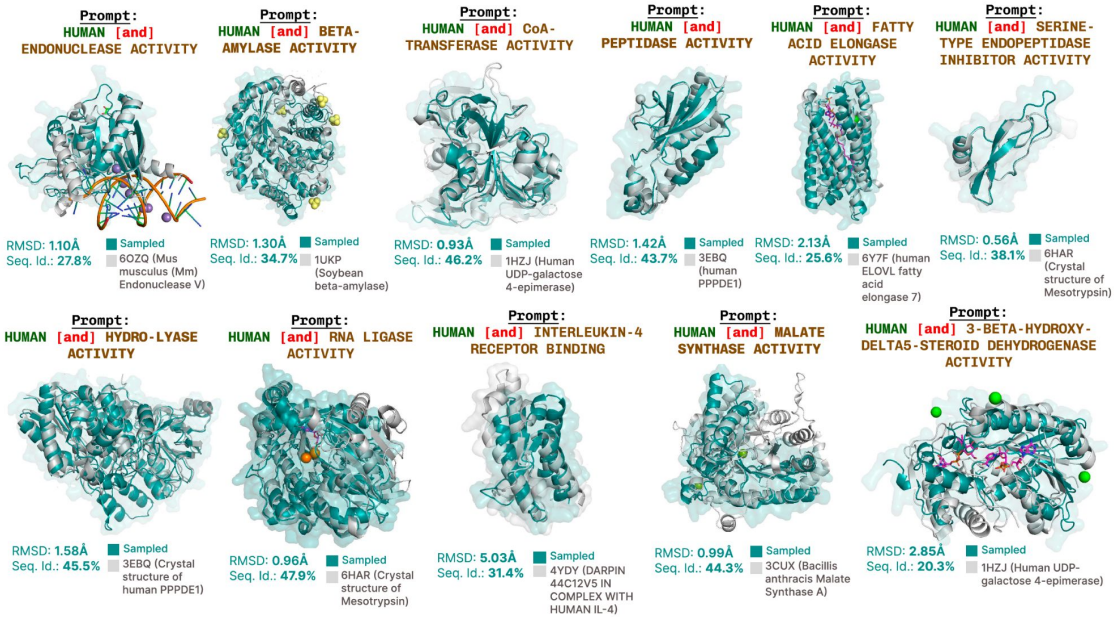




GPCRs have the expected 7-transmembrane topology, both when analyzing the sequence and structure.

Motif scaffolding without retraining





For more results, see our paper: bit.ly/plaid-proteins

Genentech
A Member of the Roche Group

Are you interested in using PLAID for your wet-lab protein designs?

Reach out to amyxlu@berkeley.edu / freyn6@gene.com

Thanks!

Berkeley
Amy X. Lu
Wilson Yan
Pieter Abbeel

Microsoft Research
Kevin Yang

Prescient Design
Sai Pooja Mahajan
Sarah Robinson
Vladimir Gligorijevic
Kyunghyun Cho
Richard Bonneau
Nathan C. Frey

Paper: bit.ly/plaid-proteins
Code: github.com/amyxlu/plaid
Weights: hf.co/amyxlu/plaid

@amyxlu amyxlu.github.io amyxlu@berkeley.edu

Appendix

Motivation: **Direct sampling from the joint distribution is natural**

Structure generation + inverse folding

$p(\text{branches})$

Generate a picture of tree branches.

Here is the image of tree branches you requested. If you need any modifications or additional details, feel free to ask!

$p(\text{apples}|\text{branches})$

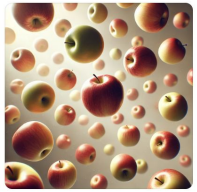
Add apples to this tree branch.

Here's the updated image with apples added to the tree branches. If you need any further adjustments, just let me know!

Sequence generation + folding

$p(\text{apples})$

Generate a picture of apples in midair.



Here is the image of apples suspended in midair. I hope it captures the surreal effect you were looking for!

$p(\text{branches}|\text{apples})$

Add branches to these apples.

Here is the updated image with branches holding the apples in place. The branche touch while maintaining the surreal effect. I hope this meets your expectations!

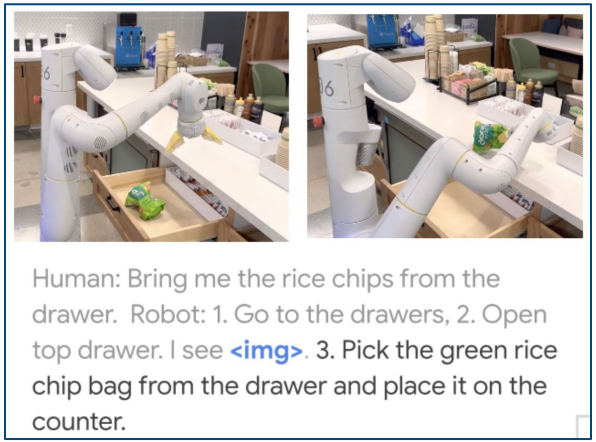
Co-generation

$p(\text{apples}, \text{branches})$

Generate a picture of an apple tree.

Here is the image of an apple tree you requested! Let me know if you need any adjustments or further assistance.

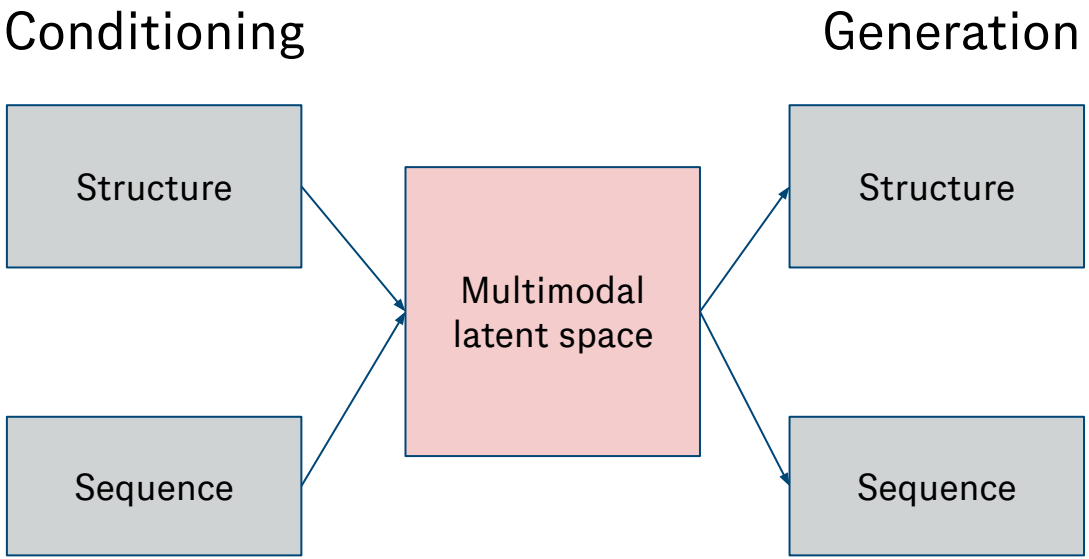
Motivation: **How can we repurpose information in pretrained models for generation?**



PaLM-E: An embodied multimodal language model.
Dreiss et al., 2023

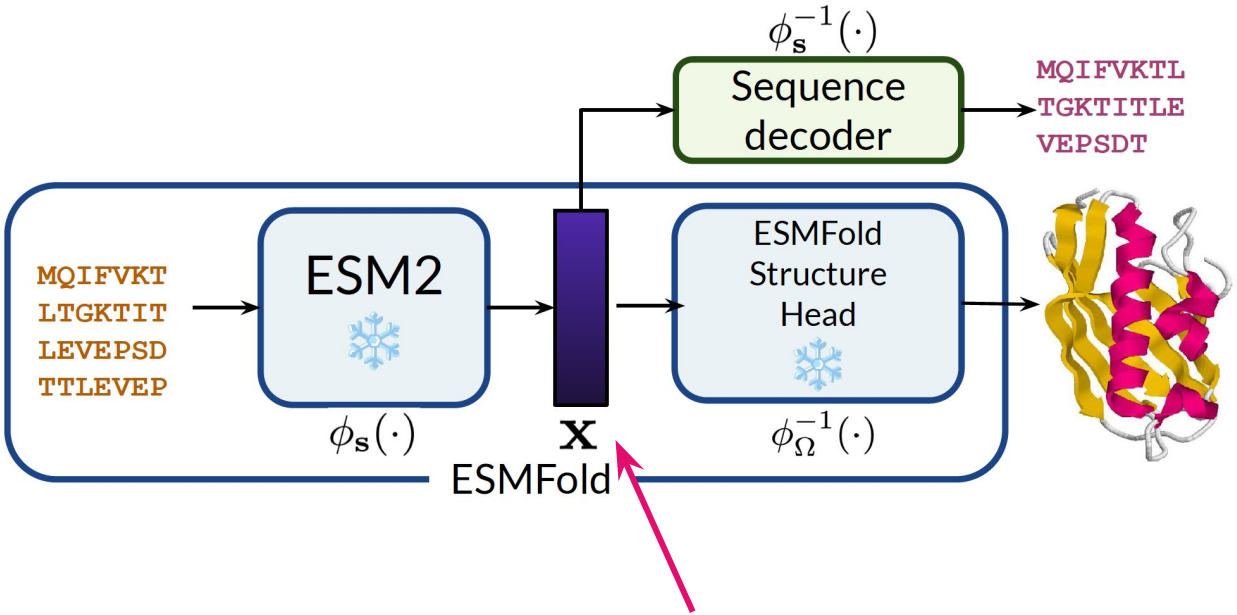
→ Can we use information captured by pretrained structure ***prediction*** models for protein ***generation***?

Motivation: **Sampling directly from the joint distribution**

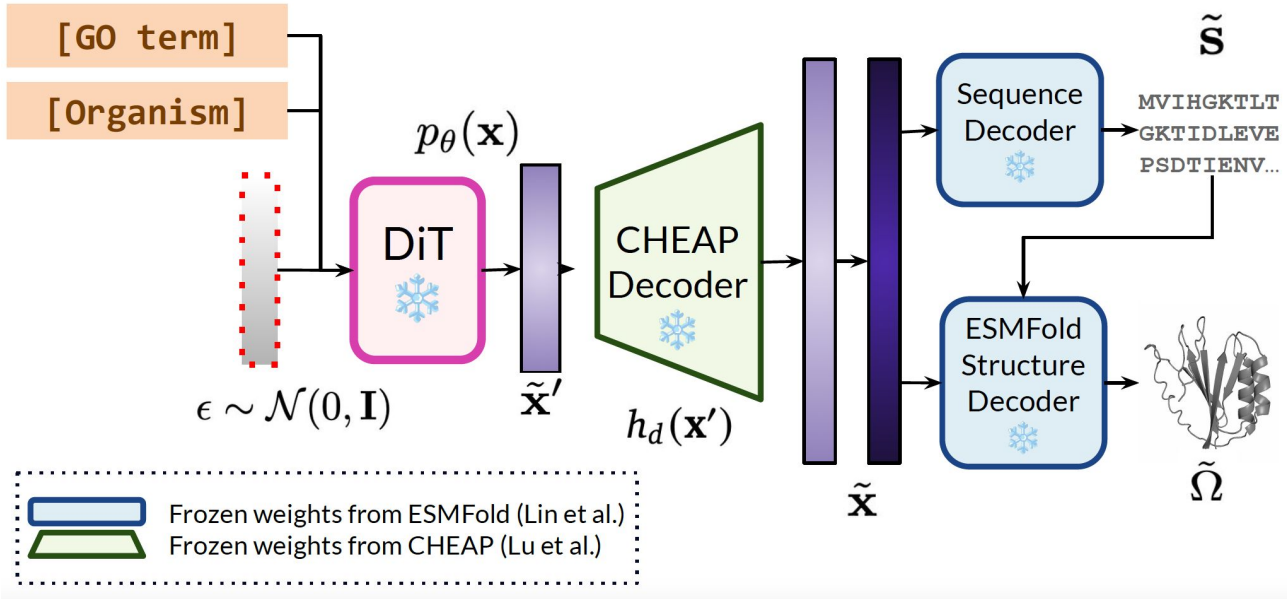


Being able to characterize a joint latent space allows flexibly conditioning by and generating either modality.

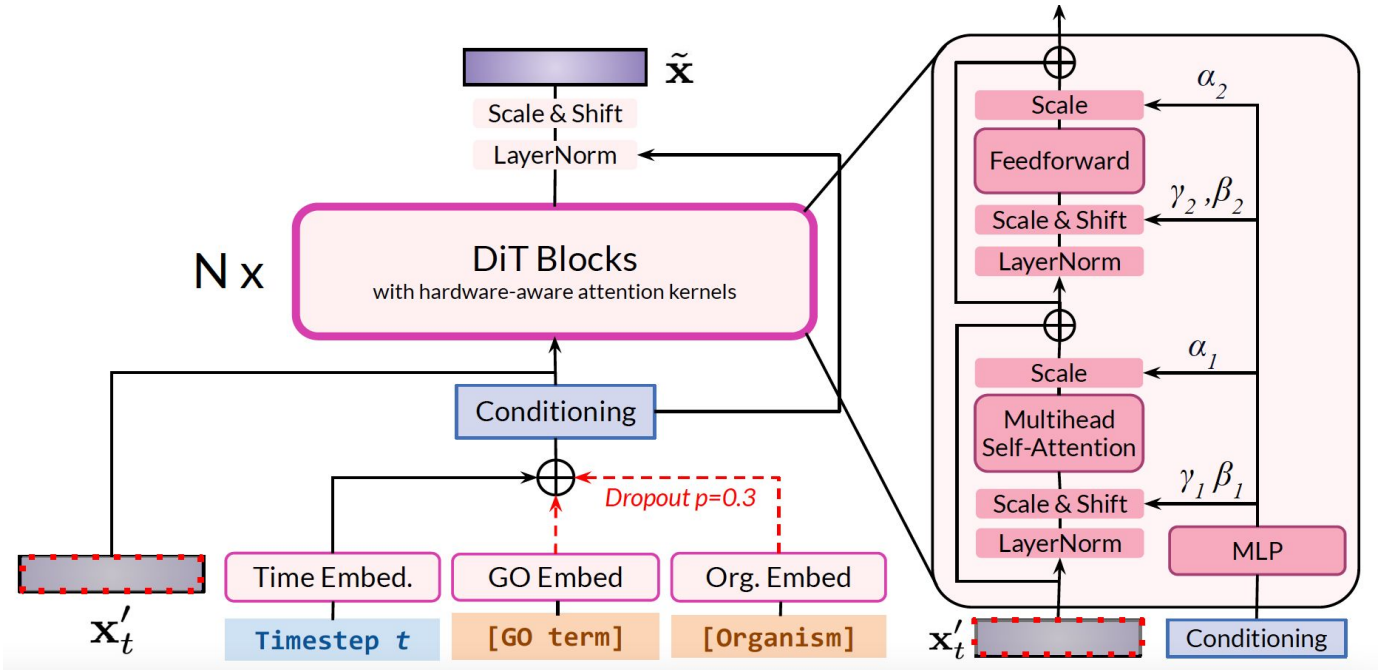
Defining the space for latent generation



PLAID: **Inference-Time Generation**

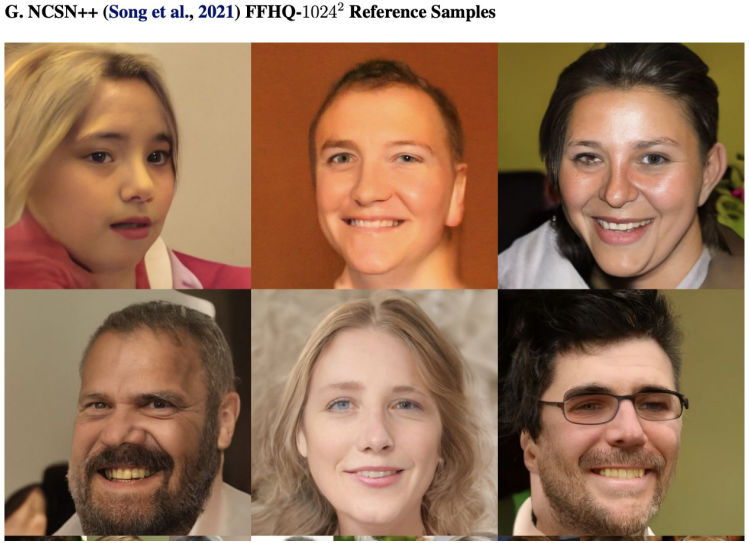


Method: **Conditioning**



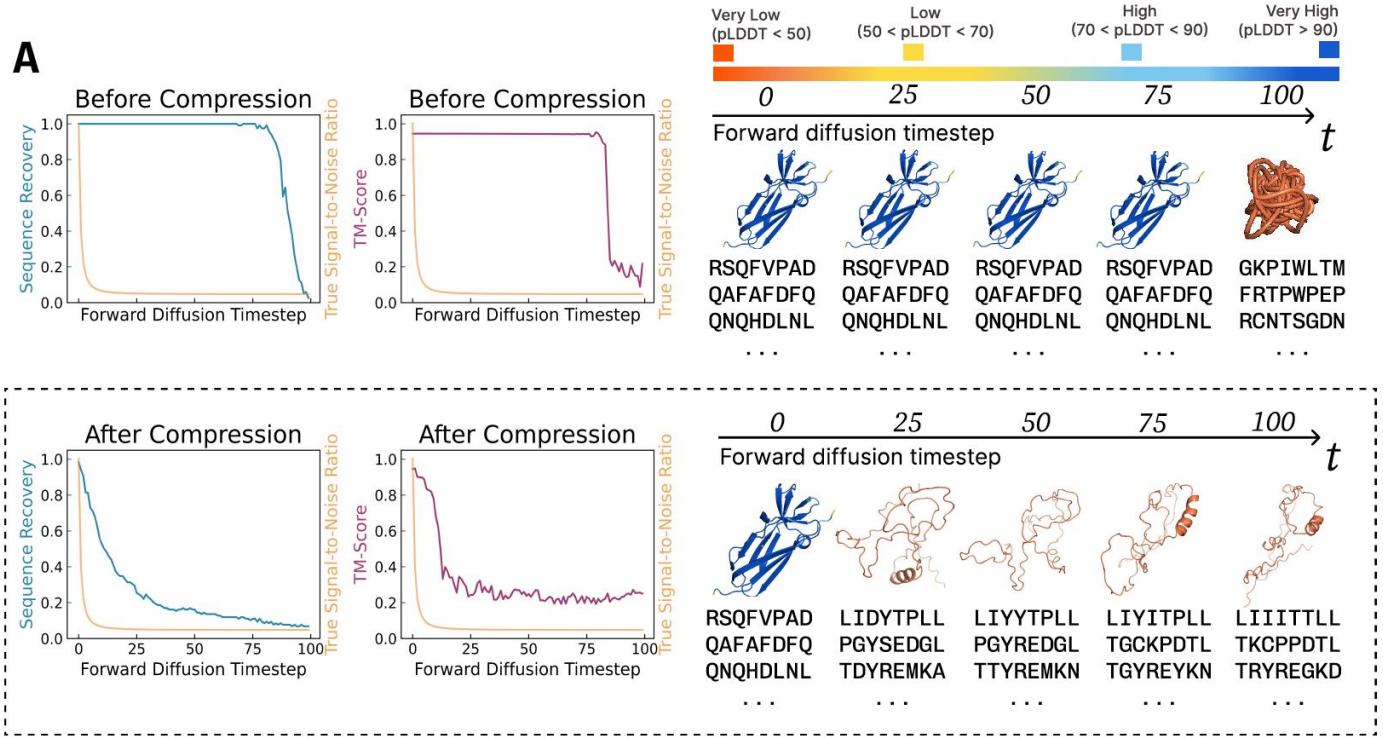
Issues and hypotheses -> CHEAP

- Latent space requires regularization
- Training data only allows for length of 128 due to memory constraints
  - Some samples show the curvatures of a beta barrel, but sequence length limits seeing a full beta barrel
    - Need to shorten the protein?
- pLDDT is not designed to assess generation from evolutionary scale datasets
  - Biased towards generative models trained on the same data as AF2, i.e. PDB
- Large latent space corresponds to high-resolution image generation
  - in LDMs, latent space is 64 x 4 x 4, as opposed to ours, which is 512 x 1024

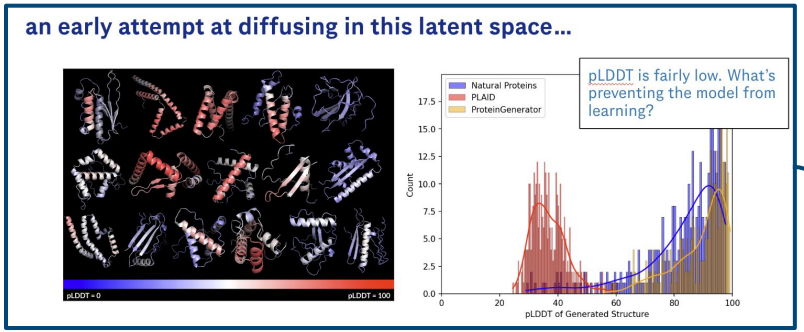


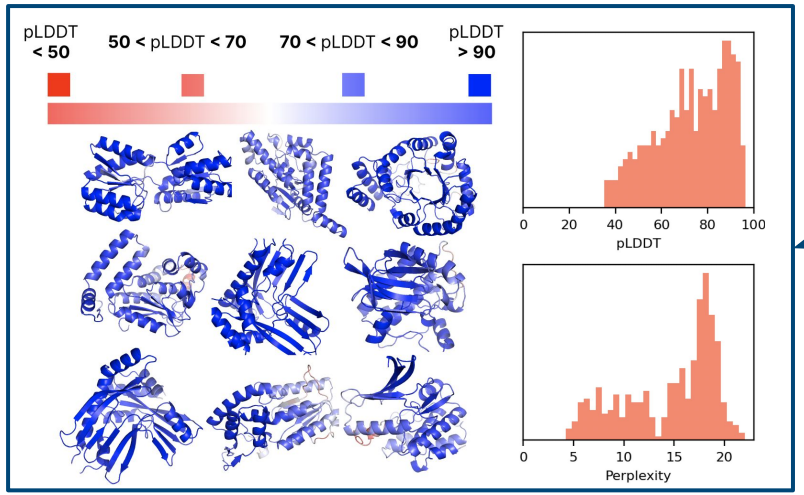
Diffusion models in their naive formulation often fail for 1024 x 1024 resolution generation.

CHEAP embeddings smooth out the latent space for generation



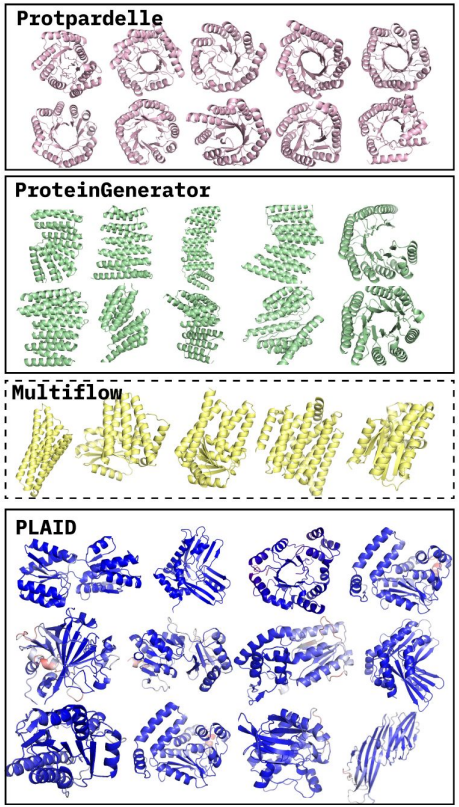
From PLAID v0.5 -> **final PLAID model:**





1. Learn diffusion model in regularized and compressed latent space
   a. mirrors the regularized autoencoder in LDM
2. Can learn on longer sequences due to CHEAP shortening
3. Use DiT instead of U-triangular self attention
   a. allows for scaling up to higher parameter counts
4. Scale up to 2B parameters with BS=2048

PLAID unconditionally generates diverse, high-quality folds



Protpardelle
>len60O_samp97
AGGGGGGGGGGGGGGGGGGGGGGLGLGLLLPPAGL...
>len60O_samp98
PPPPGGAGGGGAAAALAGGSPGGPPGGGGGGGGGGGGG...
>len60O_samp99
PPGPALPPSPGPGGVPPPPPLPPPPLPGGAPPAGGGLL...

ProteinGenerator
>len60O_OOOO97
GAAGLTAAAAVVGAAAAAGAAAAAALAAAAGAGAAAAA...
>len60O_OOOO98
AGAAGAAAAAAAAAAGAAAAGAGGGAGGAAAAAAAAAG...
>len60O_OOOO99
VAAAQAVQGAIAAAAALAATAALGLTAAGIAAPLLALV...

Multiflow
>len60O_sample_97
LLGGLLGGLLGGAAGGAGAGAAAAGGGAVGVGVAGAVT...
>len60O_sample_98
ADAATLTVGGGGTGGGGGAGGALGGAAAGGGGRVTLVV...
>len60O_sample_99
AGGGAGLAGGAGGAGGAAAAAAAAAAAAAGAGGGAAAA...

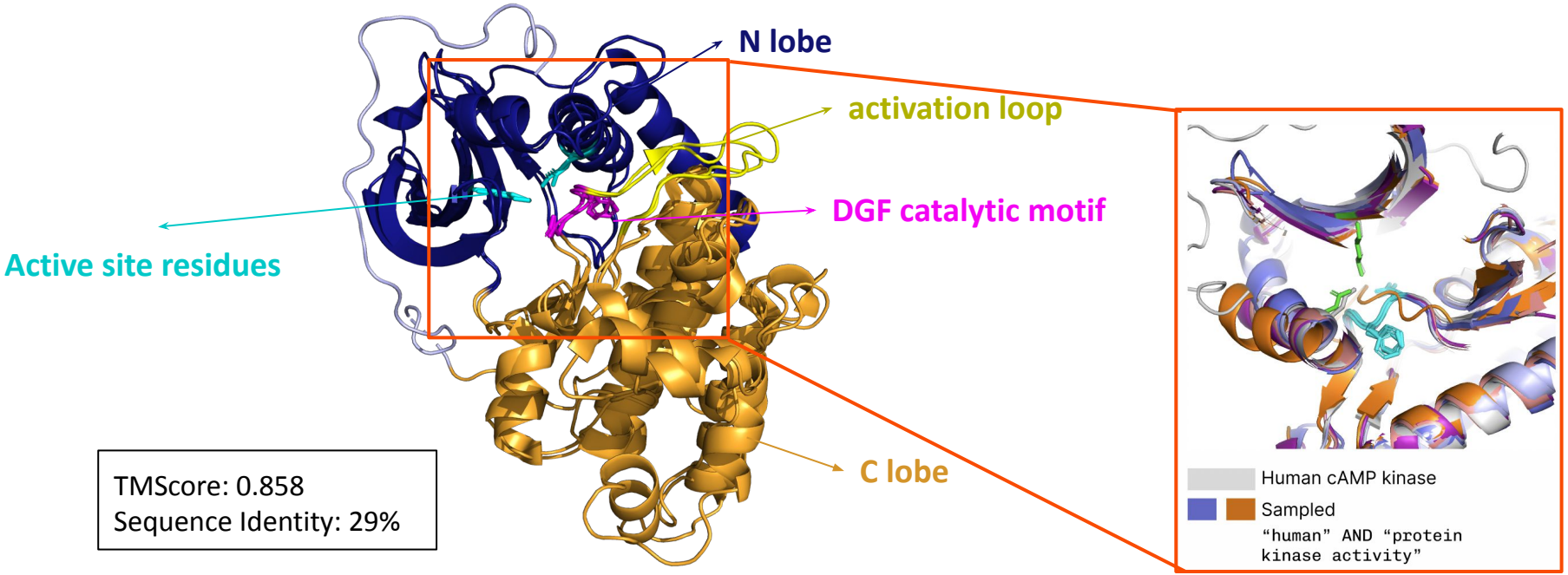
PLAID
>len60O_sample97
PDMGTVLGLAHSVGHLDFKTPDLSVADLETNLALLAAH...
>len60O_sample98
FEMFDDKGGDLWERAASSGQLLIDVAYLANNGLRDGAT...
>len60O_sample99
GNGGQARGTDDPLTHALQTLFQSAALDQSLQGDPENAV...

Genentech
A Member of the Roche Group

Examining active site conservation

prompt: **“human” AND “protein kinase activity”**

Closest Foldseek neighbor: **6cd6 (human calcium/calmodulin-dependent protein kinase kinase 1)**



Function-prompted generations learn active site sidechains