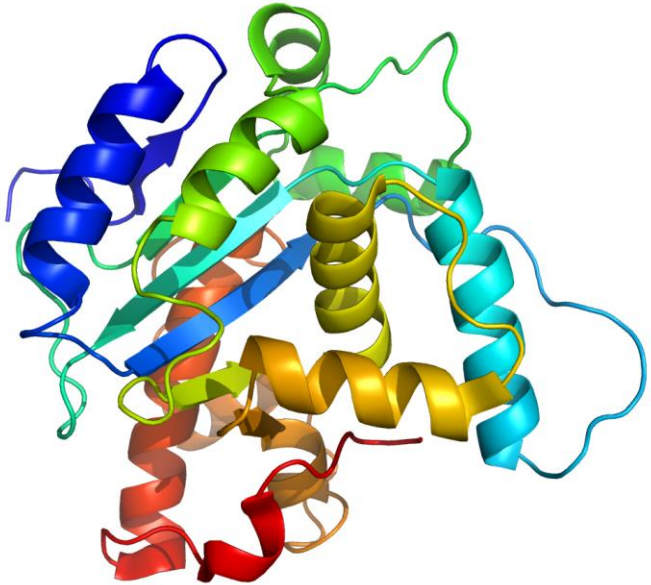
AlphaFold Meets Flow Matching for Generating Protein Ensembles

Sascha Benz

Technische Universität München

Munich, 05. November 2024



How is AlphaFold limited?

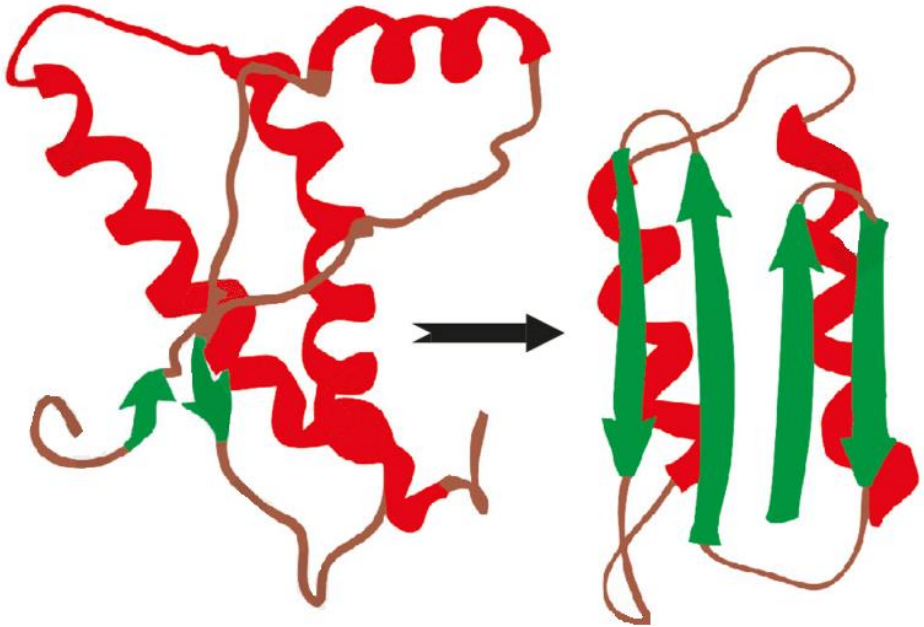


Fig. 2: Scheme of conformational change in a protein.

How is AlphaFold limited?

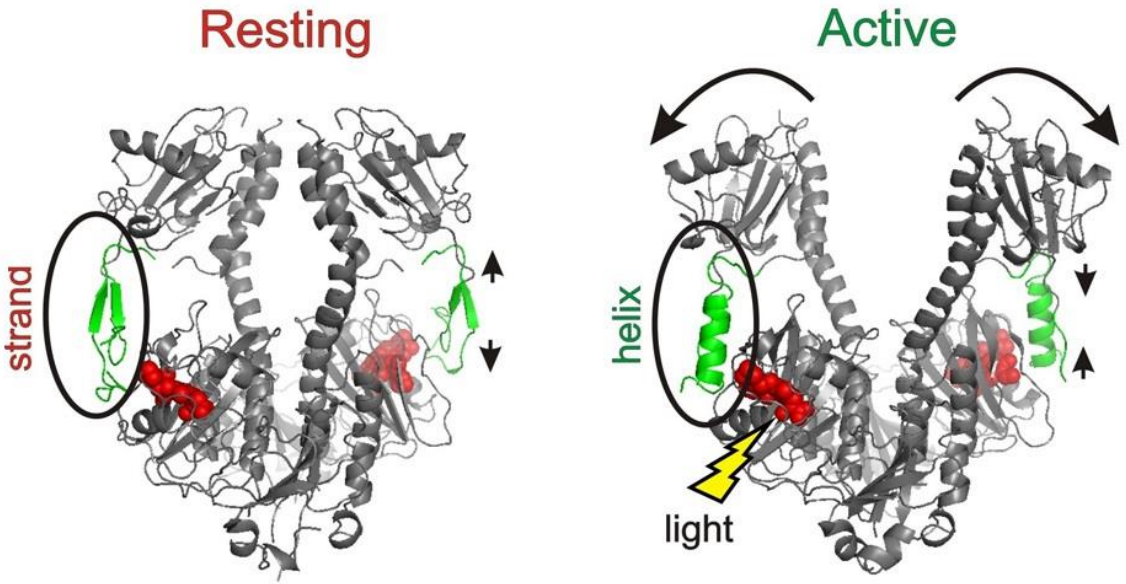


Fig. 3: *Deinococcus radiodurans* undergoing conf. change after being activated by light.

Overview

1. Motivation
2. Background
3. Methodology
4. Experiments
5. Strengths & Limitations
6. Conclusion

Background: Related Work

- Subsampling (most popular), mutagenesis, or clustering of MSA
  - *De-facto* standard methodology
- Sequence-to-structure generative models of protein ensembles
  - Perform poorer than MSA subsampling
- Learning generative models of Boltzmann distributions
  - Do not scale effectively

Flow Matching

Fig. 4: Flow matching objective.

Flow Matching

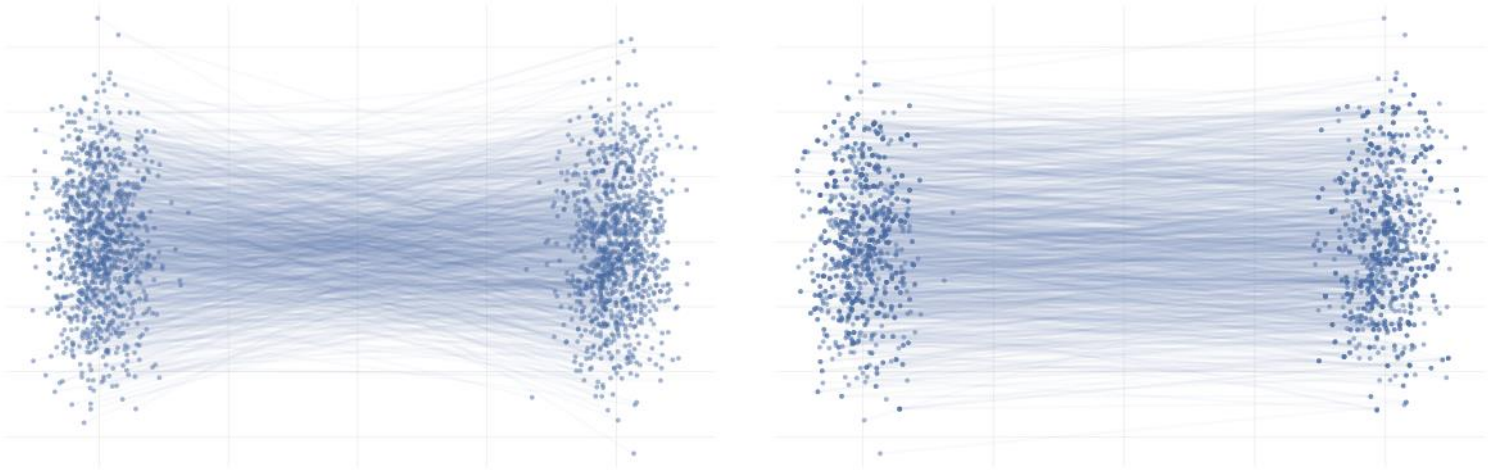


Fig. 5: Gaussian-to-Gaussian uniformly sampled pairings (left) vs. optimal transport (right).

Flow Matching

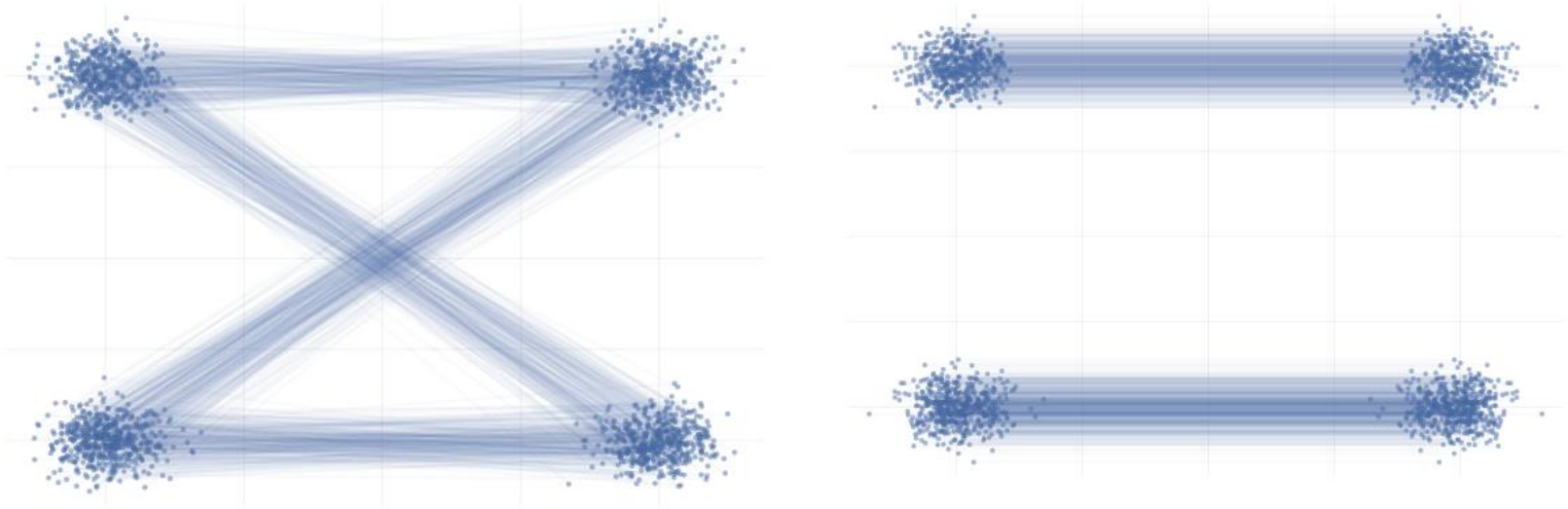


Fig. 6: Mixture of Gaussians uniformly sampled pairings (left) vs. optimal transport (right).

AlphaFold

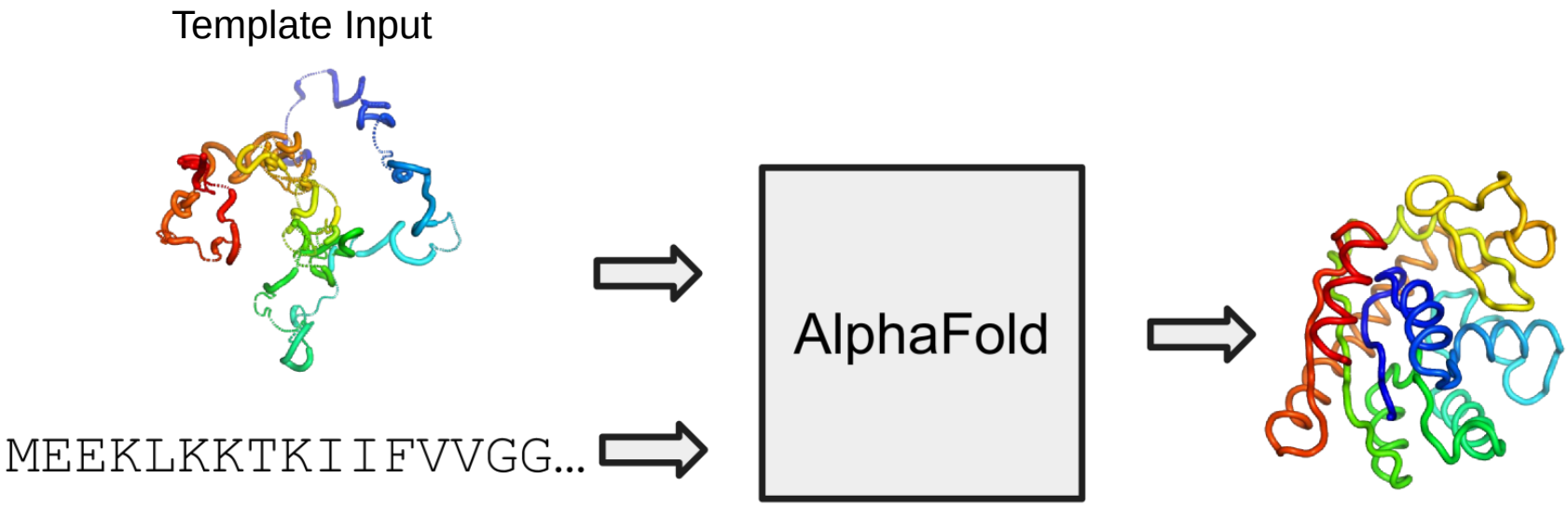


Fig. 7: AlphaFold input and output

ESMFold

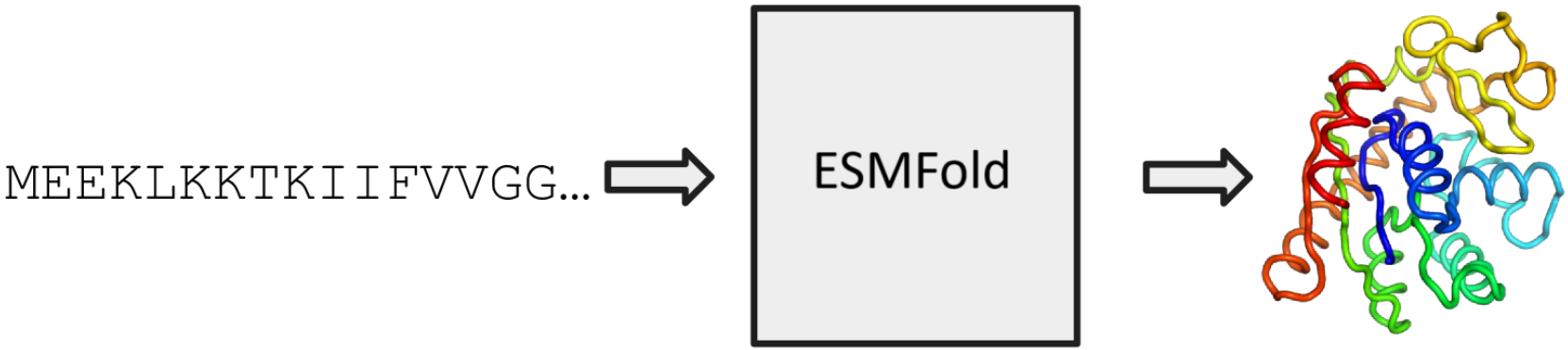


Fig. 8: ESMFold input and output.

Combination of Fold and Flow

- AlphaFold/ESMFold are regression models, not generative

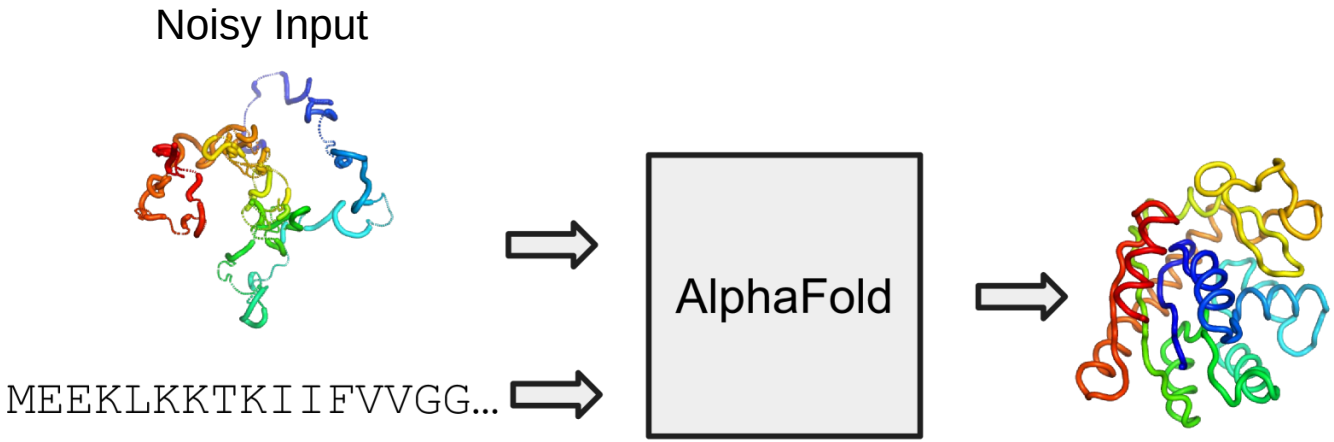


Fig. 9: Making AlphaFold generative.

Combination of Fold and Flow

- Arbitrary output orientation; no propagation possible

  → Space of protein structures $R^{3N}/SE(3)$

  → Linear Interpolation after RMSD alignment

  → Loss: $FAPE^2\left(x_0, f_\theta(x_t, t)\right)$

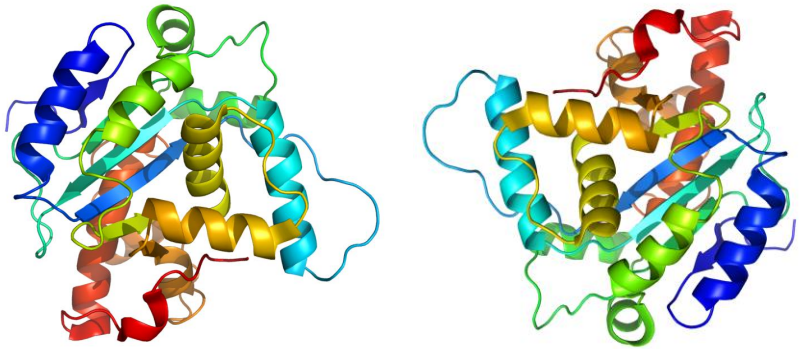


Fig. 10: Mirrored protein.

Flow Matching with AlphaFold2/ESMFold

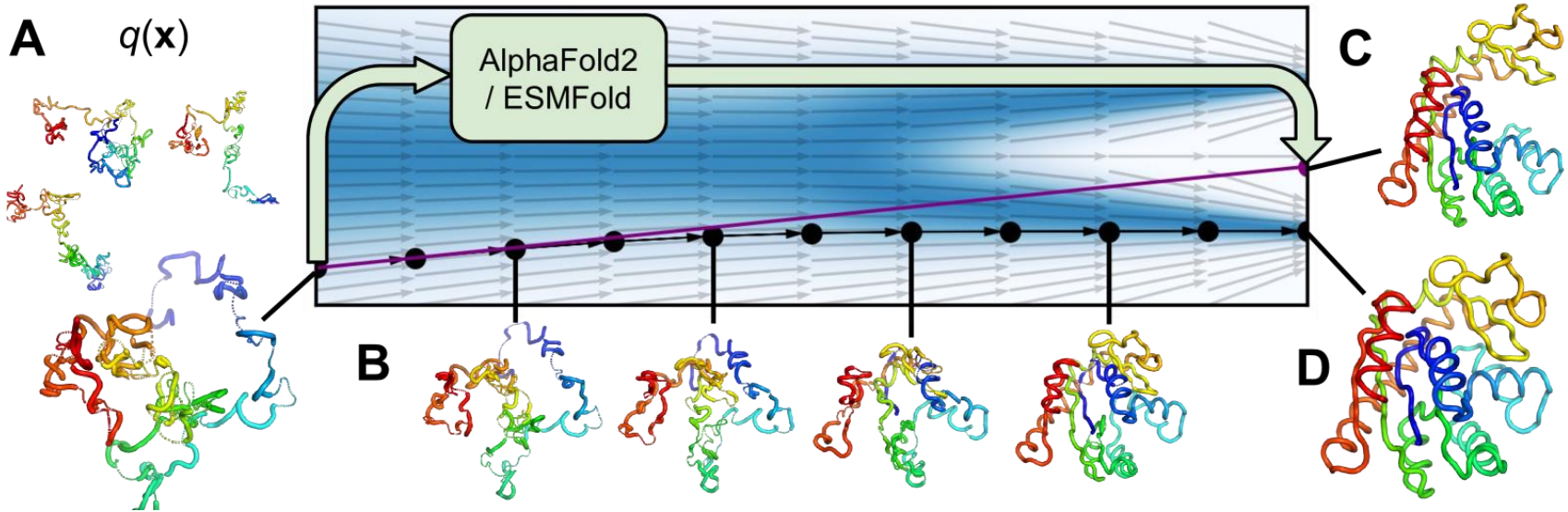
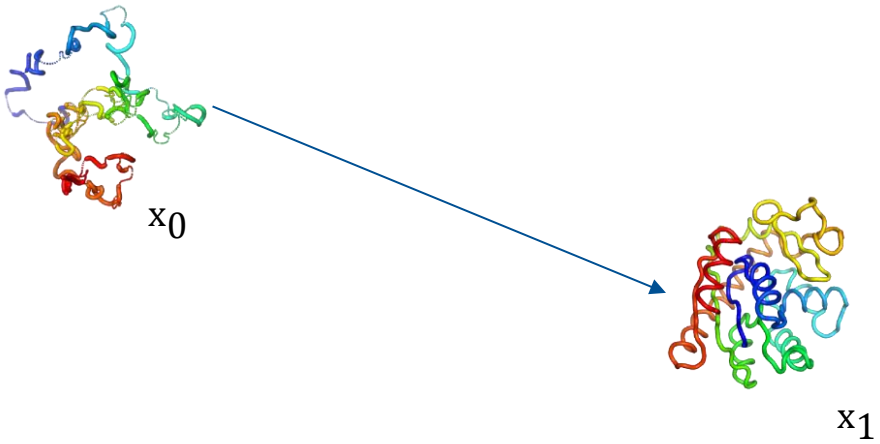


Fig. 11: Flow Matching with AlphaFold step-by-step.

AlphaFlow Training

Prior

Data



Input: Training examples of structures, sequences, and MSAs $\{(S_i, A_i, M_i)\}$

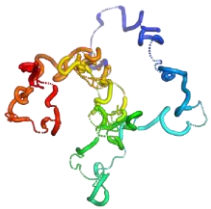
for all (S_i, A_i, M_i) **do**

Extract $x_1 \leftarrow \text{BetaCarbons}(S_i)$

Sample $x_0 \sim \text{HarmonicPrior}(\text{length}(A_i))$

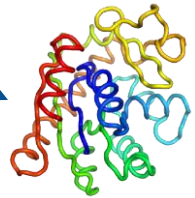
AlphaFlow Training

Prior



x_0

Data



x_1

Input: Training examples of structures, sequences, and MSAs $\{(S_i, A_i, M_i)\}$

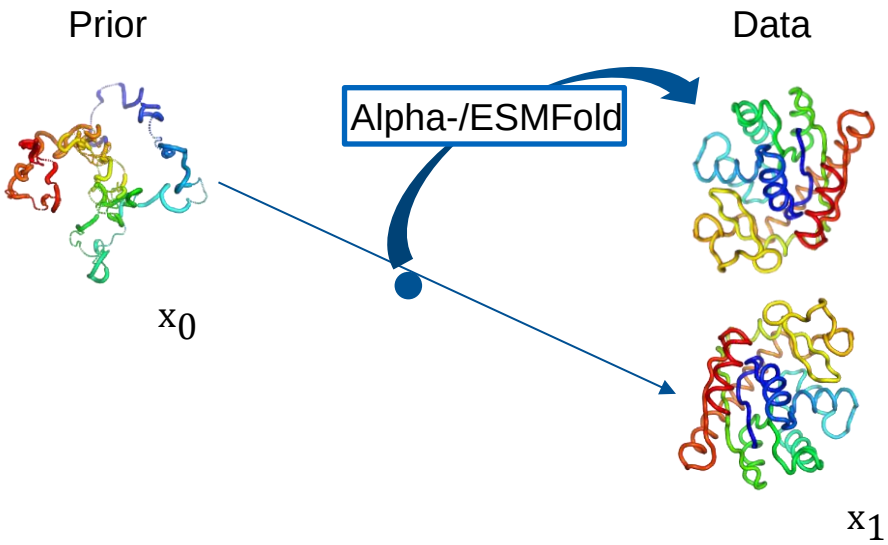
for all (S_i, A_i, M_i) **do**

Extract $x_1 \leftarrow \text{BetaCarbons}(S_i)$

Sample $x_0 \sim \text{HarmonicPrior}(\text{length}(A_i))$

Align $x_0 \leftarrow \text{RMSDAlign}(x_0, x_1)$

AlphaFlow Training



Input: Training examples of structures, sequences, and MSAs $\{(S_i, A_i, M_i)\}$

for all (S_i, A_i, M_i) **do**

Extract $x_1 \leftarrow \text{BetaCarbons}(S_i)$

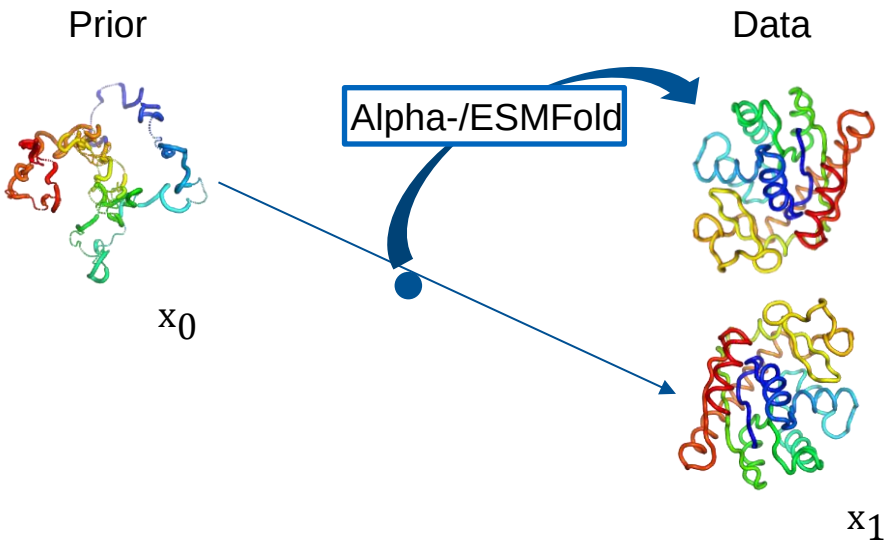
Sample $x_0 \sim \text{HarmonicPrior}(\text{length}(A_i))$

Align $x_0 \leftarrow \text{RMSDAlign}(x_0, x_1)$

Sample $t \sim \text{Uniform}[0, 1]$

Interpolate $x_t \leftarrow t \cdot x_1 + (1 - t) \cdot x_0$

AlphaFlow Training



Input: Training examples of structures, sequences, and MSAs $\{(S_i, A_i, M_i)\}$

for all (S_i, A_i, M_i) **do**

Extract $x_1 \leftarrow \text{BetaCarbons}(S_i)$

Sample $x_0 \sim \text{HarmonicPrior}(\text{length}(A_i))$

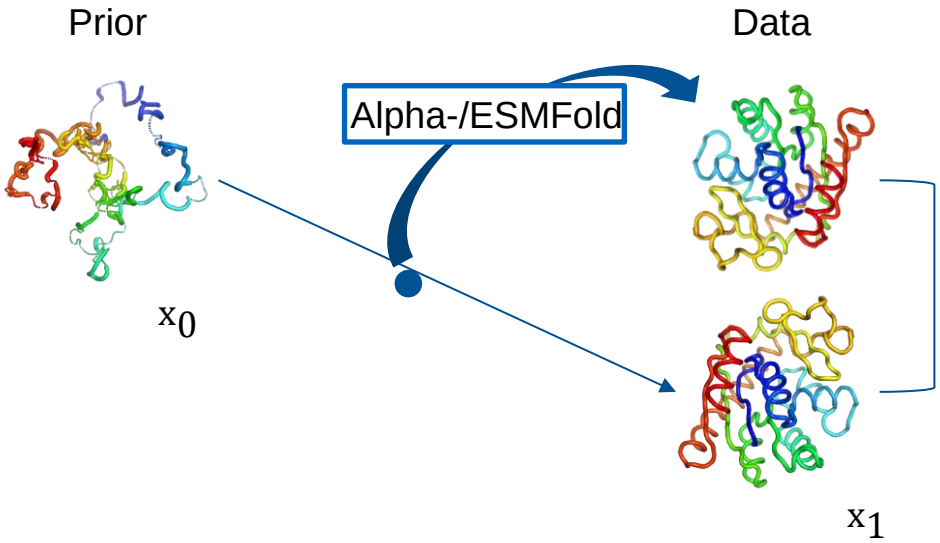
Align $x_0 \leftarrow \text{RMSDAlign}(x_0, x_1)$

Sample $t \sim \text{Uniform}[0, 1]$

Interpolate $x_t \leftarrow t \cdot x_1 + (1 - t) \cdot x_0$

Predict $S'_i \leftarrow \text{AlphaFold}(A_i, M_i, x_t, t)$

AlphaFlow Training



Input: Training examples of structures, sequences, and MSAs $\{(S_i, A_i, M_i)\}$

for all (S_i, A_i, M_i) **do**

Extract $x_1 \leftarrow \text{BetaCarbons}(S_i)$

Sample $x_0 \sim \text{HarmonicPrior}(\text{length}(A_i))$

Align $x_0 \leftarrow \text{RMSDAlign}(x_0, x_1)$

Sample $t \sim \text{Uniform}[0, 1]$

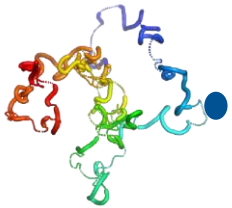
Interpolate $x_t \leftarrow t \cdot x_1 + (1 - t) \cdot x_0$

Predict $S'_i \leftarrow \text{AlphaFold}(A_i, M_i, x_t, t)$

Optimize loss $L = \text{FAPE}^2(S'_i, S_i)$

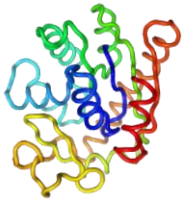
AlphaFlow Inference

Prior



x_0

Data

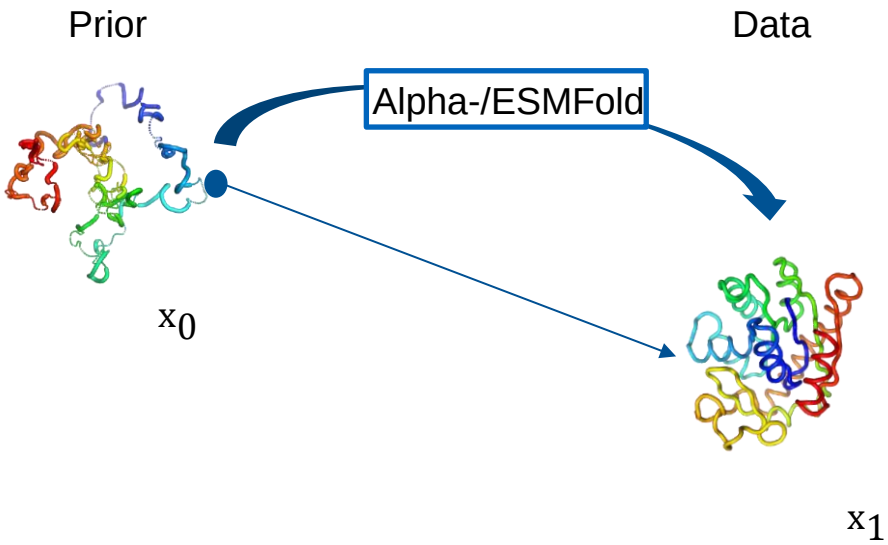


x_1

Input: Sequence and MSA (A, M)
Output: Sampled all-atom structure S‘

$\text{Sample } x_0 \sim \text{HarmonicPrior}(\text{length}(A))$

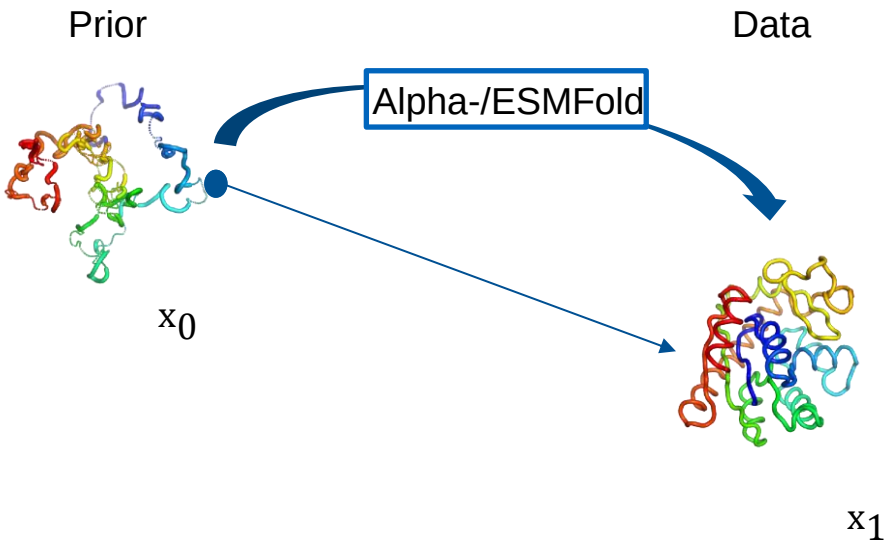
AlphaFlow Inference



Input: Sequence and MSA (A, M)
Output: Sampled all-atom structure S‘

Sample $x_0 \sim \text{HarmonicPrior}(\text{length}(A))$
for n ← 0 to N − 1 **do**
Let $t \leftarrow n/N$ and $s \leftarrow t + 1/N$
Predict $S' \leftarrow \text{AlphaFold}(A, M, x_t, t)$

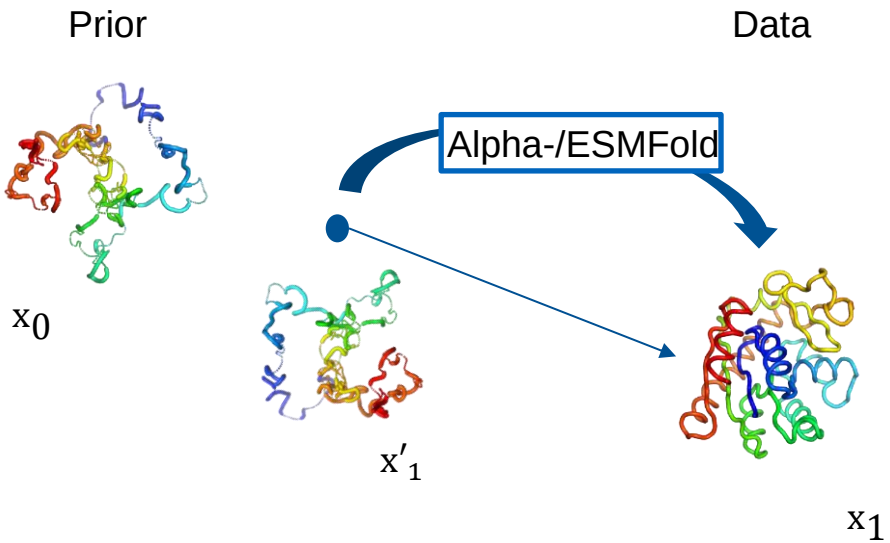
AlphaFlow Inference



Input: Sequence and MSA (A, M)
Output: Sampled all-atom structure S‘

Sample $x_0 \sim \text{HarmonicPrior}(\text{length}(A))$
for $n \leftarrow 0$ to $N - 1$ **do**
  Let $t \leftarrow n/N$ and $s \leftarrow t + 1/N$
  Predict $S' \leftarrow \text{AlphaFold}(A, M, x_t, t)$
  if $n = N - 1$ **then**
    return S'
  end if
  Extract $x'_1 \leftarrow \text{BetaCarbons}(S')$
  Align $x_t \leftarrow \text{RMSDAlign}(x_t, x'_1)$

AlphaFlow Inference



Input: Sequence and MSA (A, M)
Output: Sampled all-atom structure S‘

Sample $x_0 \sim \text{HarmonicPrior}(\text{length}(A))$
for $n \leftarrow 0$ to $N-1$ **do**
  Let $t \leftarrow n/N$ and $s \leftarrow t + 1/N$
  Predict $S' \leftarrow \text{AlphaFold}(A, M, x_t, t)$
  if $n = N-1$ **then**
    return S'
  end if
  Extract $x'_1 \leftarrow \text{BetaCarbons}(S')$
  Align $x_t \leftarrow \text{RMSDAlign}(x_t, x'_1)$
  Interpol. $x_s \leftarrow (s-t)/(1-t) \cdot x'_1 + (1-s)/(1-t) \cdot x_t$

AlphaFlow Inference

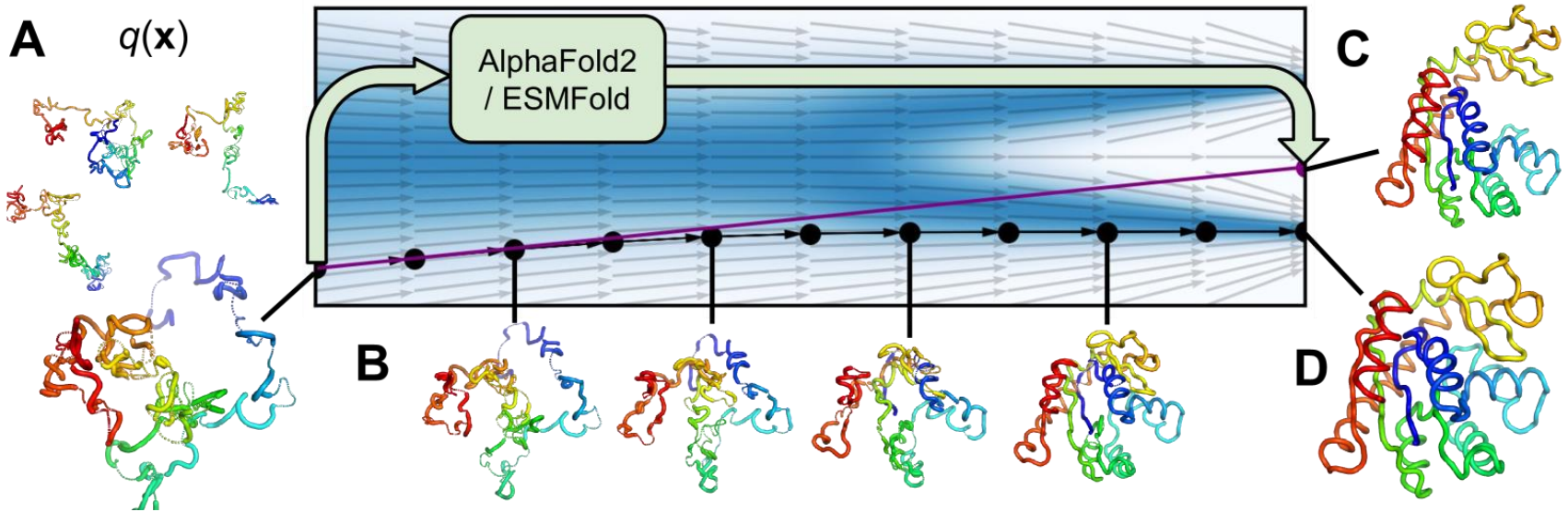
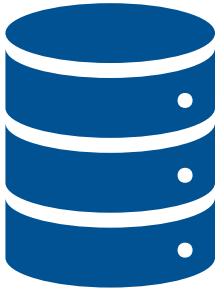
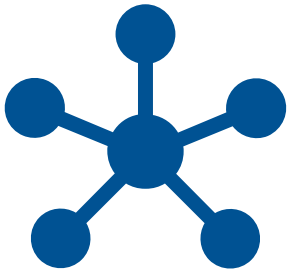


Fig. 11: Flow Matching with AlphaFold step-by-step.

Conformational States vs. Molecular Dynamics



Conformational states from Protein Data Bank (PDB)



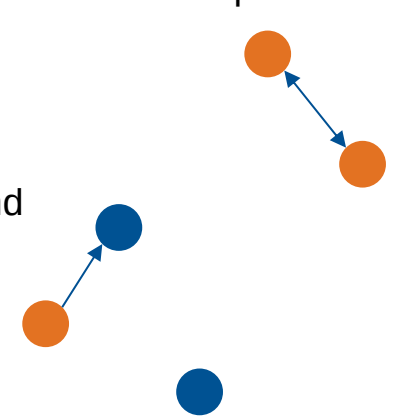
Molecular Dynamics (MD) data

PDB Metrics

Diversity
1 – lDDT between predictions

Recall
lDDT between PDB sample and nearest prediction

Precision
lDDT between prediction and nearest PDB sample

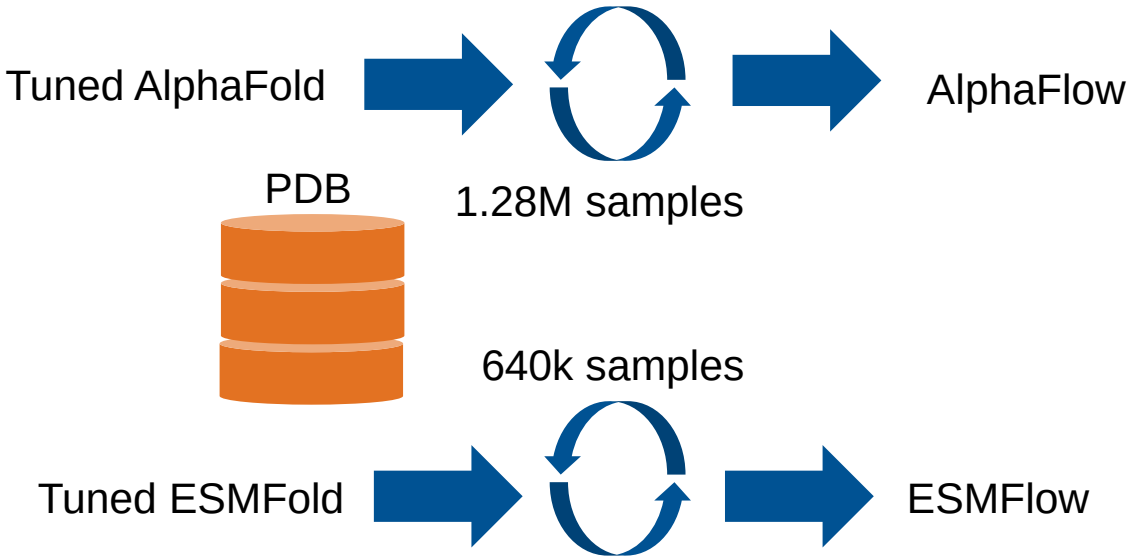


IDDT: Local Distance Difference Test

Predicted Ensemble

PDB Ensemble

PDB Ensembles



PDB Ensembles

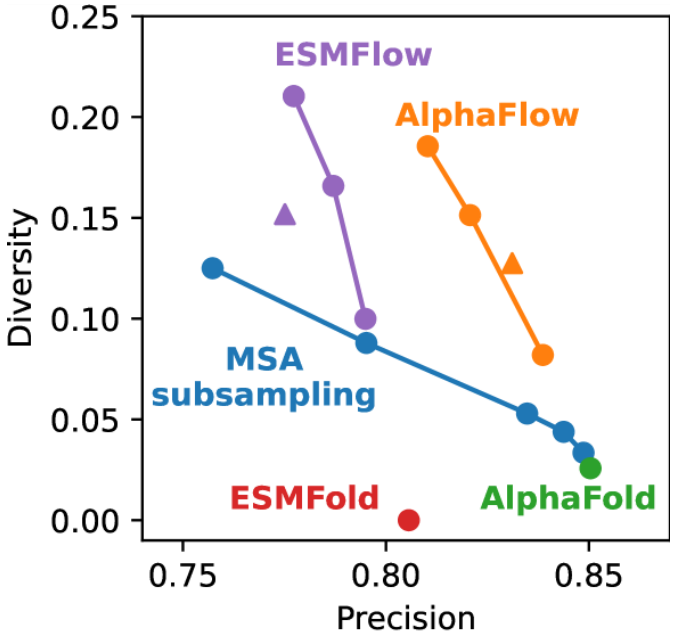


Fig. 12: Comparison of Diversity w.r.t. Precision

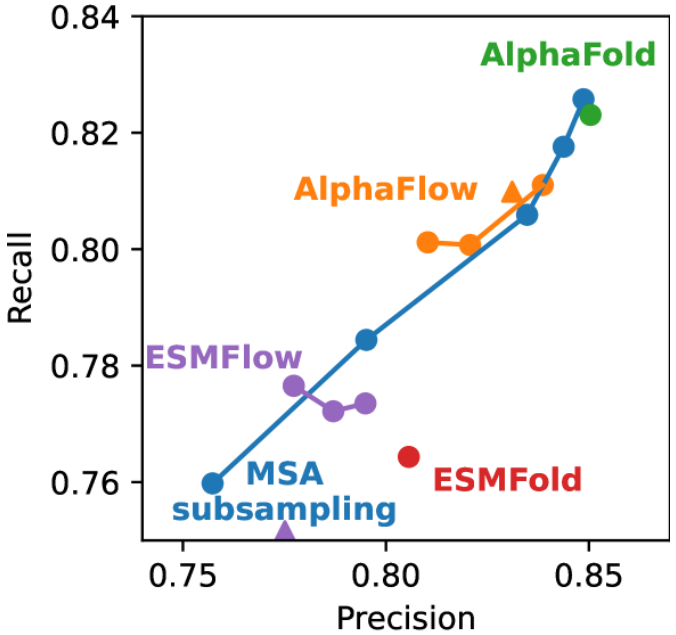


Fig. 13: Comparison of Recall w.r.t. Precision

MD Ensembles

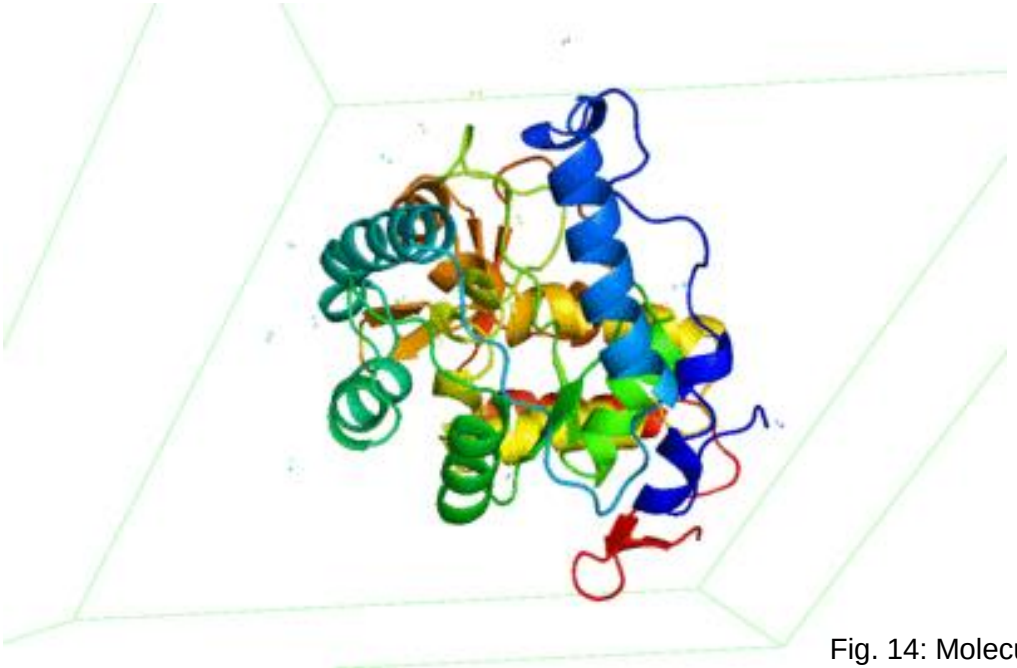


Fig. 14: Molecular dynamics simulation

MD Ensembles

Templates

- Protein sequence
- Single protein structure

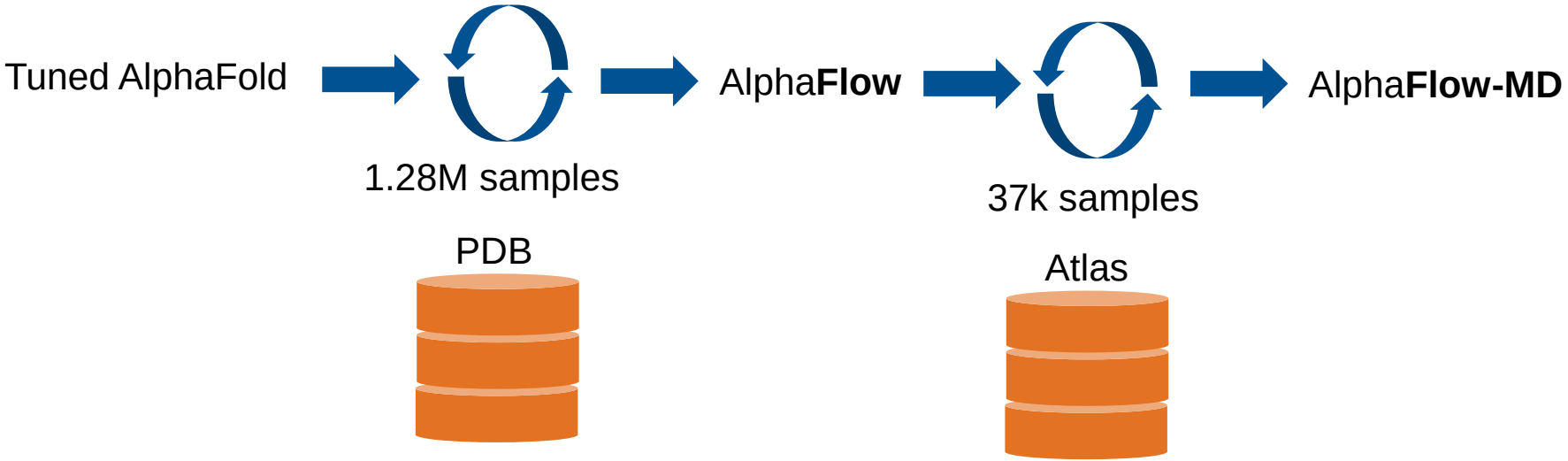
- Distribution of protein ensembles

Without templates

- Protein sequence

- Distribution of protein ensembles

MD Ensembles



MD Ensembles: Predicting Flexibility

	AlphaFlow-MD		MSA subsampling				AlphaFold	AFMD+Templates	
	Full	Distilled	32	48	64	256		Full	Distilled
Pairwise RMSD (= 2.90)	**2.89**	1.94	4.40	2.34	1.67	0.72	0.58	2.18	1.73
Pairwise RMSD r ↑	0.48	0.48	0.03	0.12	0.22	0.15	0.10	**0.94**	0.92
All-atom RMSF (=1.70)	**1.68**	1.28	5.38	2.29	1.17	0.49	0.31	1.31	1.00
Global RMSF r	0.60	0.54	0.13	0.23	0.29	0.26	0.21	**0.91**	0.89
Per-target RMSF r	0.85	0.81	0.51	0.52	0.51	0.55	0.52	**0.90**	0.88

Table 1: Comparing performance based on flexibility

MD Ensembles: Distributional Accuracy

	AlphaFlow-MD		MSA subsampling				AlphaFold	AFMD+Templates	
	Full	Distilled	32	48	64	256		Full	Distilled
Root mean $\mathcal{W}2$-dist. ↓	2.61	3.70	6.15	5.32	4.28	3.62	3.58	**1.95**	2.18
↪ Translation contrib. ↓	2.28	3.10	5.22	3.92	3.33	2.87	2.86	**1.64**	1.74
↪ Variance contrib. ↓	1.30	1.52	3.55	2.49	2.24	2.24	2.27	**1.01**	1.25
MD PCA $\mathcal{W}2$-dist. ↓	1.52	1.73	2.44	2.30	2.23	1.88	1.99	**1.25**	1.41
Joint PCA $\mathcal{W}2$-dist. ↓	2.25	3.05	5.51	4.51	3.57	3.02	2.86	**1.58**	1.68

Table 2: Comparing performance based on accuracy

MD Ensembles: Ensemble Observables

	AlphaFlow-MD		MSA subsampling				AlphaFold	AFMD+Templates	
	Full	Distilled	32	48	64	256		Full	Distilled
Weak contacts J ↑	**0.62**	0.52	0.40	0.40	0.37	0.30	0.27	0.62	0.51
Transient contacts J ↑	0.41	0.28	0.23	0.26	0.27	0.27	0.28	**0.47**	0.42
Exposed residue J ↑	**0.50**	0.48	0.34	0.37	0.37	0.33	0.32	0.50	0.47
Exposed MI matrix ρ ↑	**0.25**	0.14	0.14	0.11	0.10	0.06	0.02	0.25	0.18

Table 3: Comparing performance based on ensemble observables

MD Ensembles: Computation Time

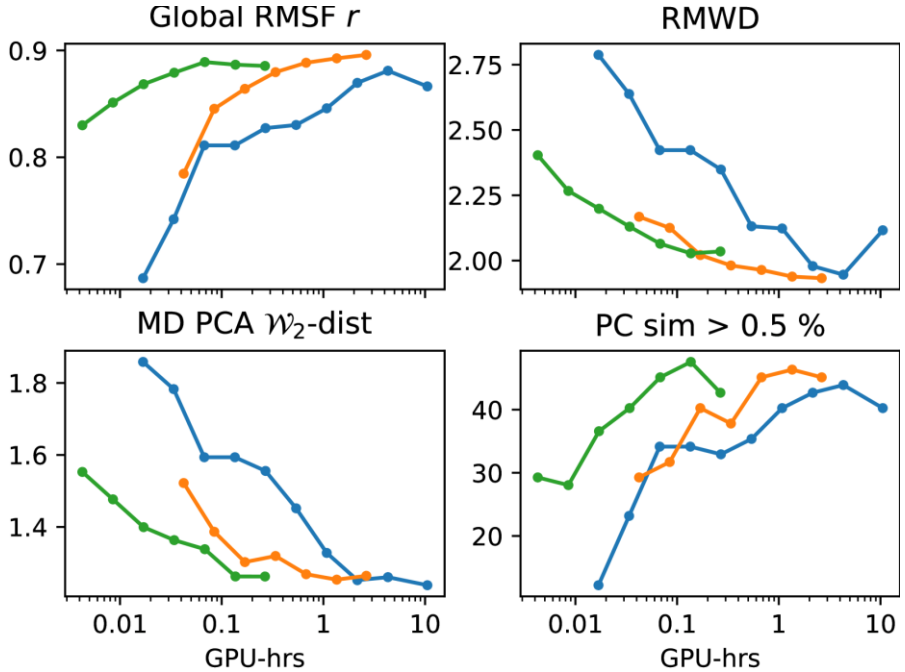
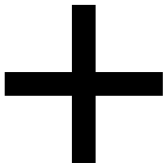


Fig. 15: Metrics w.r.t. Computation Time

Strengths

- Increase in diversity on static (PDB) and accuracy of dynamic ensembles (MD)
- Generalization to unseen proteins
- Computationally faster than MD
- Propose conformational structures (further evaluation via MD)
- Bridges gap between static structure prediction and dynamic nature of proteins

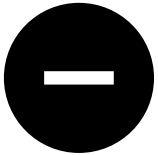


Limitations (Can not…)

- Generate dynamics

→ Match long MD simulations

- Sample energy landscape of the protein



Conclusion

“AlphaFlow accelerates one specific application of MD
but does it in a transferable manner.“, Bowen Jing

References

- Conformational action with light: https://www.aps.anl.gov/APS-Science-Highlight/2014/Lights-Conformational-Change-Action
- Flow Marching: https://doi.org/10.48550/arXiv.2210.02747
- AlphaFold Meets Flow Matching for Generating Protein Ensembles: https://doi.org/10.48550/arXiv.2402.04845
- AlphaFold: https://doi.org/10.1038/s41586-021-03819-2
- MD: https://www.drugdesign.org/chapters/molecular-dynamics/#introduction
- AlphaFlow and ESMFlow on Github: https://github.com/bjing2016/alphaflow
- LDDT: https://doi.org/10.1093/bioinformatics/btt473
- Parts of this presentation were inspired by Bowen Jing‘s talk: https://portal.valencelabs.com/events/post/alphafold-meets-flow-matching-for-generating-protein-ensembles-tlWJ4oC7Xrnim0s

Figure References

1 https://foodscience-techn.blogspot.com/2014/07/structure-of-protein.html
2 https://www.researchgate.net/figure/Diagram-of-conformational-changes-during-the-transformation-of-a-natural-prion-protein_fig1_337289618
3 https://www.aps.anl.gov/APS-Science-Highlight/2014/Lights-Conformational-Change-Action
4 - 6 https://mlg.eng.cam.ac.uk/blog/2024/01/20/flow-matching.html
7 - 9 adapted, 11 – 15, table 1-3 from https://doi.org/10.48550/arXiv.2402.04845
10 adapted from https://foodscience-techn.blogspot.com/2014/07/structure-of-protein.html