Diffusion models for Protein Generation

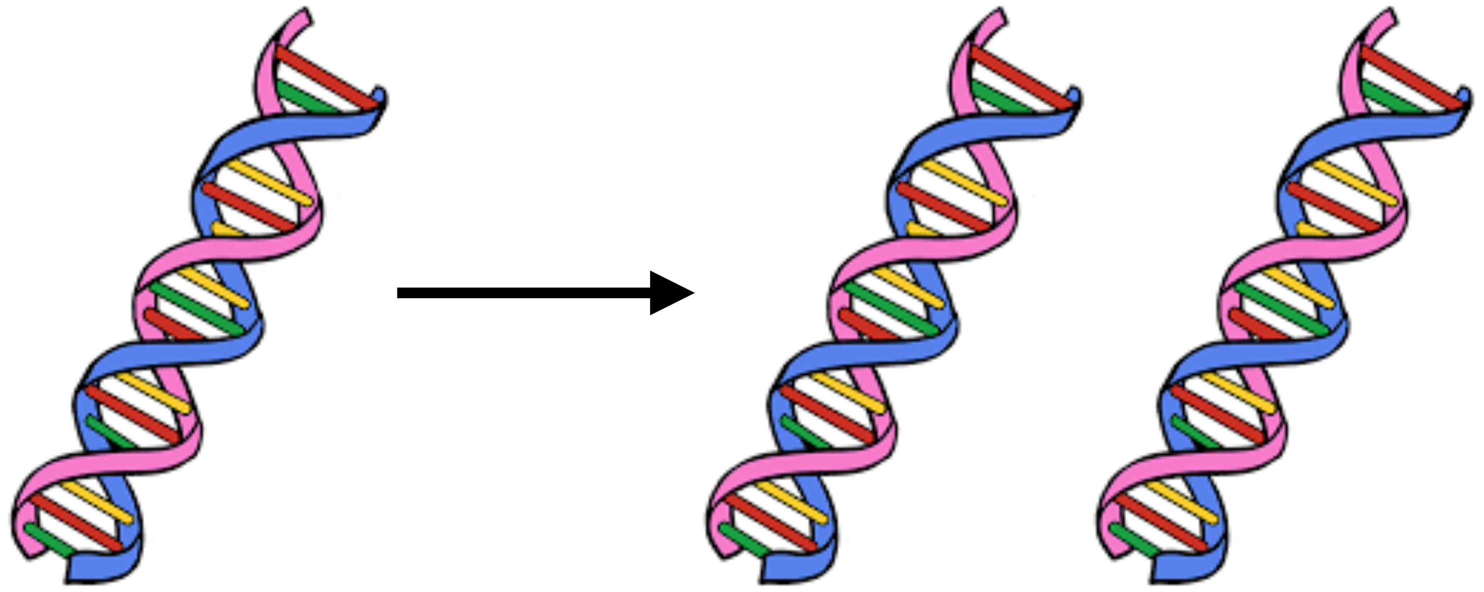
Jason Yim

MIT | JAMEEL CLINIC

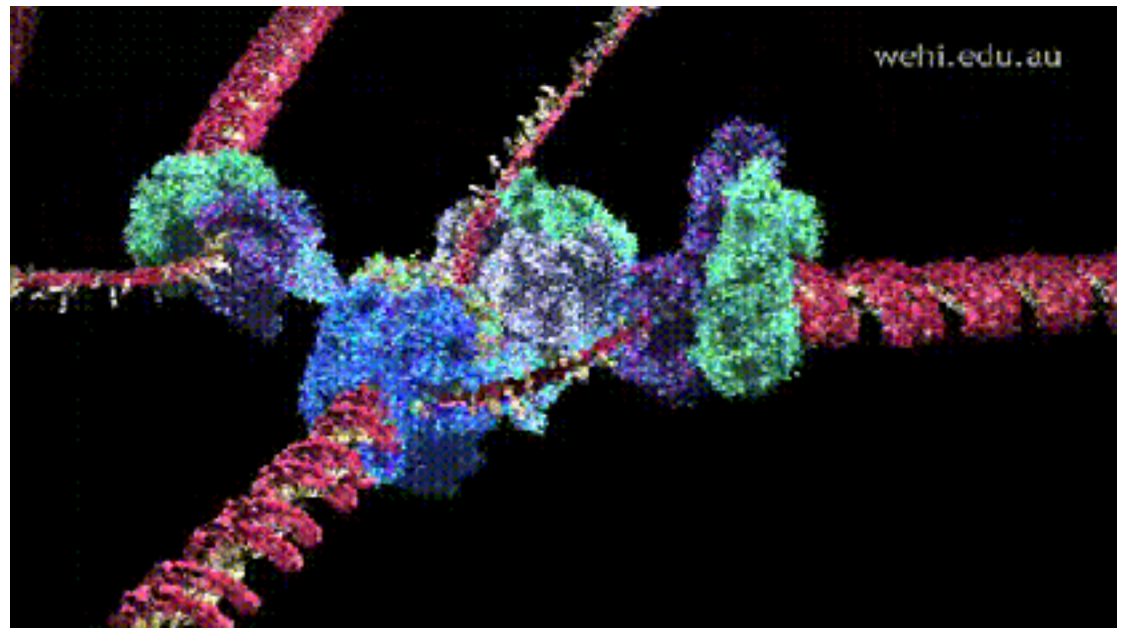
MIT CSAIL

What do proteins do?

- Nature has evolved proteins to perform necessary functions for life.



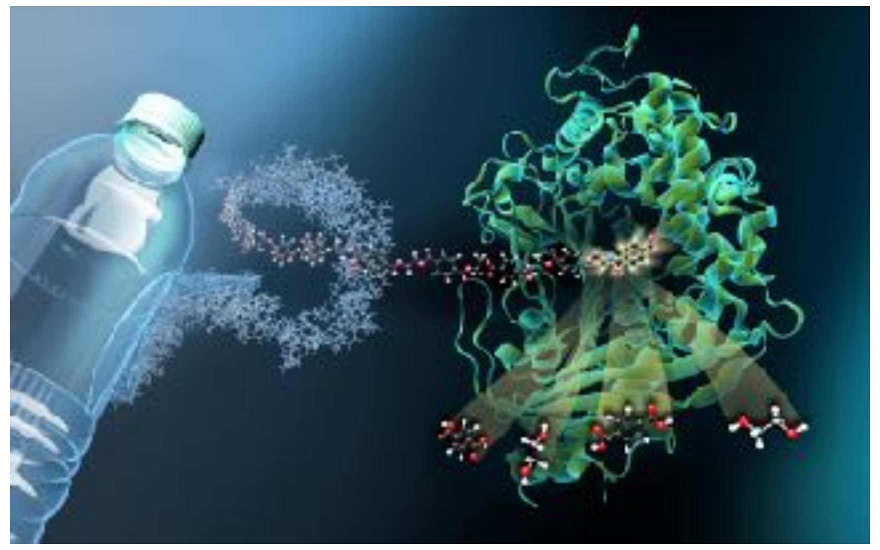
For example, DNA replication



Proteins performing DNA replication

- Humans have engineered proteins for specific needs.

Vaccine & drugs



Plastic degrading enzymes

Genome editing

Image: Marsbars via iStock

Image: Martin Künsting/HZB

Image: Amanda Heidt via The Scientist

Why AI for proteins?

Eroom's Law

(Moore's law backwards!)

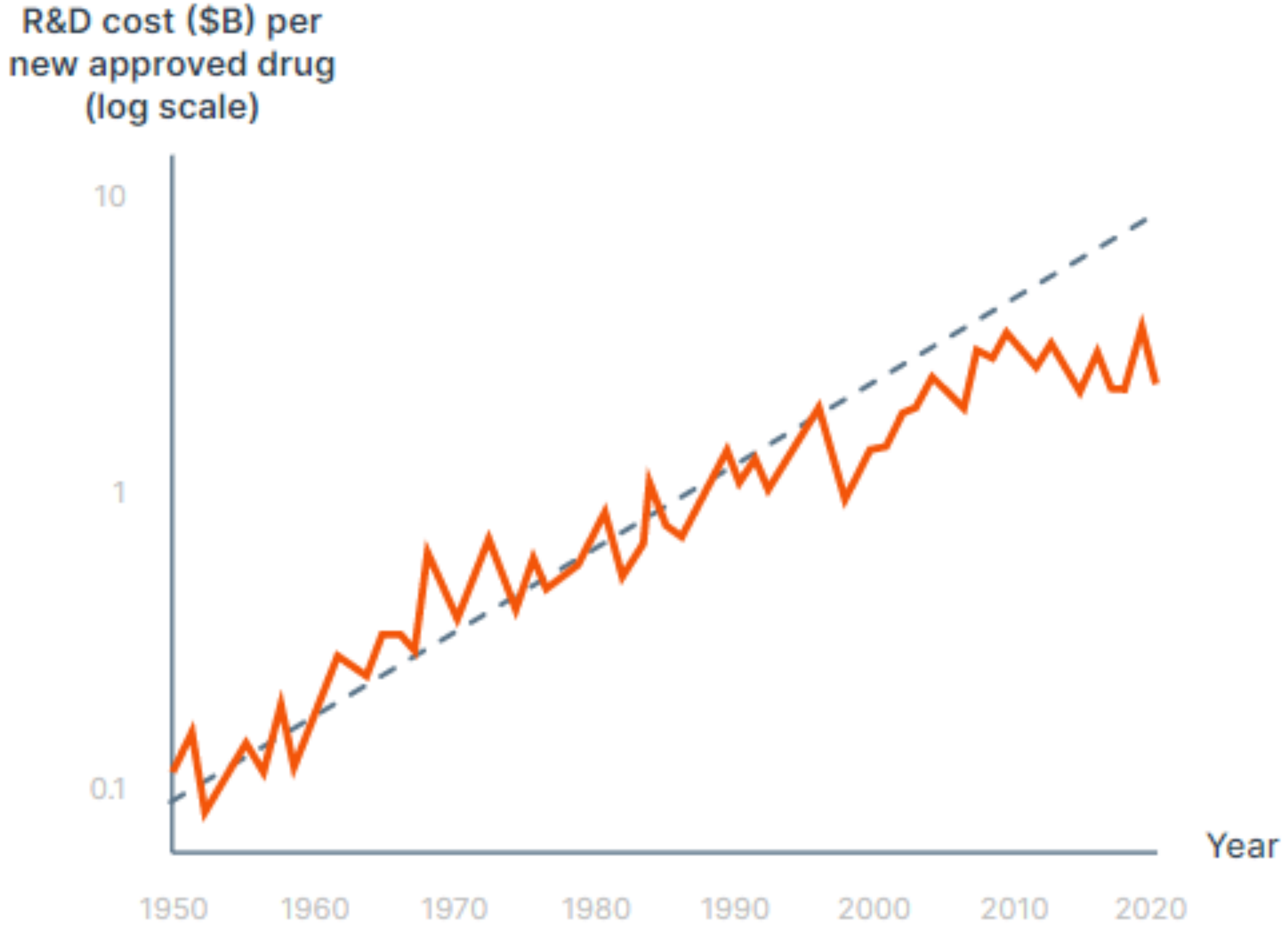
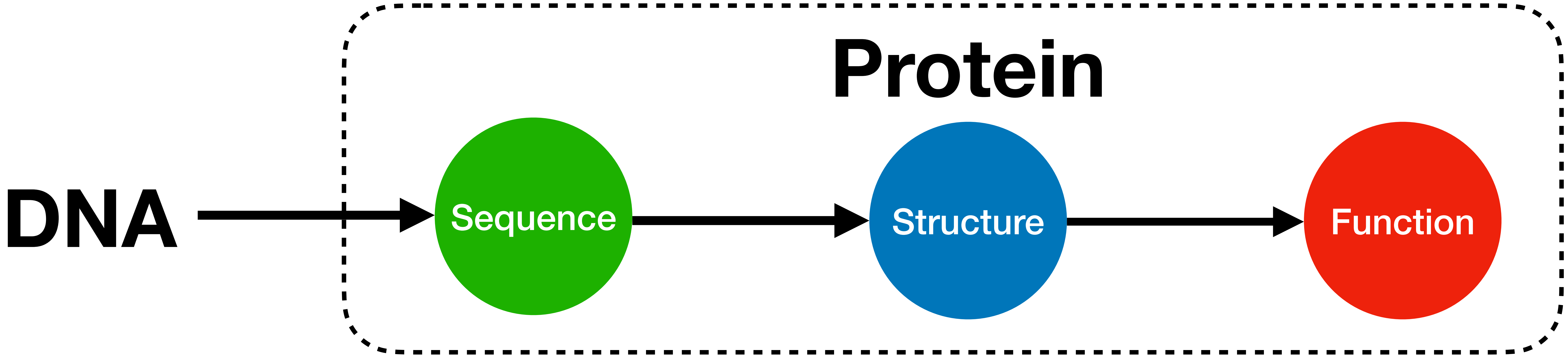
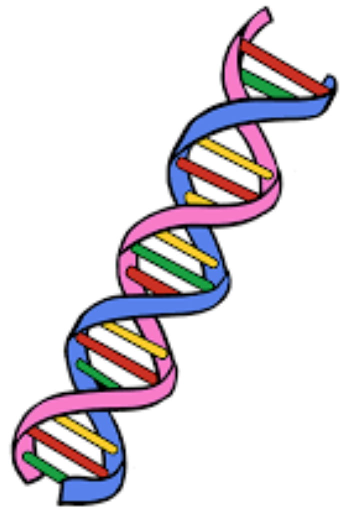


Image: Lindus Health

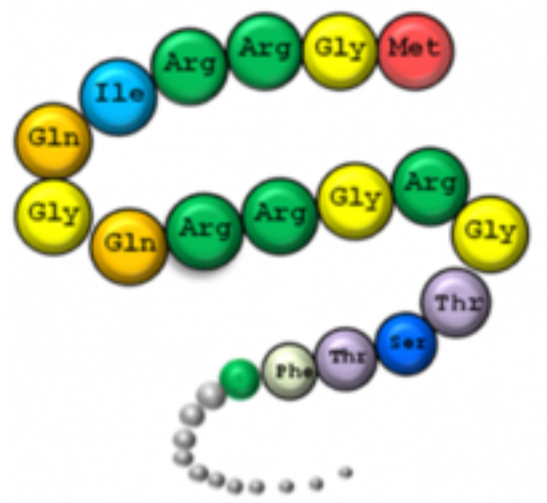
- Drug development crisis.
  - Takes **~10 years** and **~$2.6 billion** to make a single drug.
  - **Can AI accelerate this timeline?**

Protein modeling

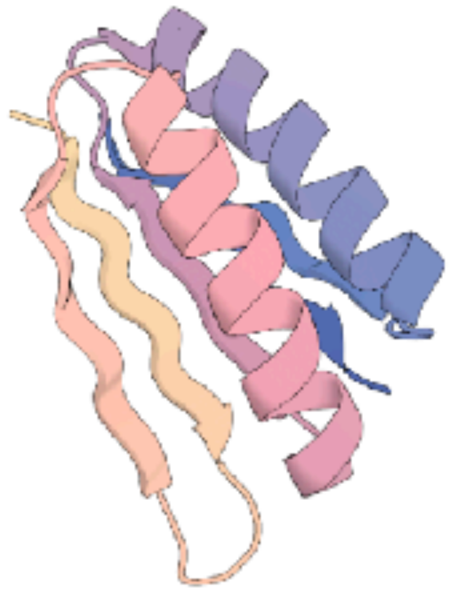




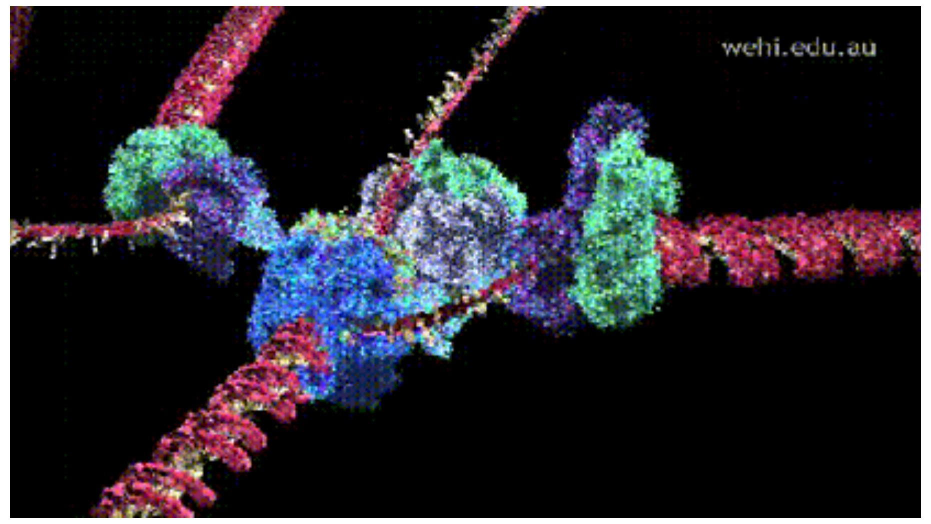
4 letter vocabulary



20 letter vocabulary



3D coordinates

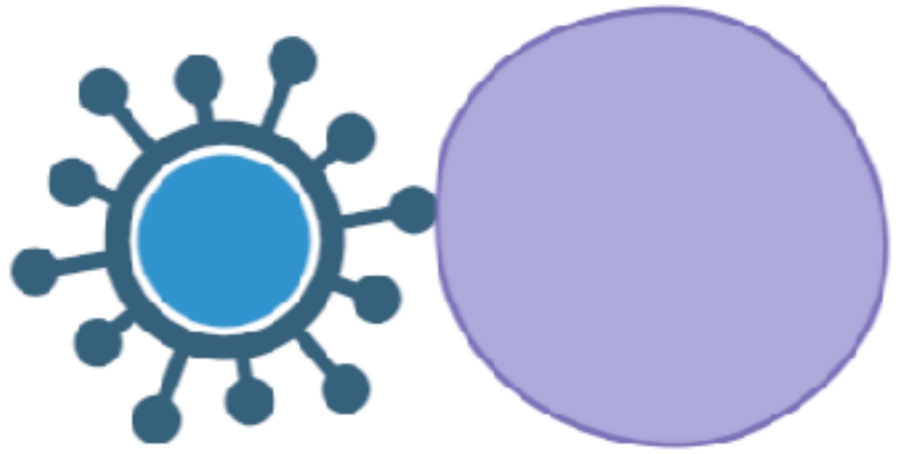


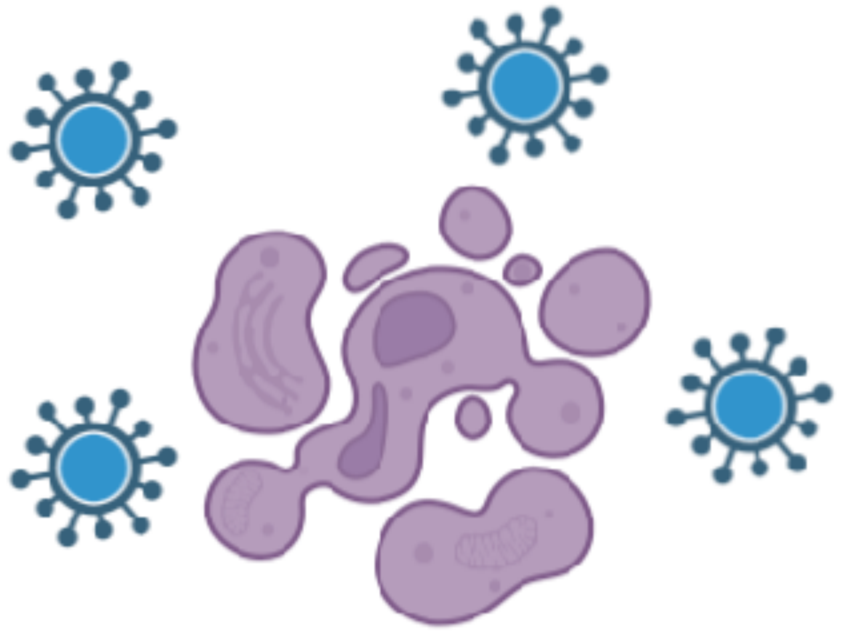
wehi.edu.au

Binding, Reactions

Protein function: simplified example

- **How do viruses work?**

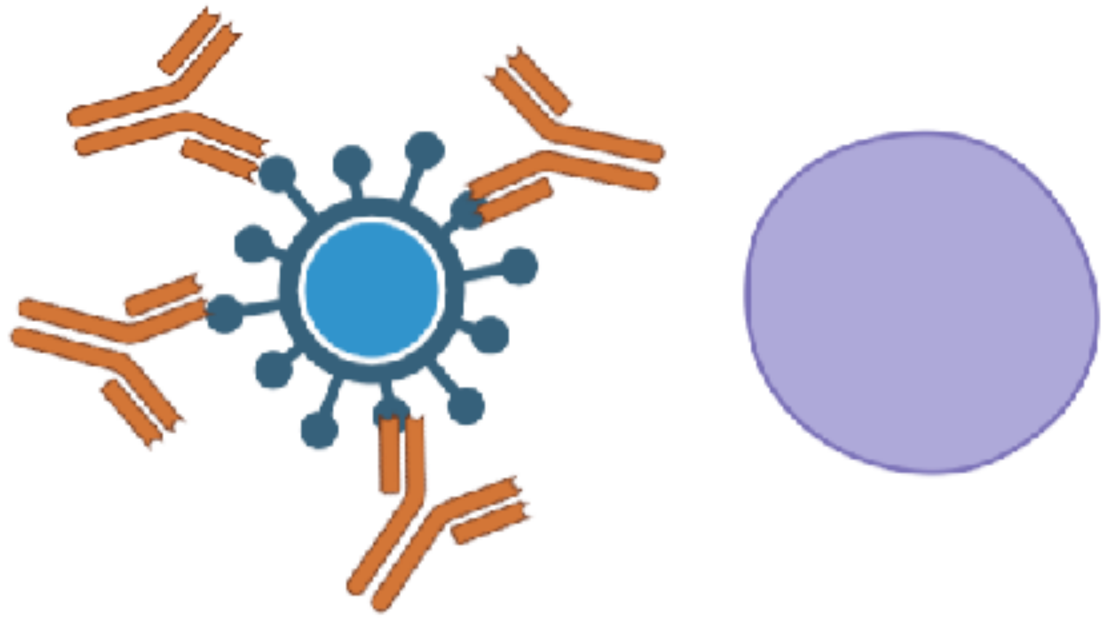




Cell death, virus replication, human gets sick

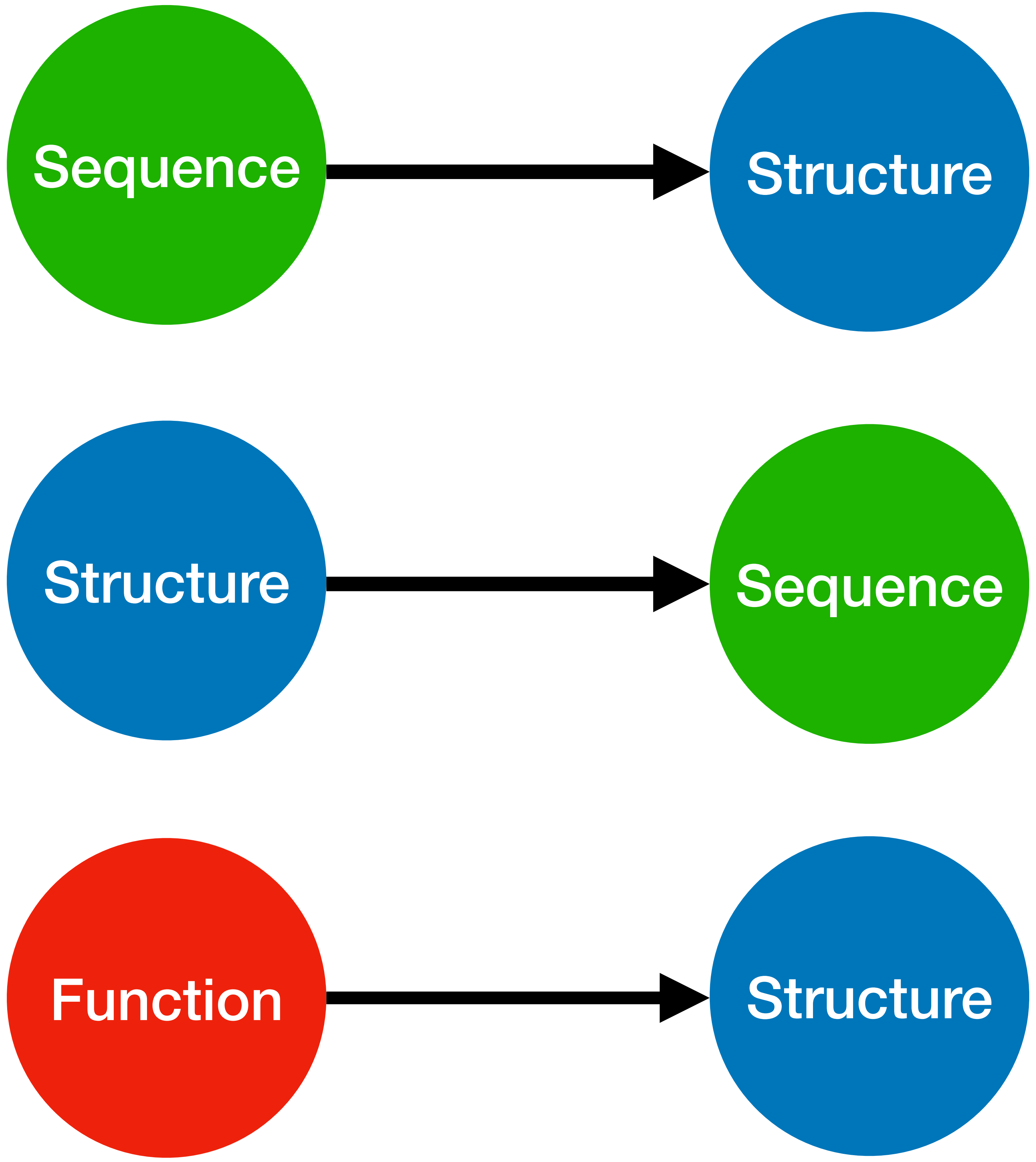
- **How does protein binding stop viruses?**

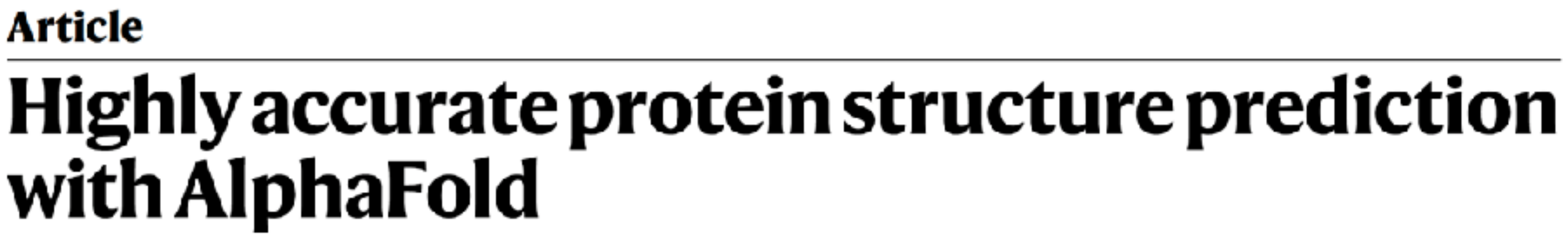
Designed antibodies



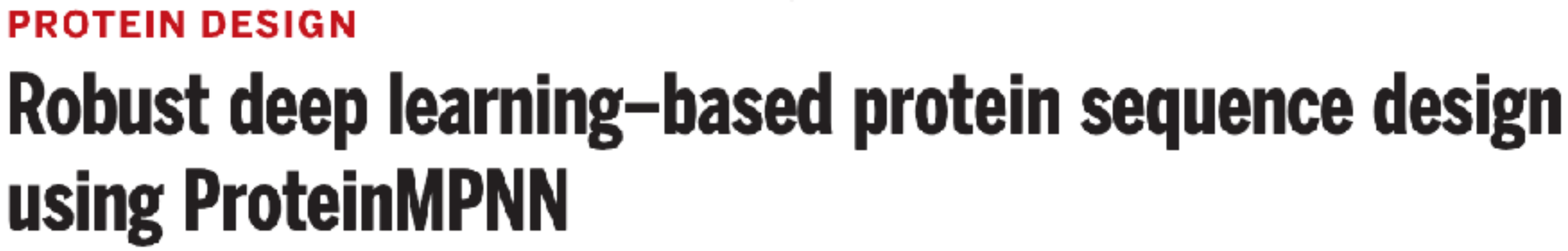
Prevent infection (antibodies have other functions as well)

Machine learning is revolutionizing protein design

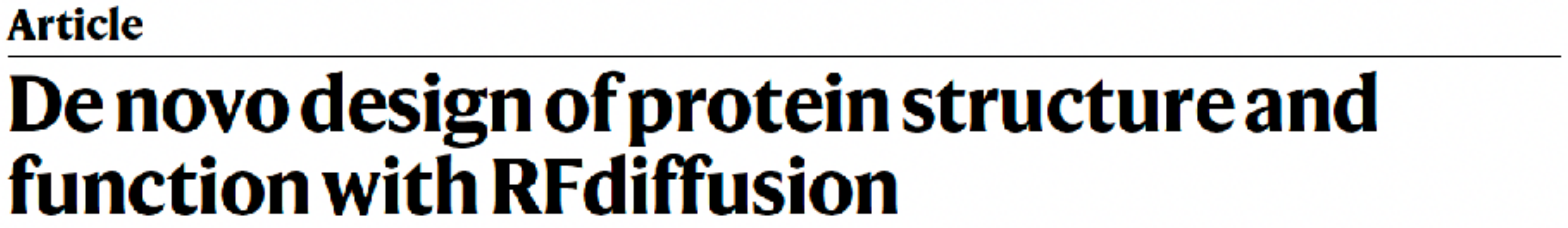



Article

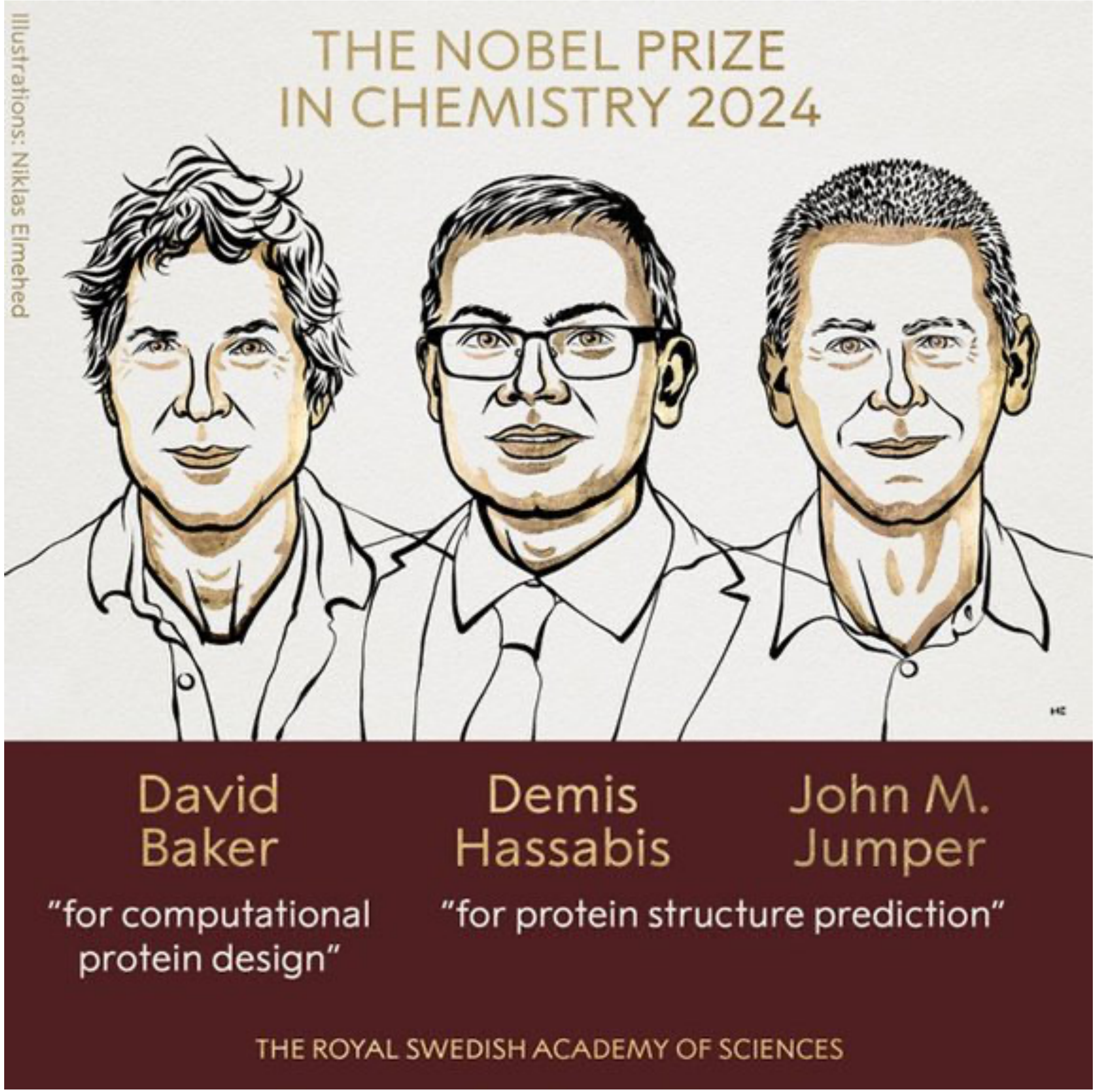
Highly accurate protein structure prediction with AlphaFold


PROTEIN DESIGN

Robust deep learning–based protein sequence design using ProteinMPNN


Article

De novo design of protein structure and function with RFdiffusion



Illustrations: Niklas Elmehed

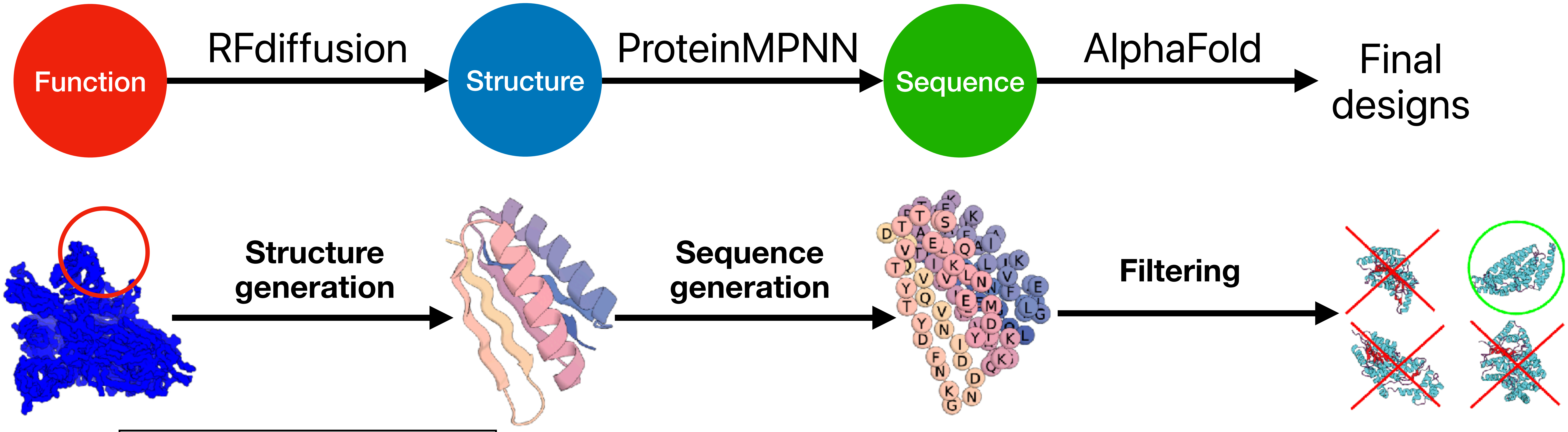
THE NOBEL PRIZE IN CHEMISTRY 2024

David Baker	Demis Hassabis	John M. Jumper
"for computational protein design"	"for protein structure prediction"	

THE ROYAL SWEDISH ACADEMY OF SCIENCES

Generative *De Novo* Protein Design

<u>Pipeline</u>



Function

RFdiffusion

Structure

ProteinMPNN

Sequence

AlphaFold

Final designs

Structure generation

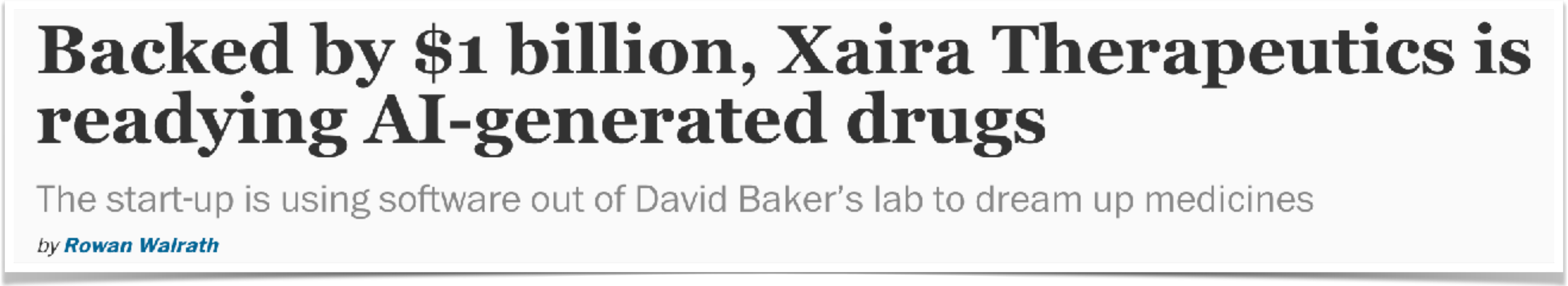
Sequence generation

Filtering

Diffusion models have enabled this step

Sequence needs to be known in order to produce the protein

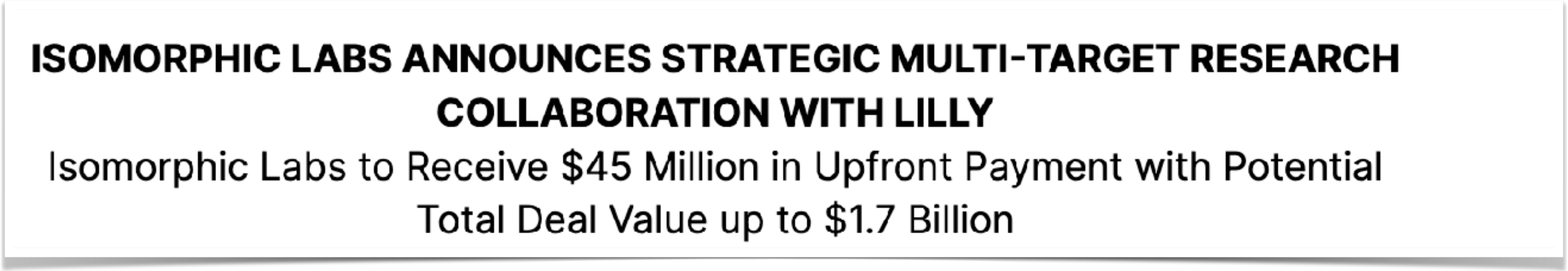
Generative AI is coming to biology



Backed by $1 billion, Xaira Therapeutics is readying AI-generated drugs

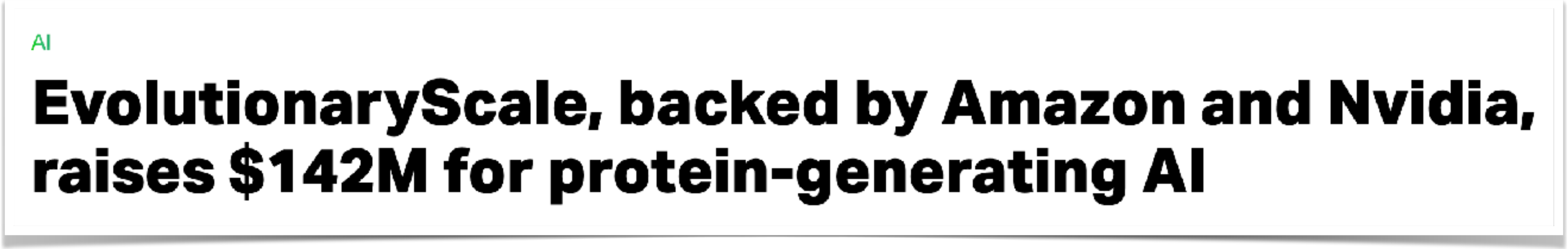
The start-up is using software out of David Baker's lab to dream up medicines

by ***Rowan Walrath***



ISOMORPHIC LABS ANNOUNCES STRATEGIC MULTI-TARGET RESEARCH COLLABORATION WITH LILLY

Isomorphic Labs to Receive $45 Million in Upfront Payment with Potential Total Deal Value up to $1.7 Billion



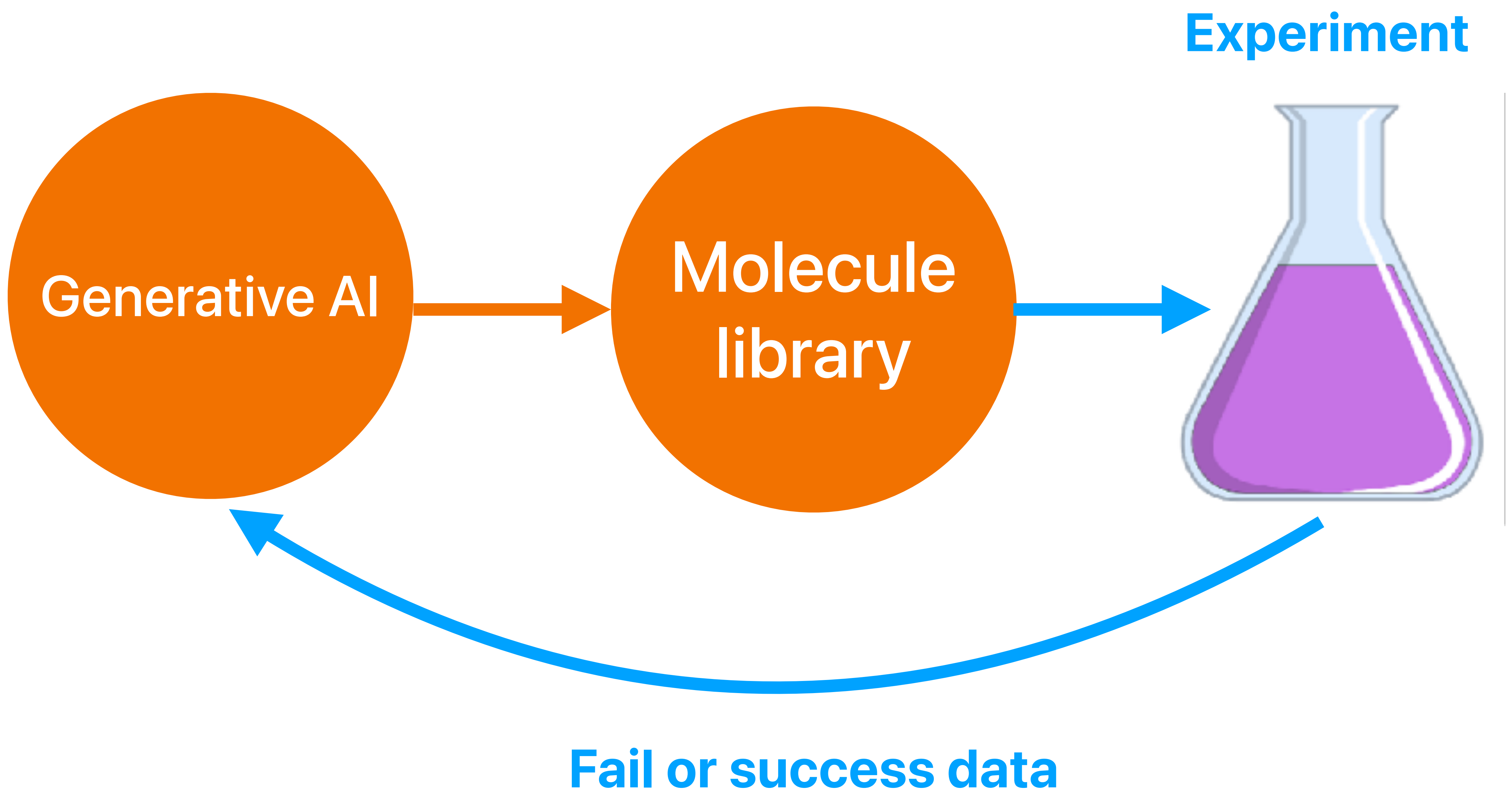
AI

EvolutionaryScale, backed by Amazon and Nvidia, raises $142M for protein-generating AI

Goal of AI in biomolecular design

Combining **generation** and **optimization** into one pipeline with AI.

1. **Generation**: Fast production of novel molecular libraries.
2. **Optimization**: Efficient fine-tuning from experiments.



Overview

1. **Protein structure generation**
    - FrameDiff [1]
2. **Generative protein design**
    - RFdiffusion [4]
3. **Co-design and sequence generation**
    - MultiFlow [3]
4. **Outlook**

References provided at end

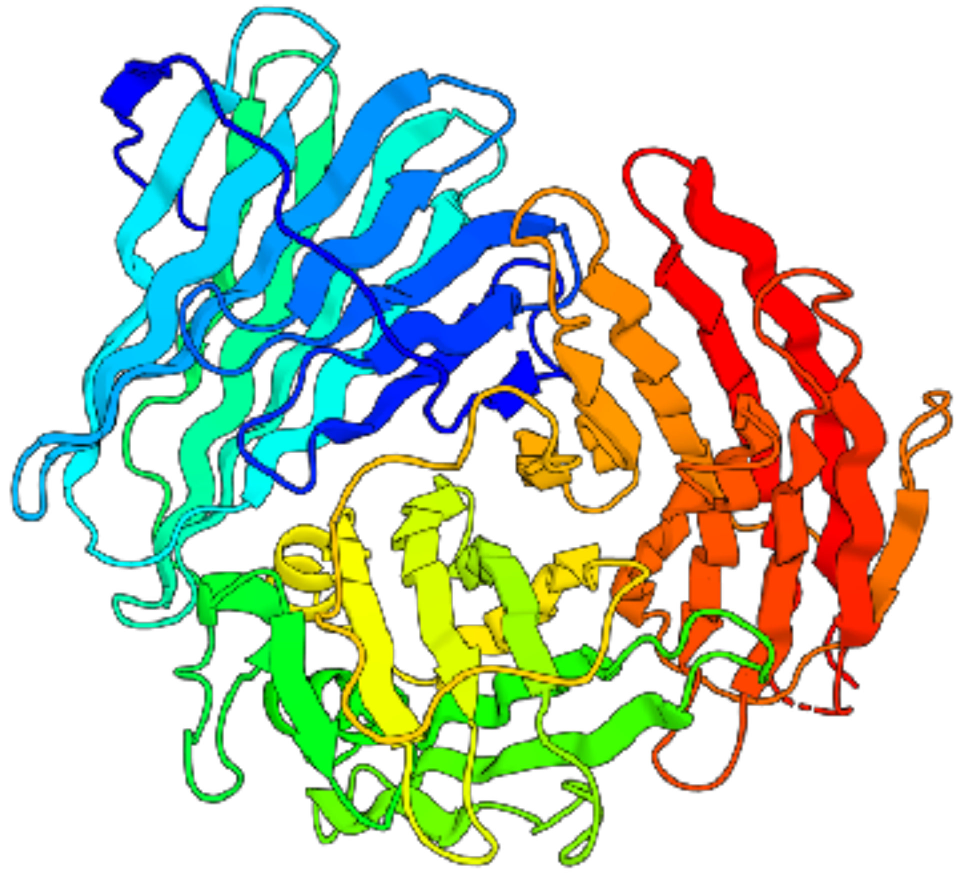
Goal: Diffusion for Protein Structure

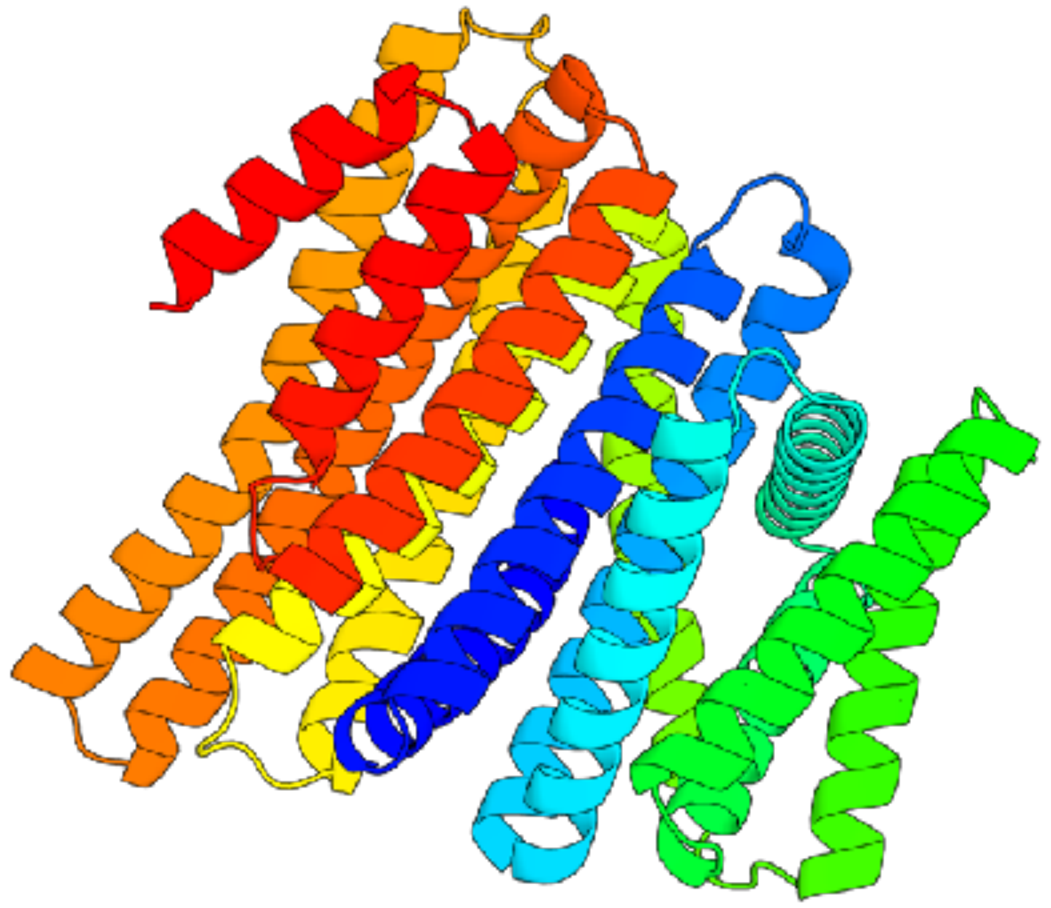
1. Generate **high quality** structures.

2. Generate **diverse** structures.

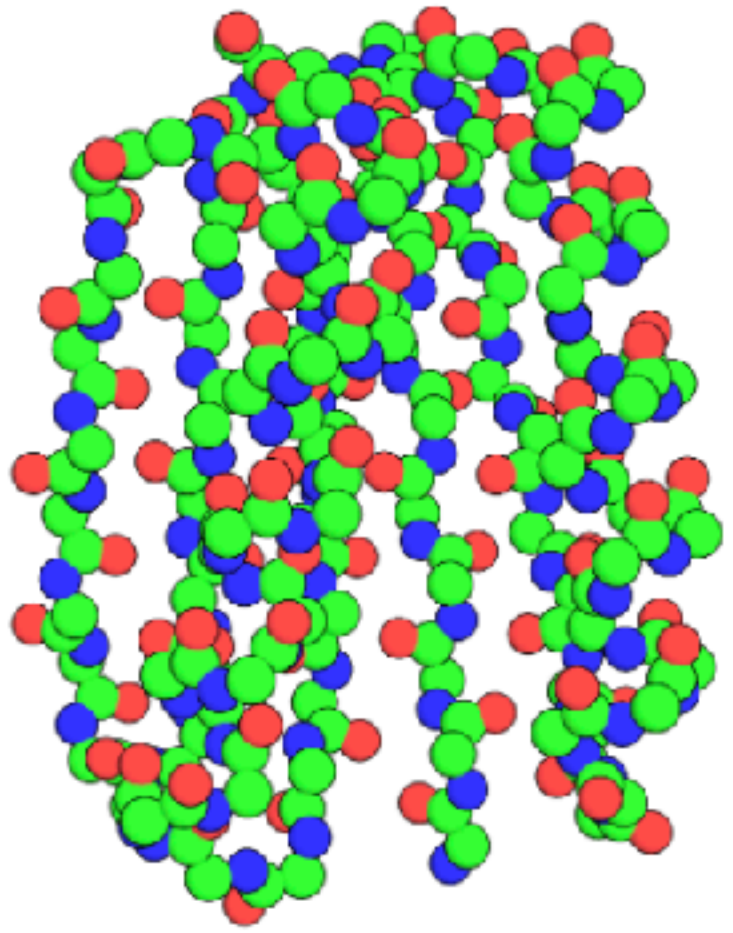
3. Generate **novel** structures.

4. Generate **functional** structures.





How to model a protein structure?



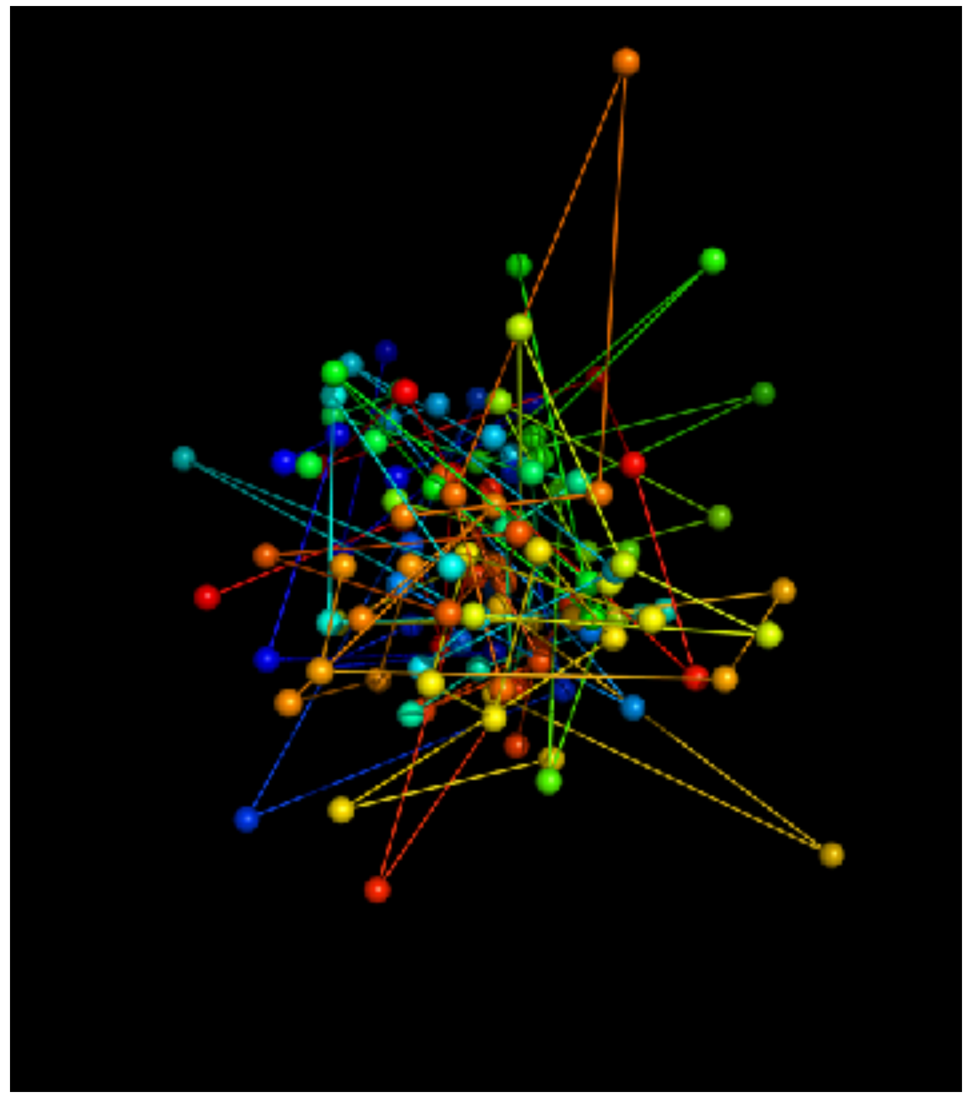
Option #1: model 3D coordinates of every atom.

- + **Precise control over atom placement.**
- - **Bonds are not fixed.**

We tried a version of this as the first step in 2022

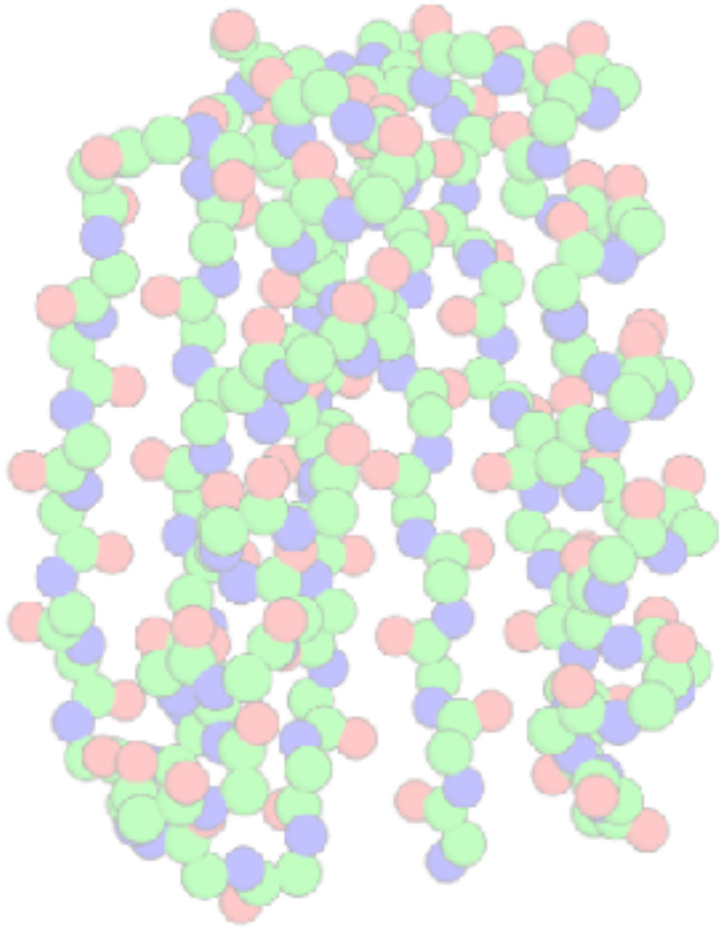
Issues:

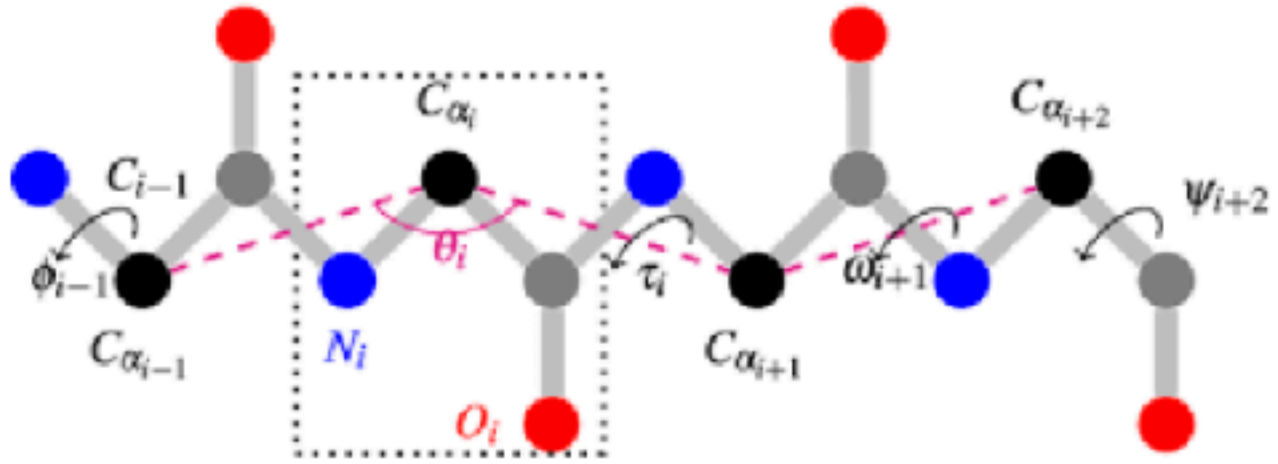
- Difficult to scale, bad performance.
- Latest works shows it is possible to scale.

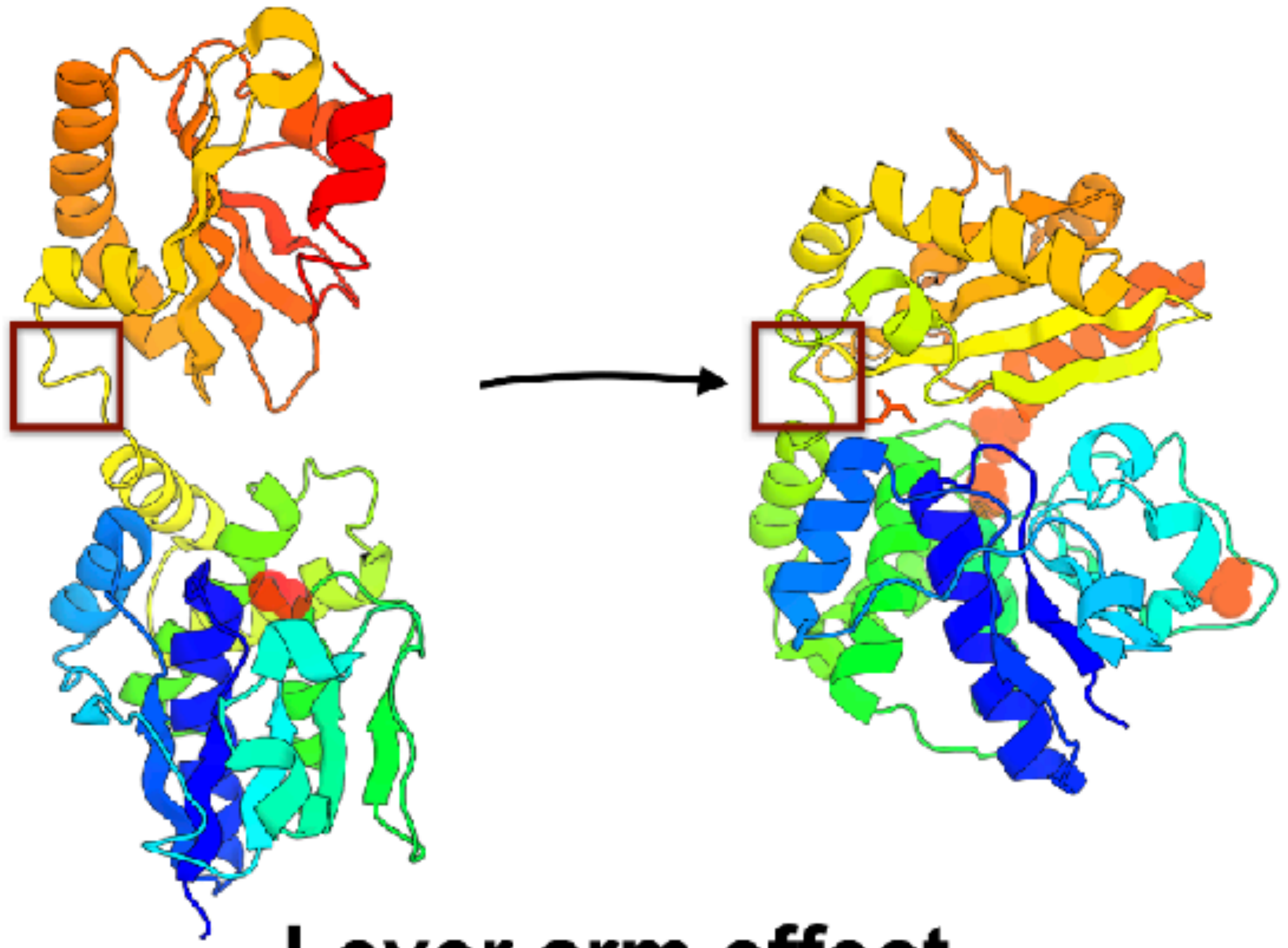


Trippe, Yim et al 2022, "Diffusion Probabilistic Modeling of Protein Backbones in 3D for the Motif-Scaffolding Problem"

How to model a protein structure?







Lever arm effect

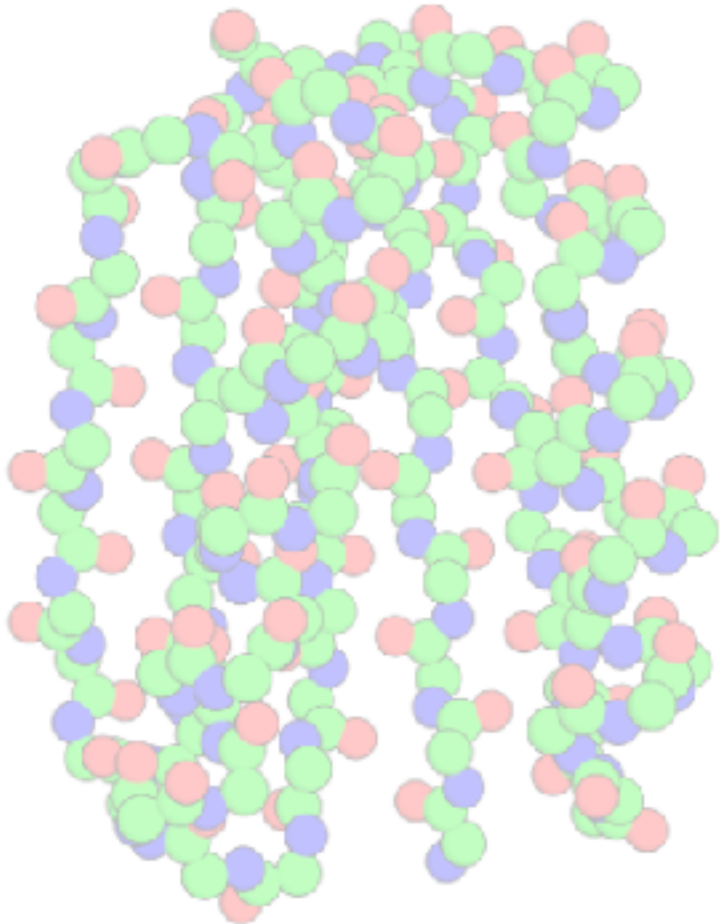
Option #1: model 3D coordinates of every atom.

- + **Precise control over atom placement.**
- - **Bonds are not fixed.**

Option #2: model only torsion angles.

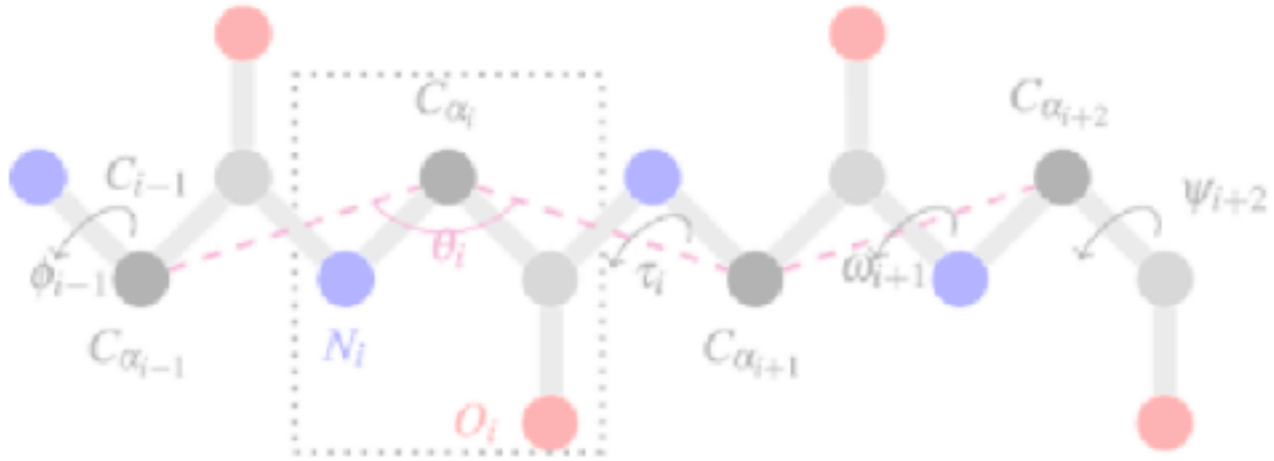
- + **Bonds are fixed.**
- - **Hard to control atom placement.**

How to model a protein structure?



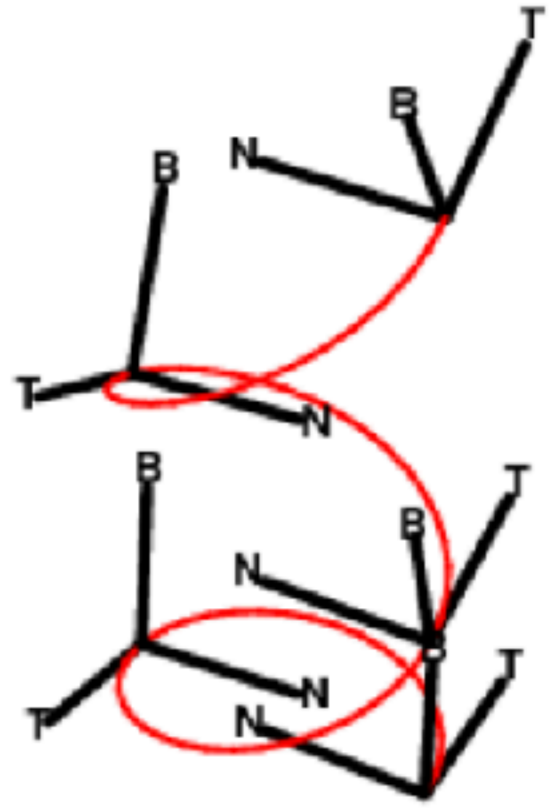
Option #1: model 3D coordinates of every atom.

- + **Precise control over atom placement.**
- - **Bonds are not fixed.**



Option #2: model only torsion angles.

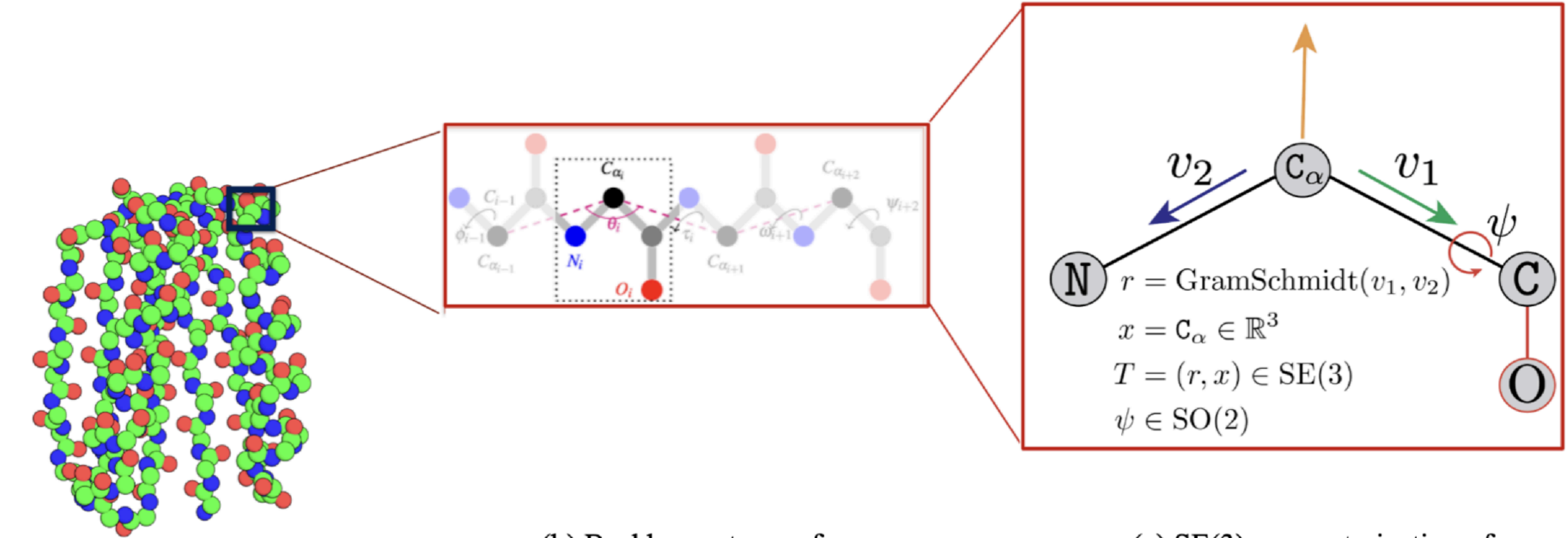
- + **Bonds are fixed.**
- - **Hard to control atom placement.**



Option #3: model with frames along a chain.

- + **3 out of 4 bonds are fixed.**
- + **Precise control over frame placement**

Background: Protein Frames



(a) Protein backbone atoms

(b) Backbone atoms of a single residue

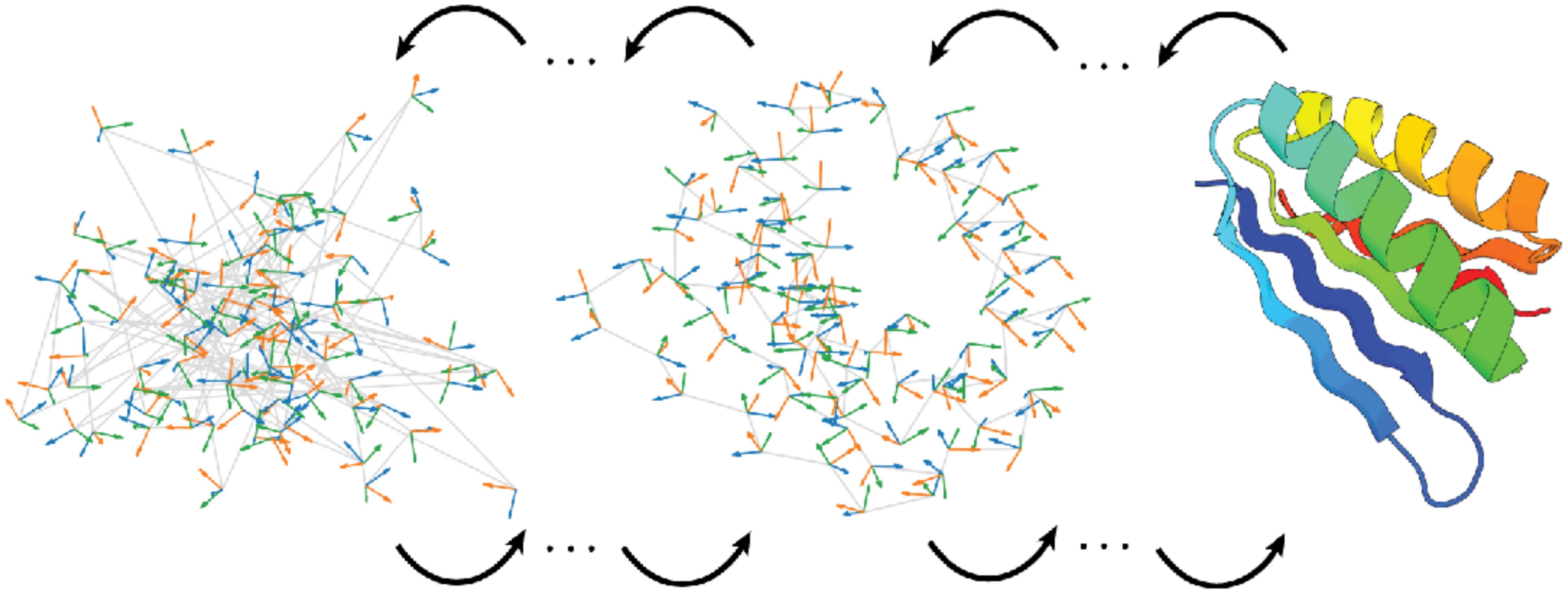
(c) SE(3) parameterization of backbone atoms of a single residue

Goal: Diffusion for Protein Frames

Forward process (noising)

$$dx = f(x,t)dt + g(t)dB$$

$p_1(x)$ (noise)



$p_0(x)$ (data)

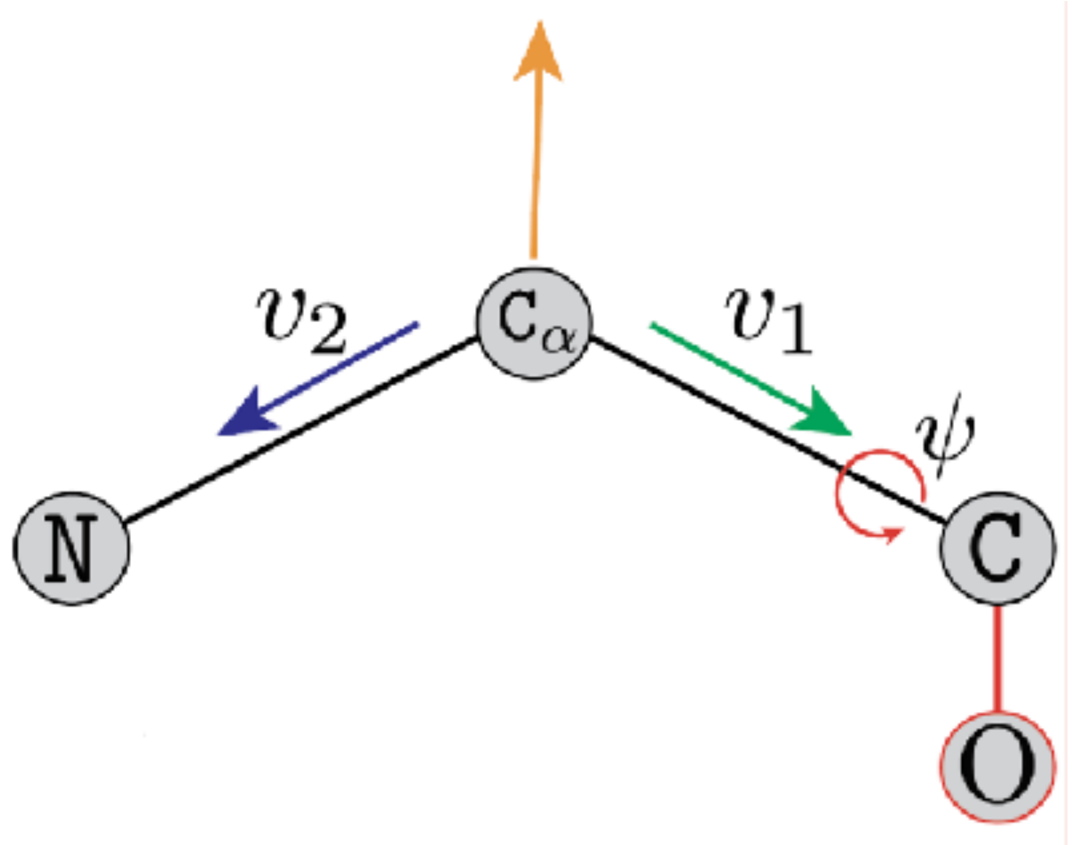
$$dx = [f(x,t) - g(t)^2 \boxed{\nabla \log p_t(x)}]dt + g(t)dB$$

Reverse process (sampling)

Learned by neural network.

Diffusion over Riemannian Manifolds

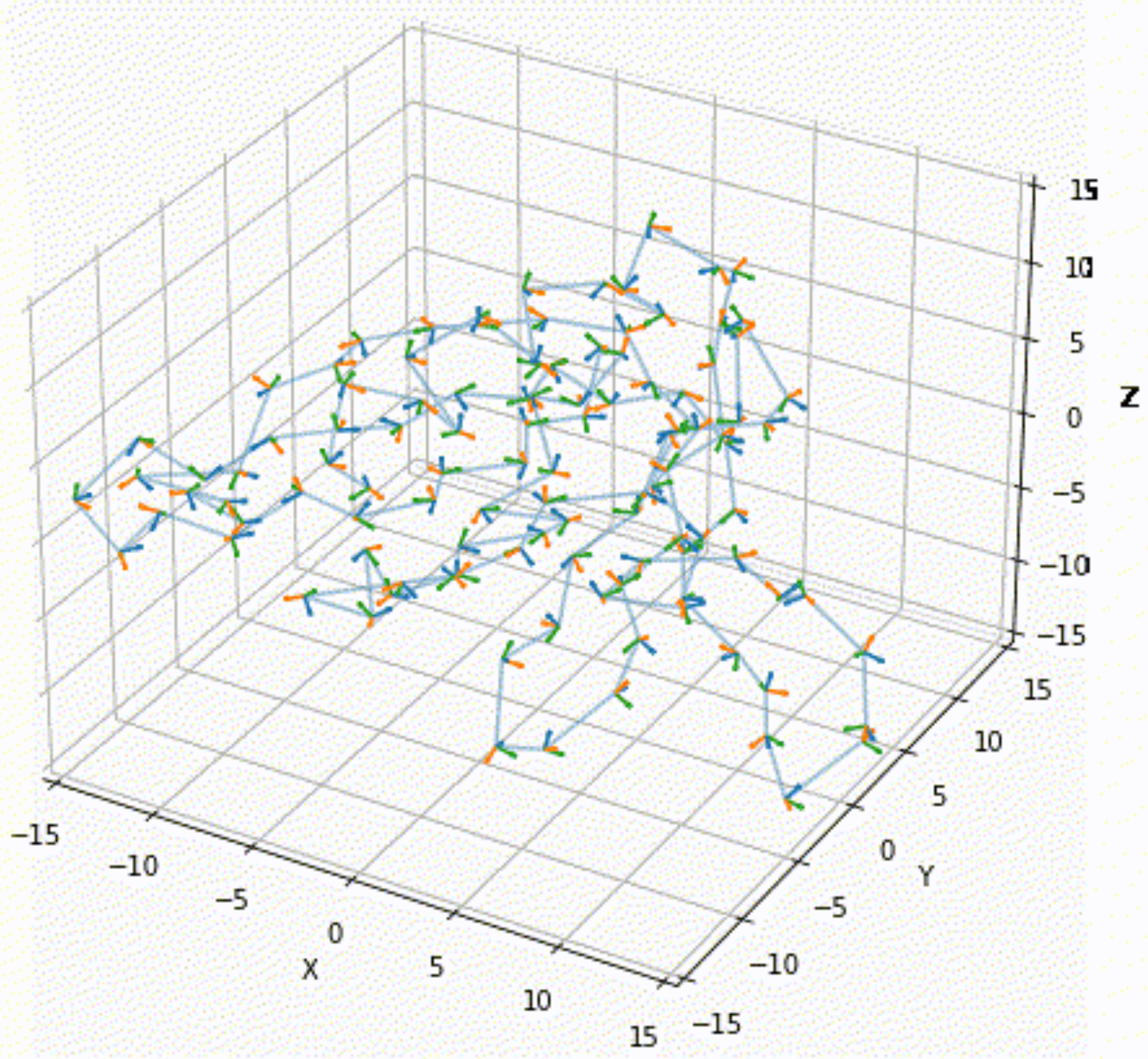
How to diffuse a frame?



Frame $(R, x) \in \mathrm{SO}(3) \times \mathbb{R}^3$

Diffuse translations $x \in \mathbb{R}^3$

Diffuse rotation $R \in \mathrm{SO}(3)$



Diffusion over Riemannian Manifolds

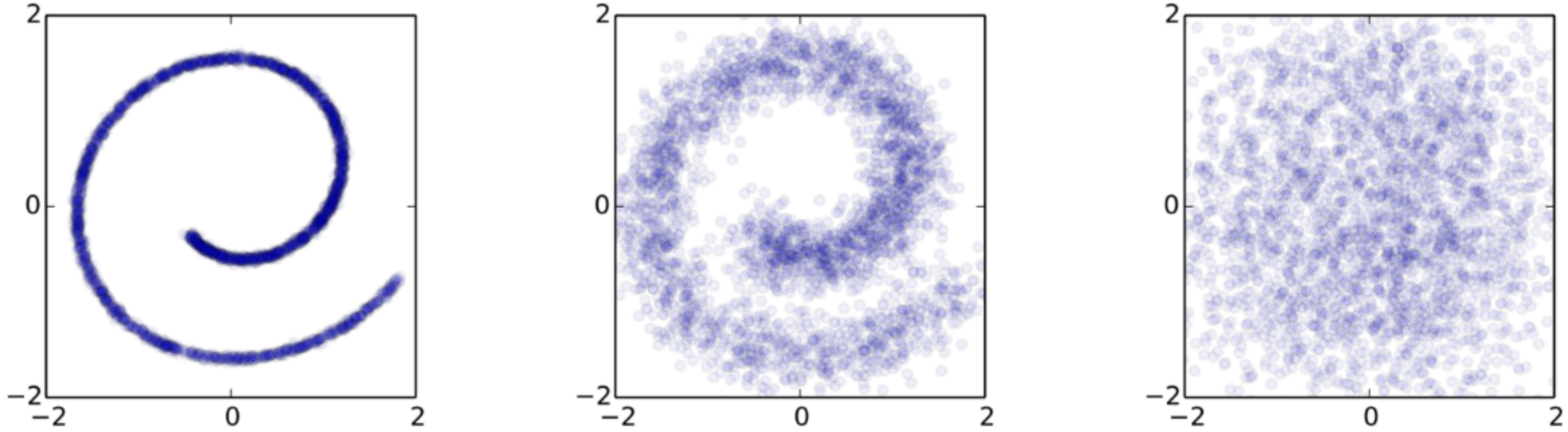
How to diffuse a frame?

(2D for visualization)

Diffuse translations $x \in \mathbb{R}^3$

Brownian motion on $\mathbb{R}^3$

$$p_{t|0}\left(x^{(t)} \,|\, x^{(0)}\right) = \mathcal{N}(x^{(t)}; \beta(t)x^{(0)}, \sigma(t))$$



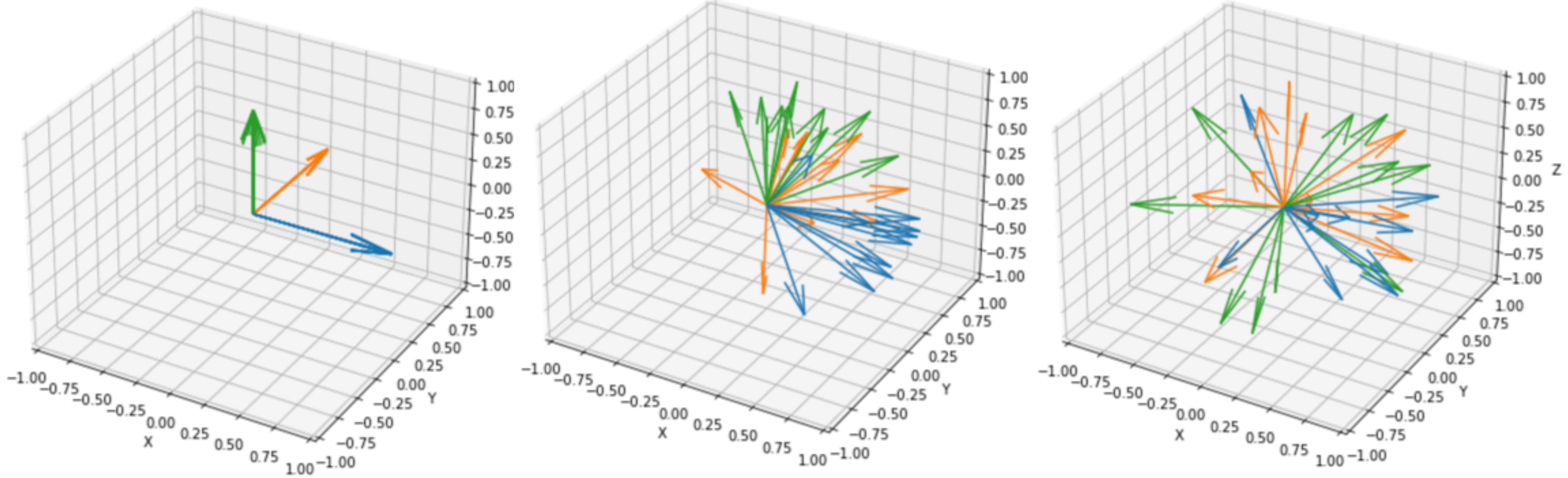
Source: Lilian Weng

Diffuse rotation $R \in \mathrm{SO}(3)$

Brownian motion on $\mathrm{SO}(3)$

$$p_{t|0}\left(R^{(t)} \,|\, R^{(0)}\right) = \mathrm{IGSO}_3(r^{(t)}; r^{(0)}, t)$$

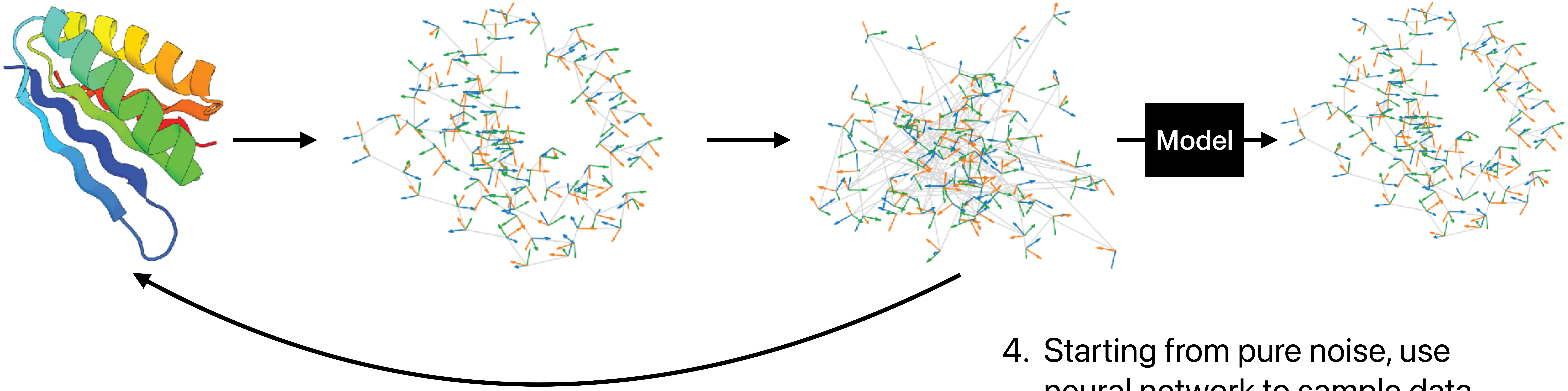
where $r^{(t)} = \mathrm{Log}(R^{(t)}),\ r^{(0)} = \mathrm{Log}(R^{(0)})$



$t = 0.0$ $t = 0.5$ $t = T$

Frame Diffusion: Training & Generation

1. Parameterize proteins
2. Corrupt
3. Train neural network to uncorrupt

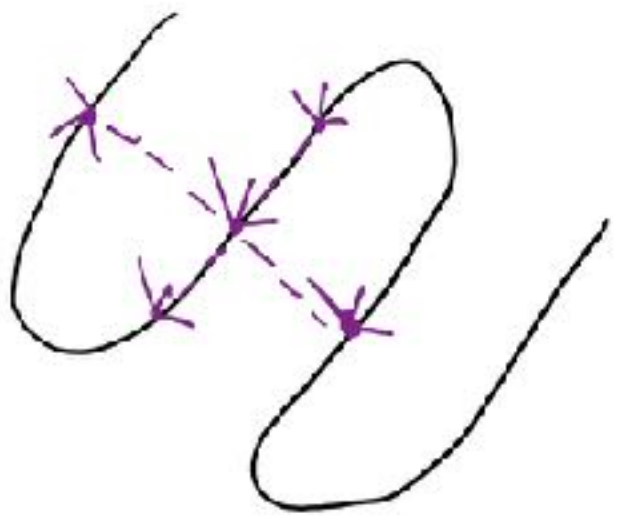


4. Starting from pure noise, use neural network to sample data.

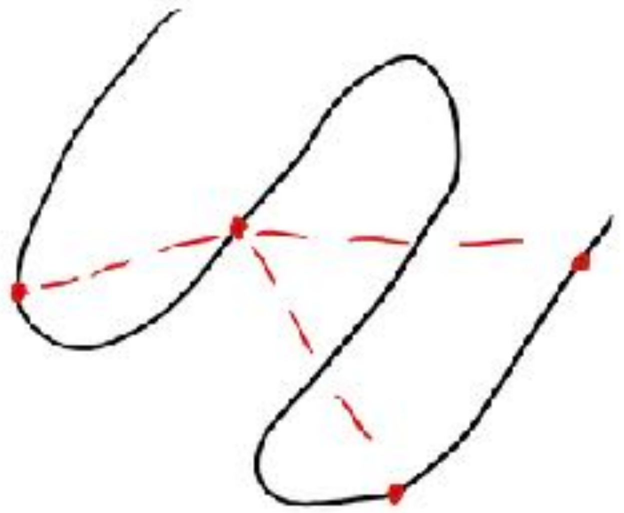
Model architecture

- Heavily inspired by AlphaFold2 architecture with two main components:

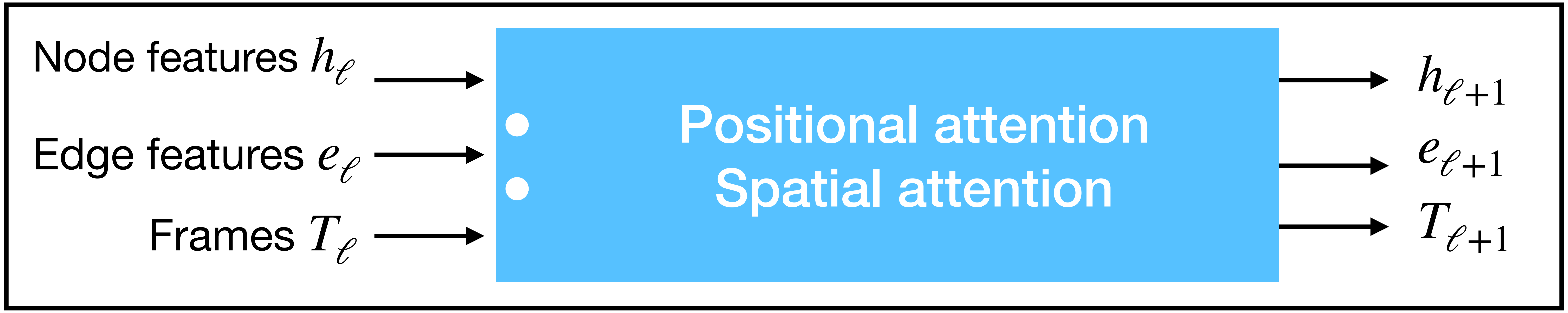
Spatial attention biases towards local residues



Positional attention allows global interactions.



Neural network



Node features h_ℓ

Edge features e_ℓ

Frames T_ℓ

- Positional attention
- Spatial attention

$h_{\ell+1}$

$e_{\ell+1}$

$T_{\ell+1}$

Single layer ℓ. Full model: stack multiple layers end-to-end.

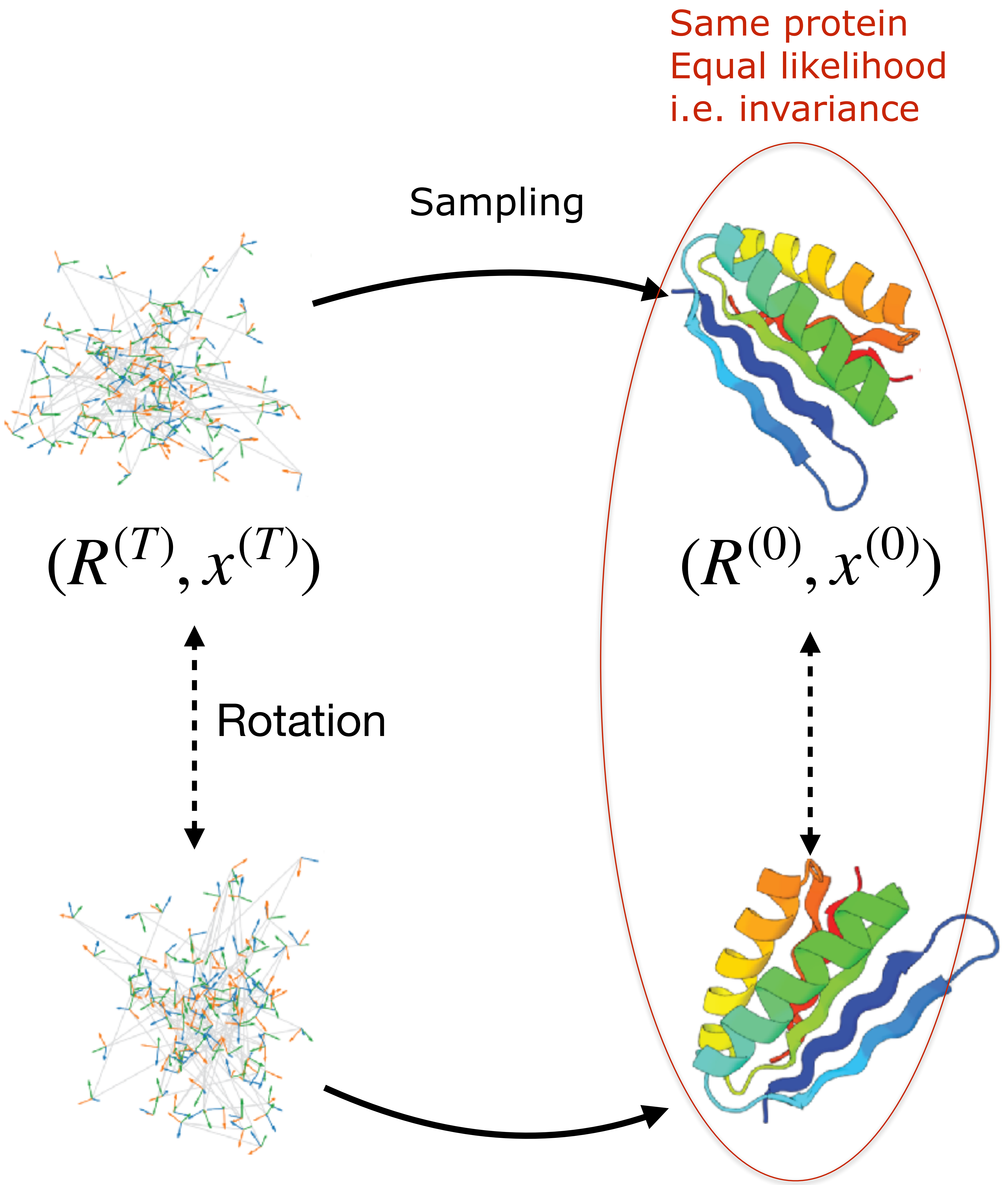
SE$(3)^N$ invariance

- Invariance requires the following:
  - By learning a SE$(3)^N$ equivariant score model.
  - Translation invariance: by zero-centering.

Needs to be equivariant

$$dx = [f(x,t) - g(t)^2 \boxed{\nabla \log p_t(x)}]dt + g(t)dB$$

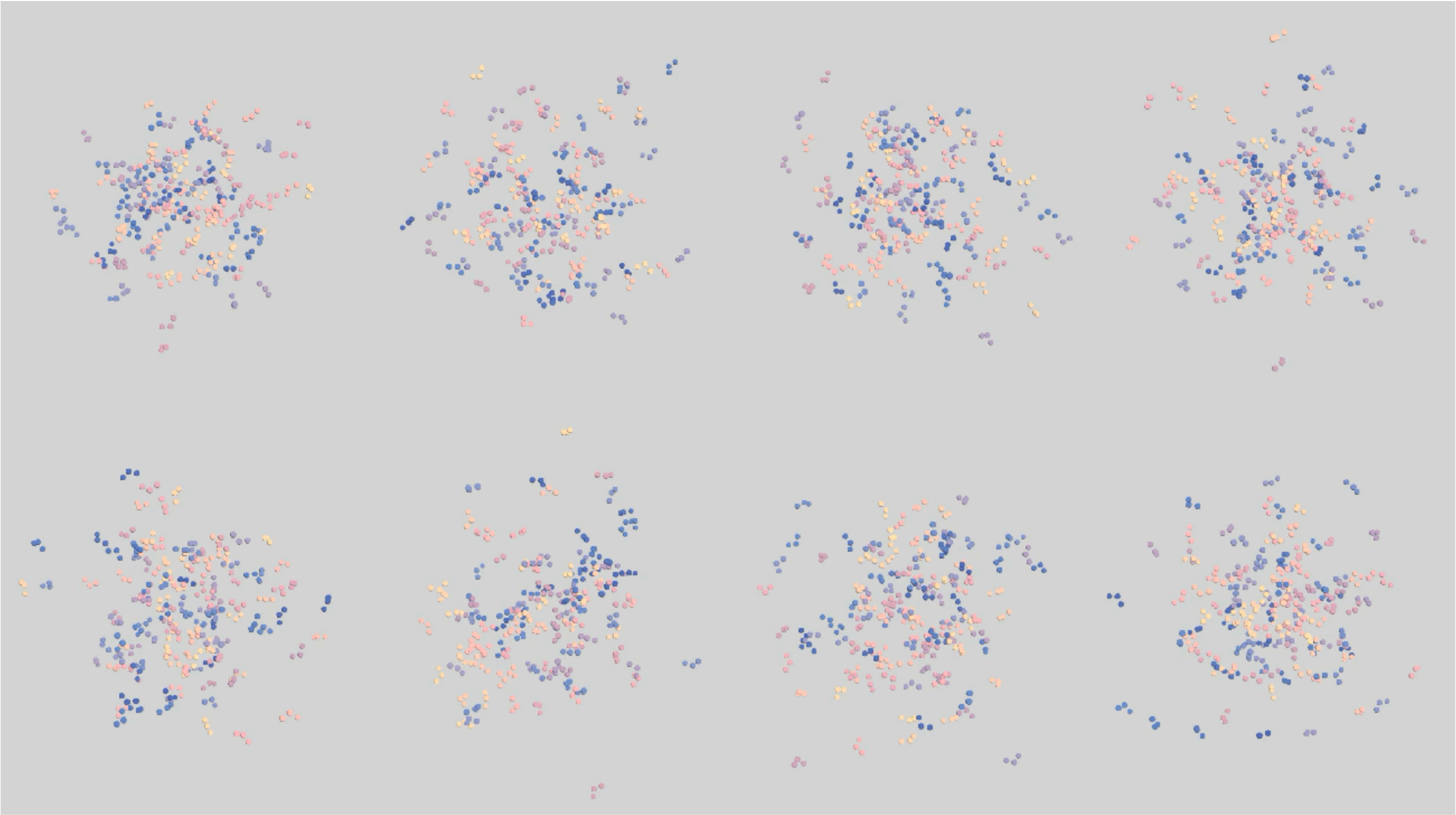
Reverse process (sampling)



Unconditional generation

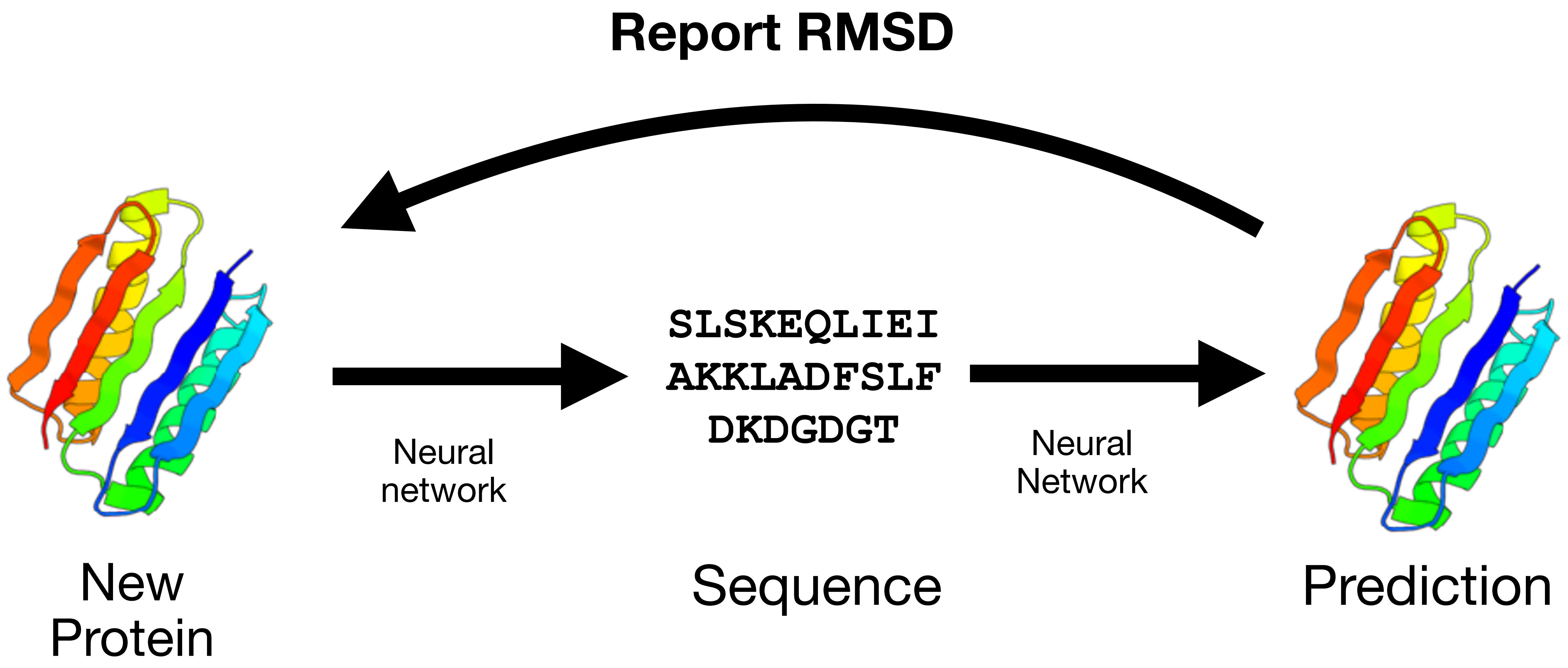
How well does the model sample realistic proteins?

- Generation from **only noise** with no other conditions.



In-silico Evaluation Metrics

- **Realism check**: could a *sequence* exist with the AI-generated *structure*.



Report RMSD

Neural network

SLSKEQLIEI
AKKLADFSLF
DKDGDGT

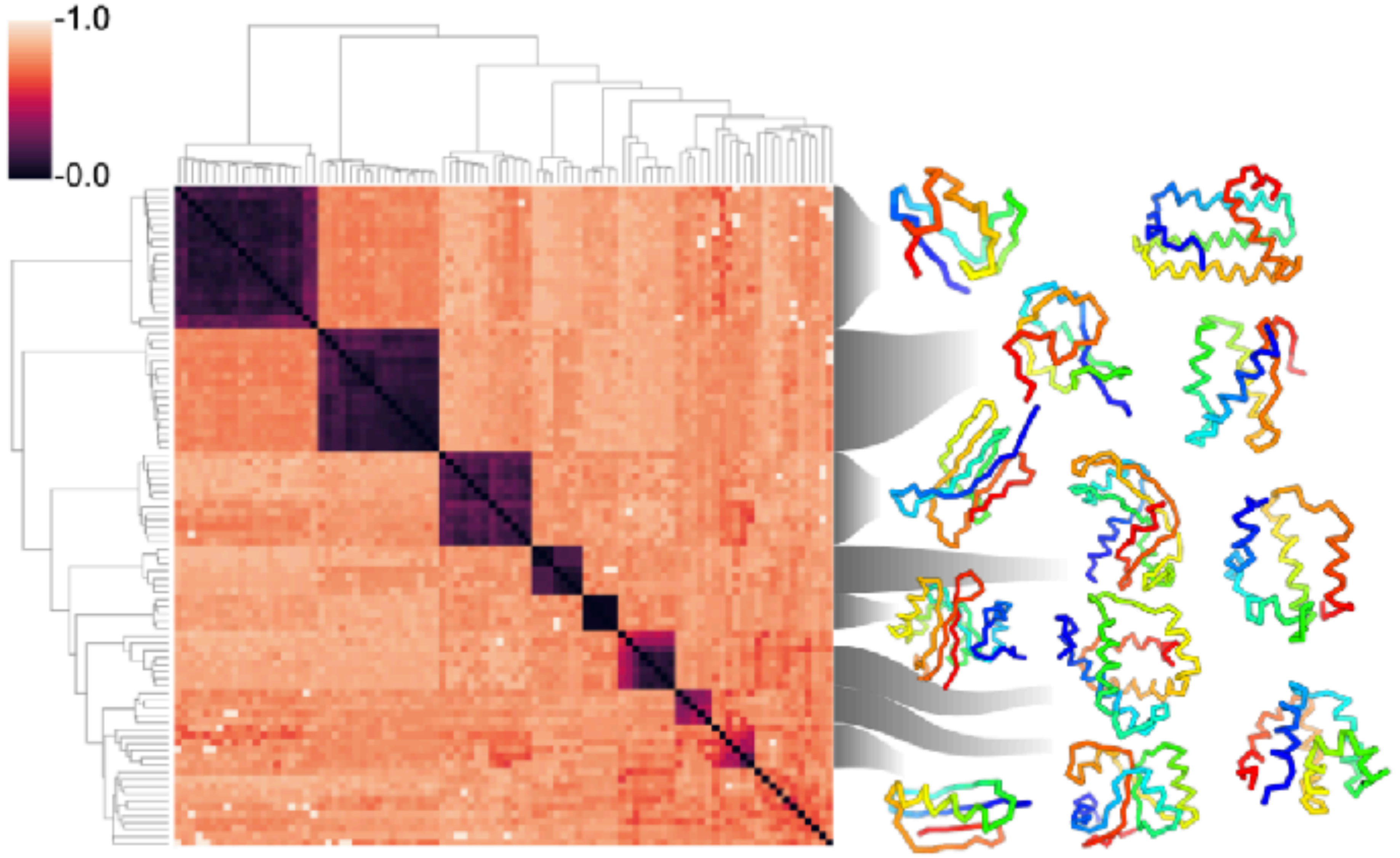
Neural Network

New Protein

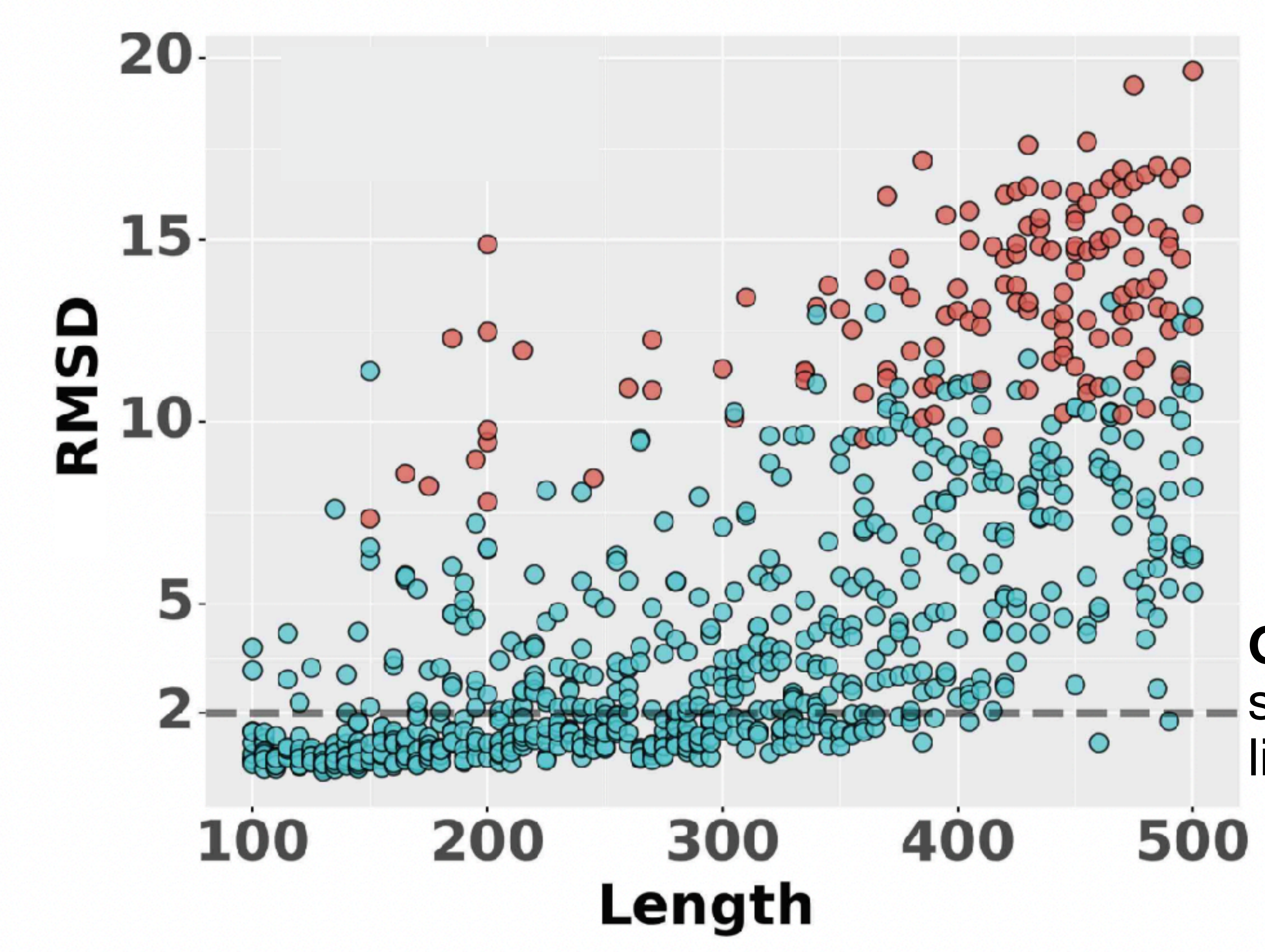
Sequence

Prediction

- **Diversity**: structurally cluster all designable backbones. Report number or fraction of clusters.



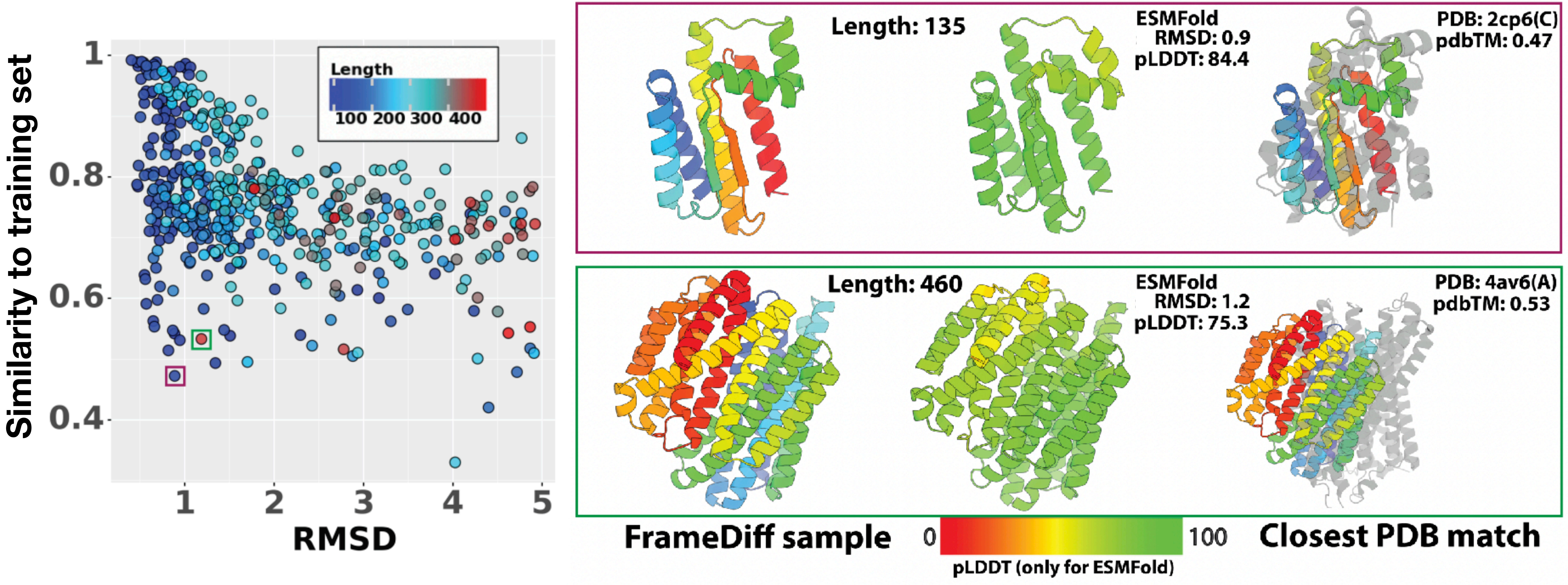
FrameDiff results



Goal: as many samples below this line.

FrameDiff results

- *In-silico* evidence of generalizing beyond PDB (training set)

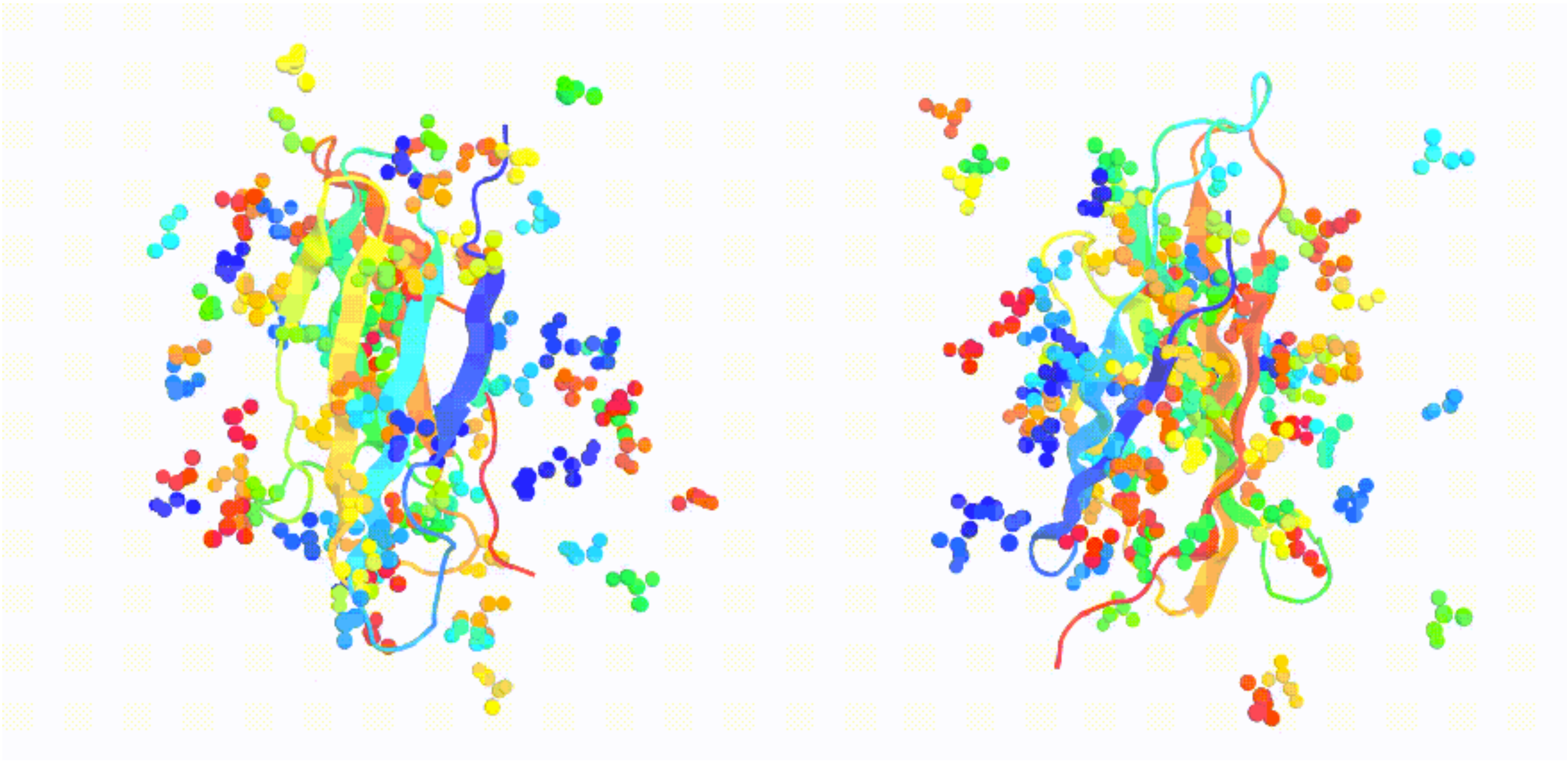


Summary: FrameDiff

Desiderata

1. Generate **high quality** structures. ✅
2. Generate **diverse** structures. ✅
3. Generate **novel** structures. ✅
4. Generate **functional** structures.

Shift to flow matching

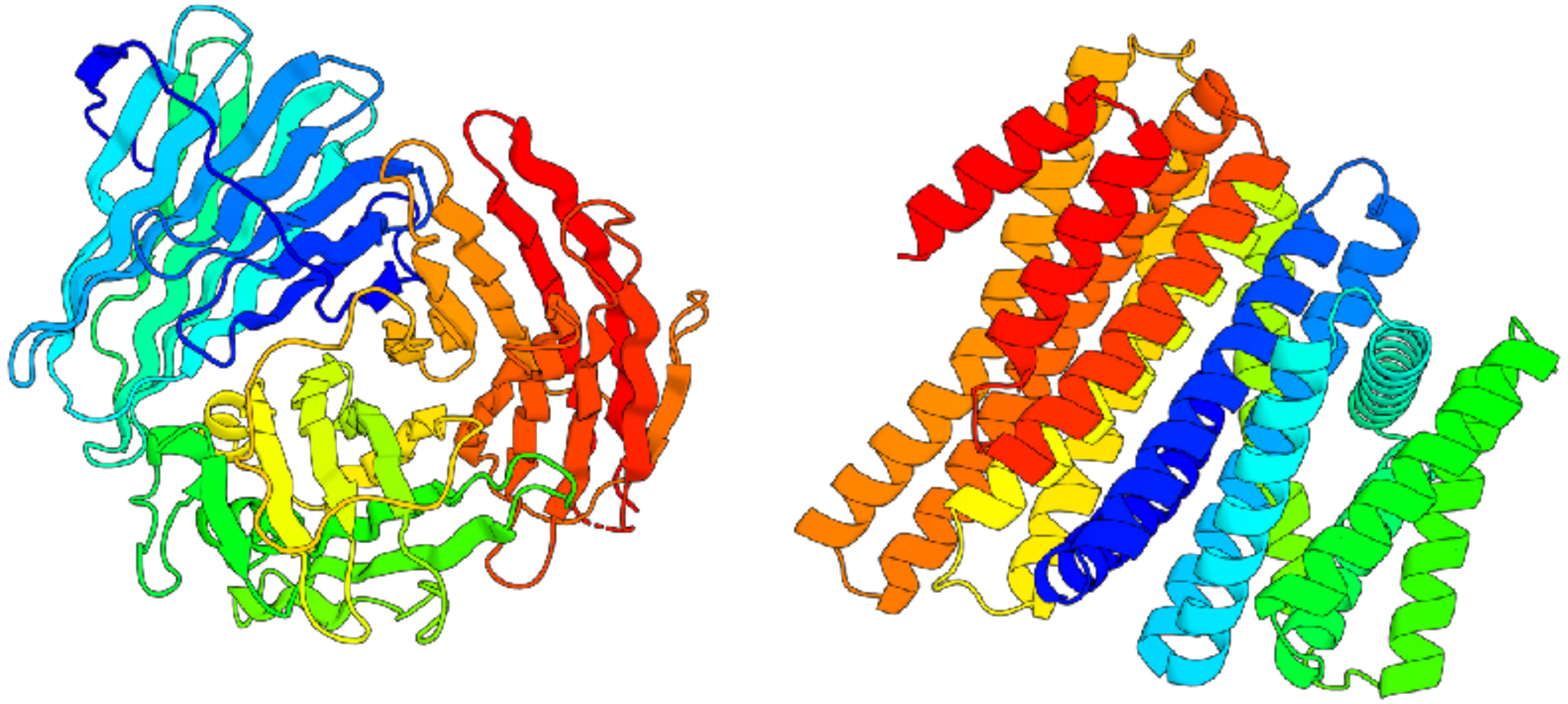


FrameDiff

FrameFlow

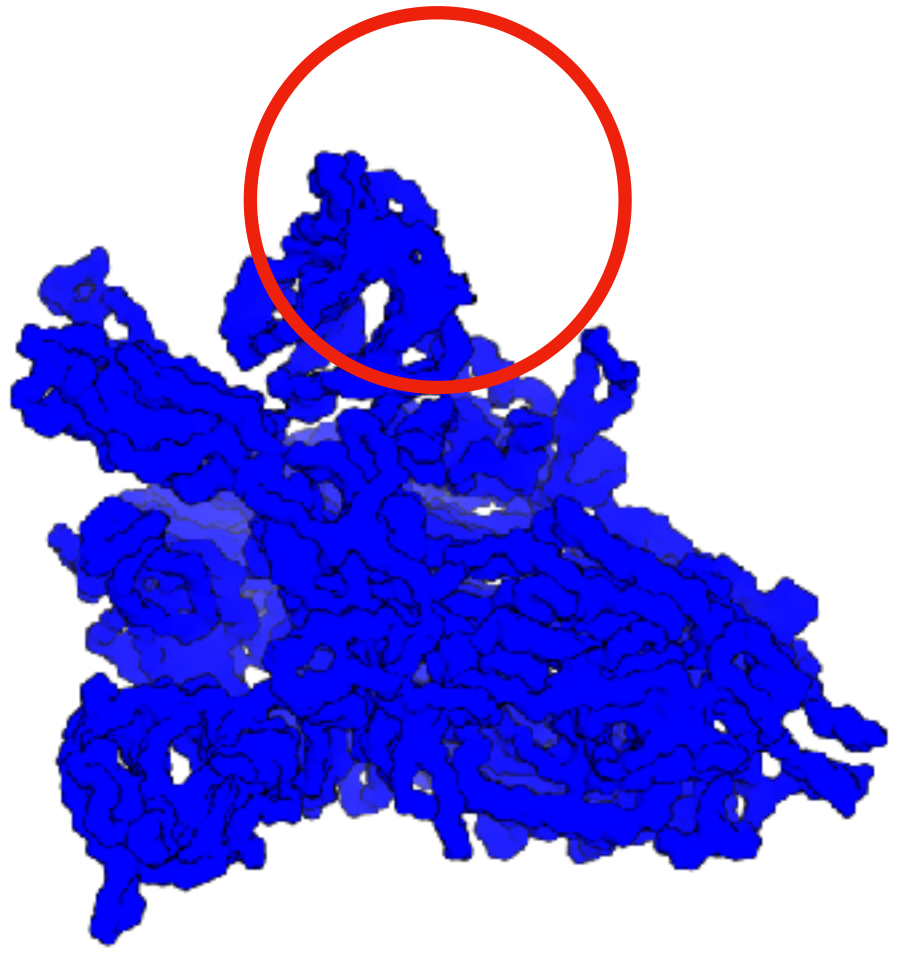
Protein generation paradigms

Unconditional generation



$P(x)$

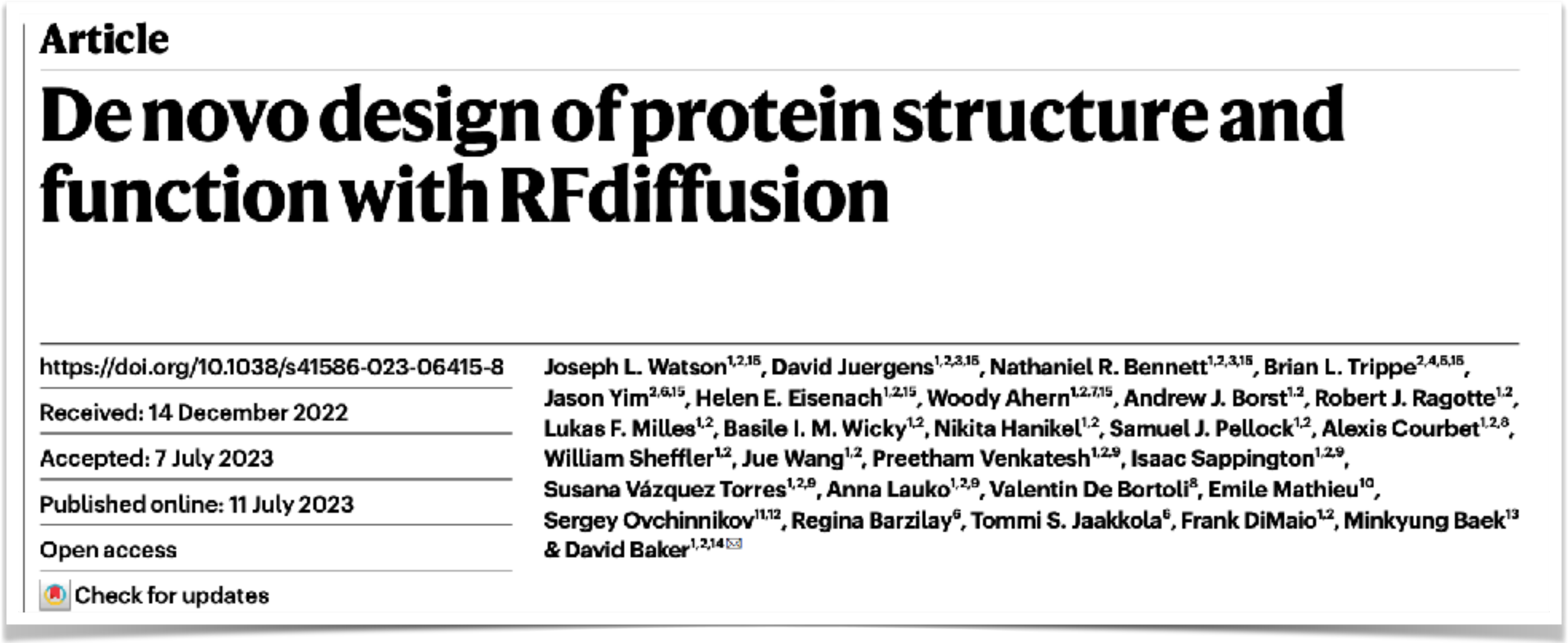
Next: Conditional generation



Condition y

$P(x \mid y)$

Diffusion model for protein design



Article

De novo design of protein structure and function with RFdiffusion

https://doi.org/10.1038/s41586-023-06415-8

Received: 14 December 2022

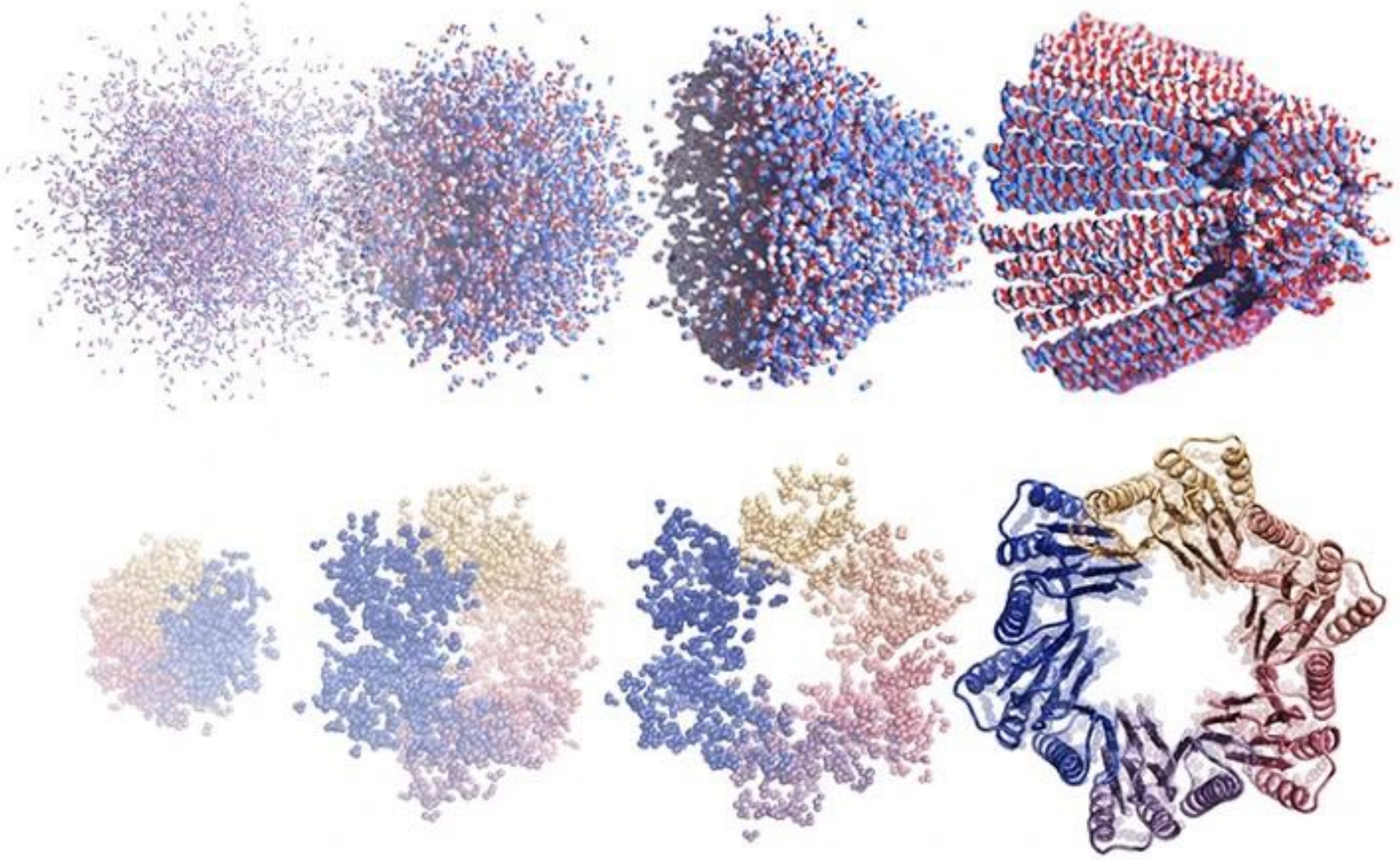
Accepted: 7 July 2023

Published online: 11 July 2023

Open access

Check for updates

Joseph L. Watson[1,2,15], David Juergens[1,2,3,15], Nathaniel R. Bennett[1,2,3,15], Brian L. Trippe[2,4,5,15], Jason Yim[2,6,15], Helen E. Eisenach[1,2,15], Woody Ahern[1,2,7,15], Andrew J. Borst[1,2], Robert J. Ragotte[1,2], Lukas F. Milles[1,2], Basile I. M. Wicky[1,2], Nikita Hanikel[1,2], Samuel J. Pellock[1,2], Alexis Courbet[1,2,8], William Sheffler[1,2], Jue Wang[1,2], Preetham Venkatesh[1,2,9], Isaac Sappington[1,2,9], Susana Vázquez Torres[1,2,9], Anna Lauko[1,2,9], Valentin De Bortoli[8], Emile Mathieu[10], Sergey Ovchinnikov[11,12], Regina Barzilay[6], Tommi S. Jaakkola[6], Frank DiMaio[1,2], Minkyung Baek[13] & David Baker[1,2,14] ✉

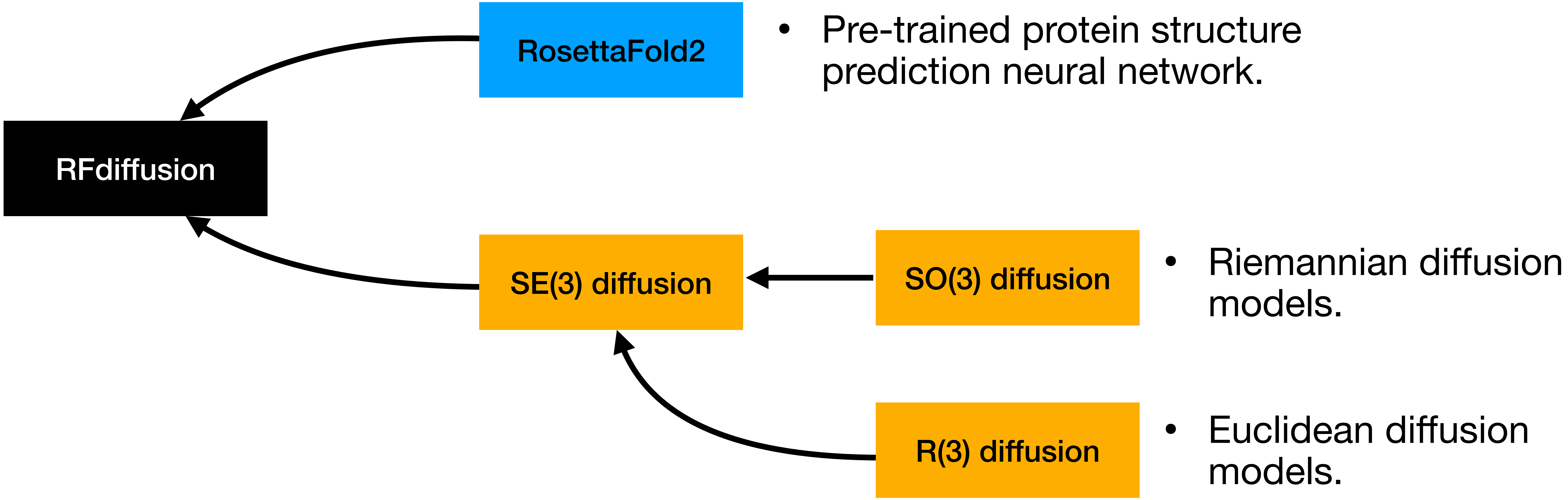


INSTITUTE FOR
Protein Design
UNIVERSITY of WASHINGTON

MIT CSAIL

RosettaFold diffusion

- RosettaFold diffusion is a culmination of multiple ideas.

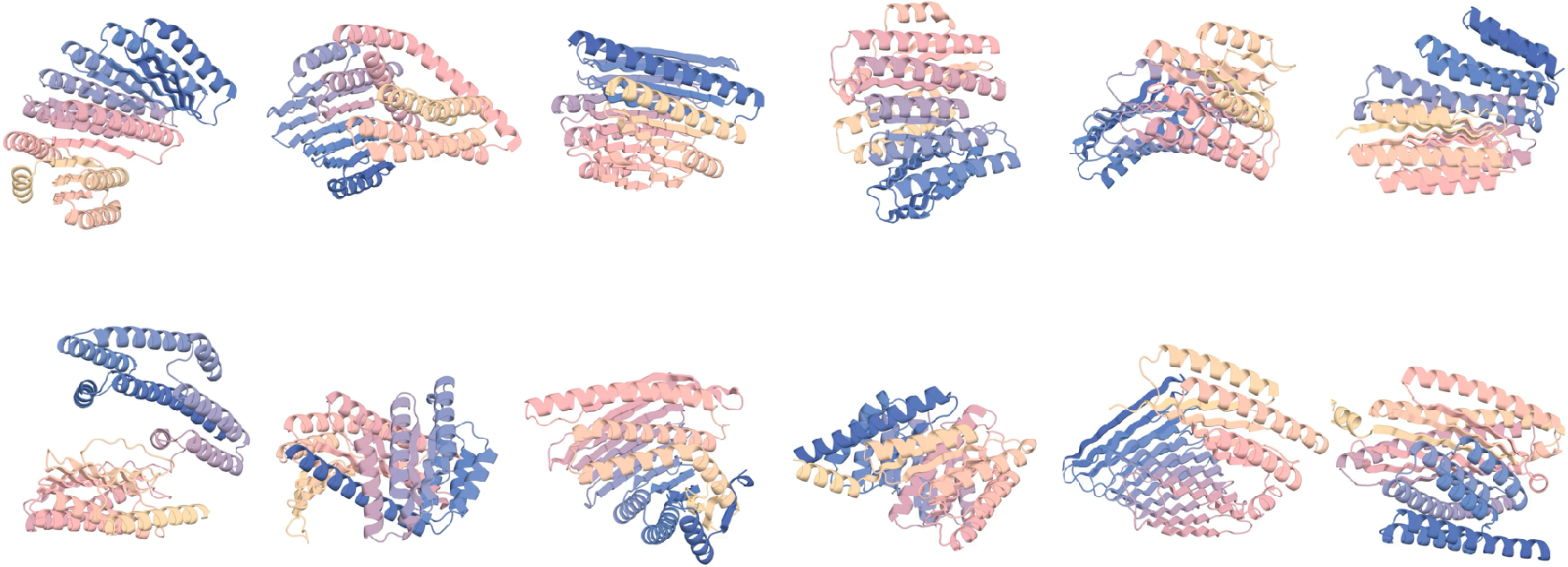


- Pre-trained protein structure prediction neural network.

- Riemannian diffusion models.

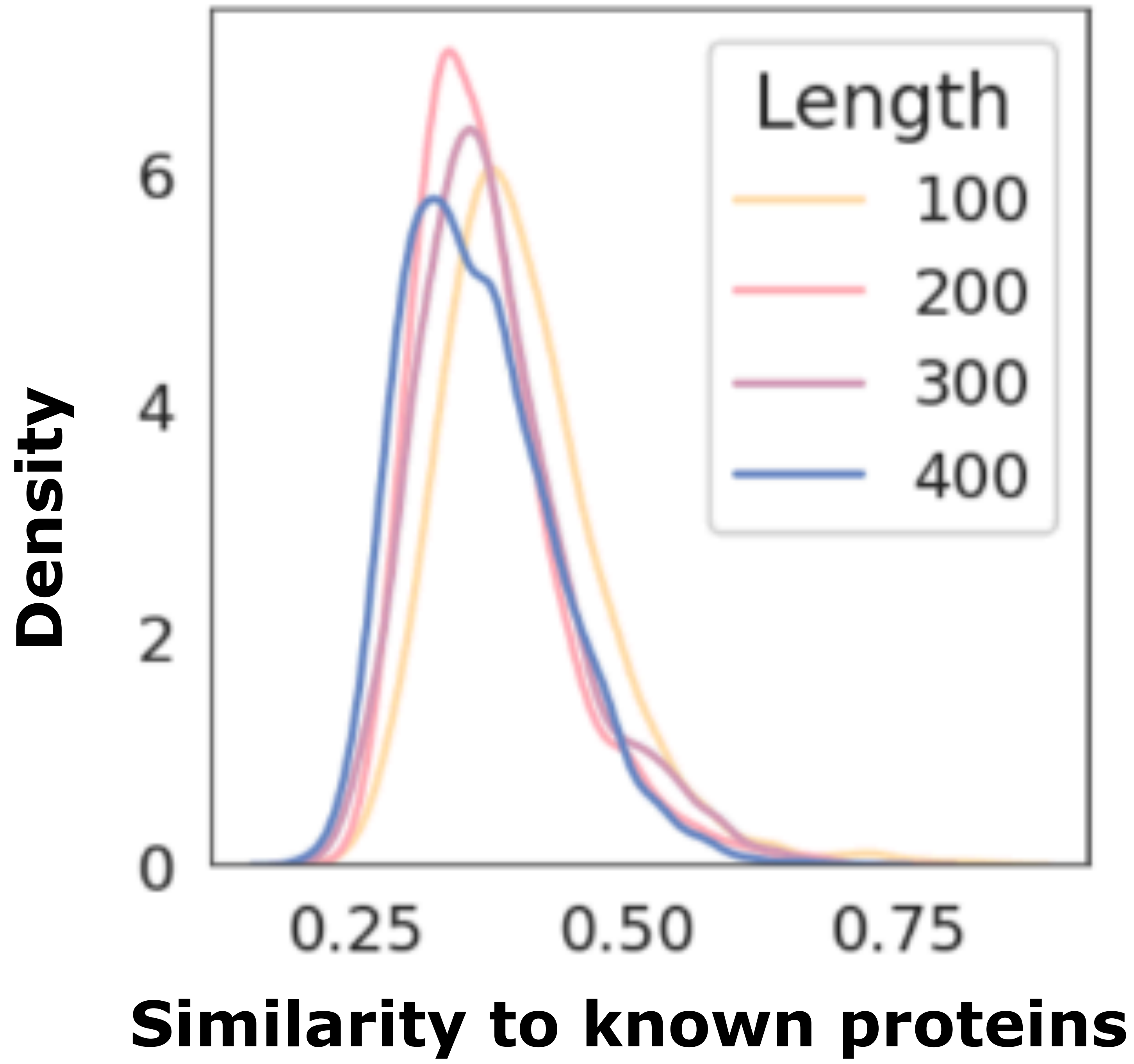
- Euclidean diffusion models.

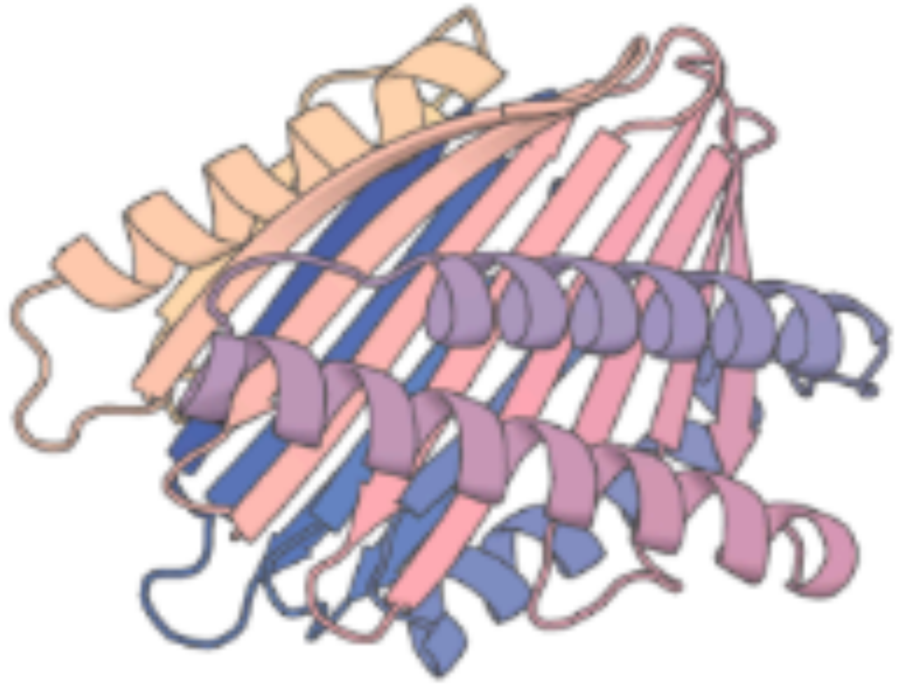
Pre-training improves unconditional generation



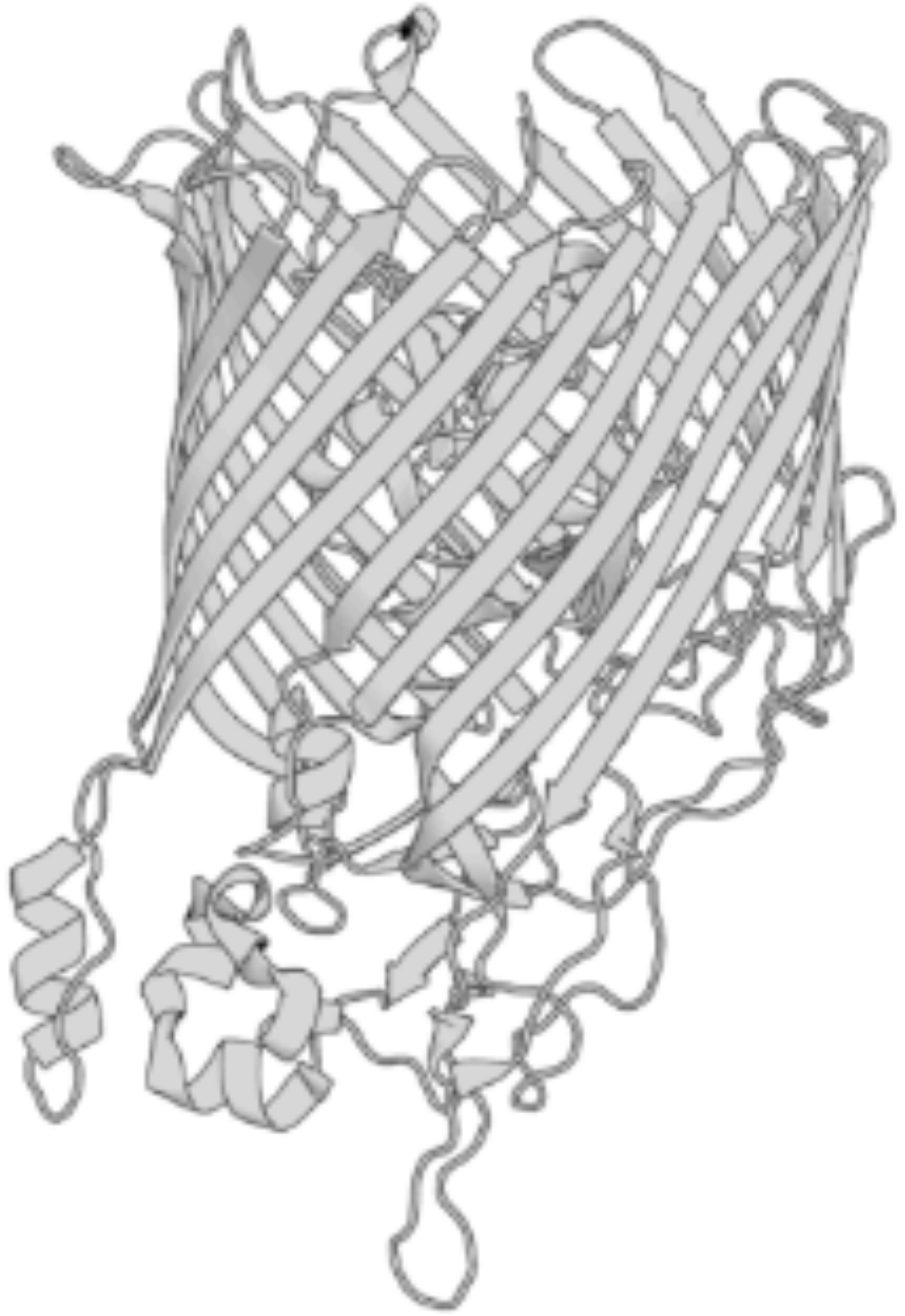
Quantifying novelty

Similarity to closest example in PDB

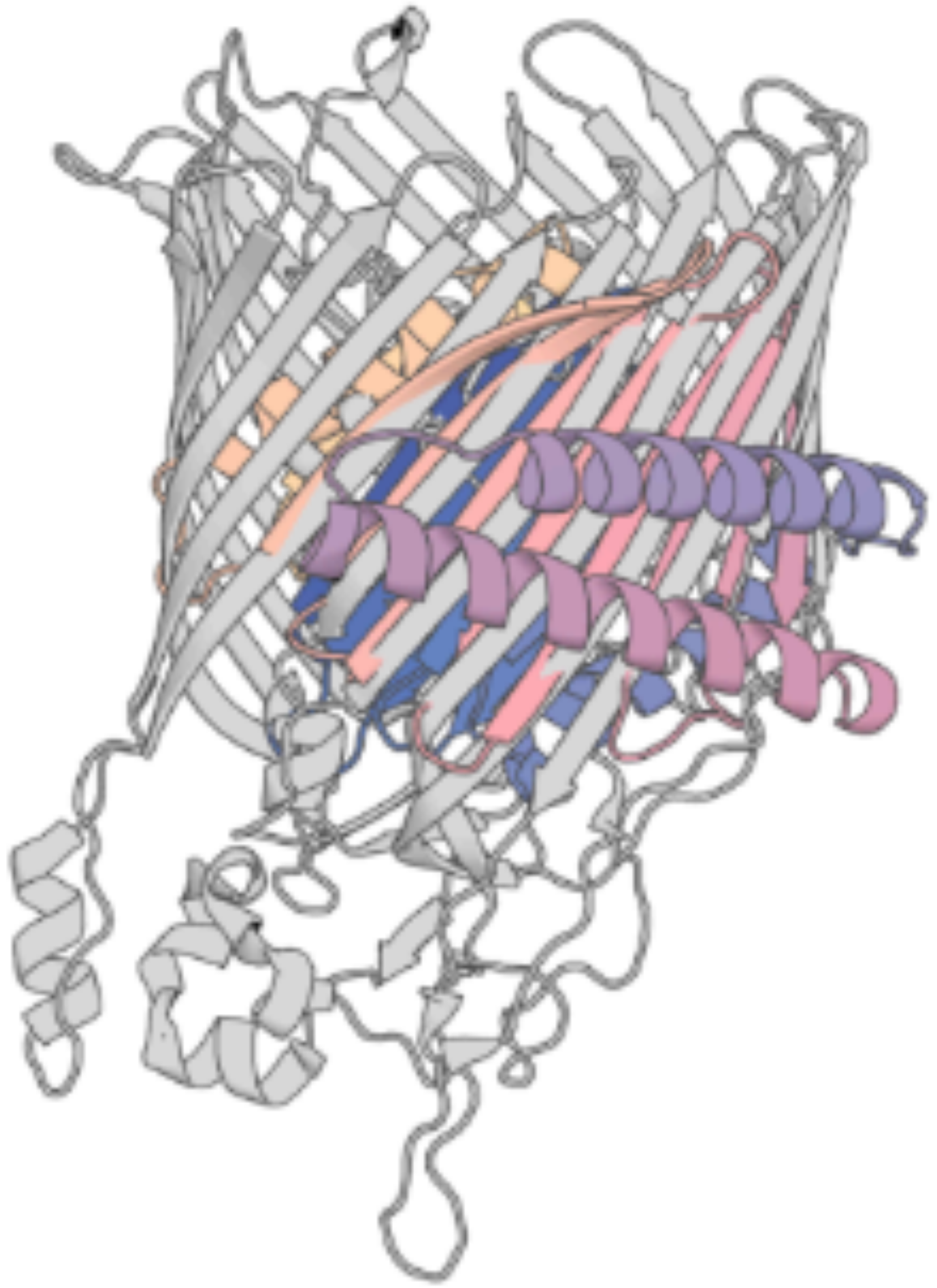




AI generated protein

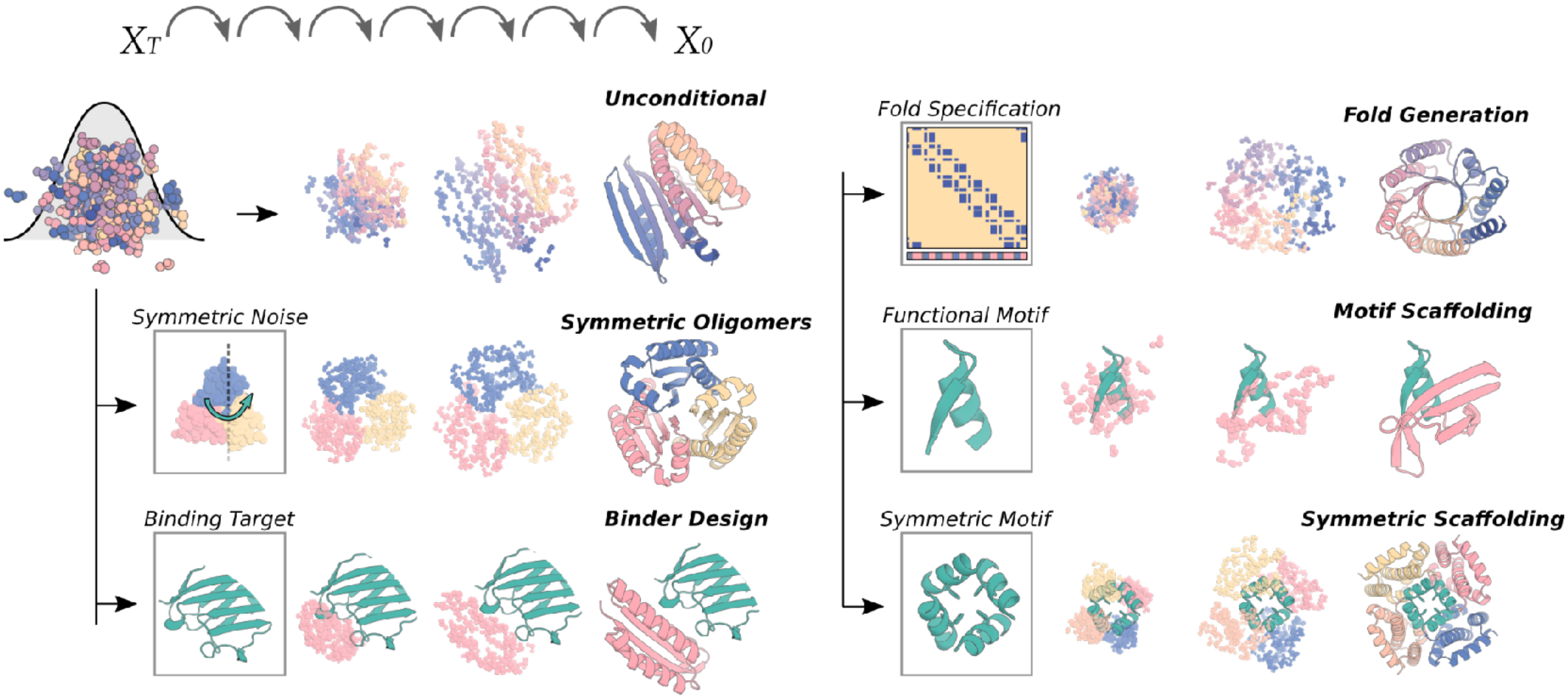


Most similar known protein



Superimposition

Conditional generation



Conditional diffusion guidance

How to guide structures towards specific functions and higher quality?

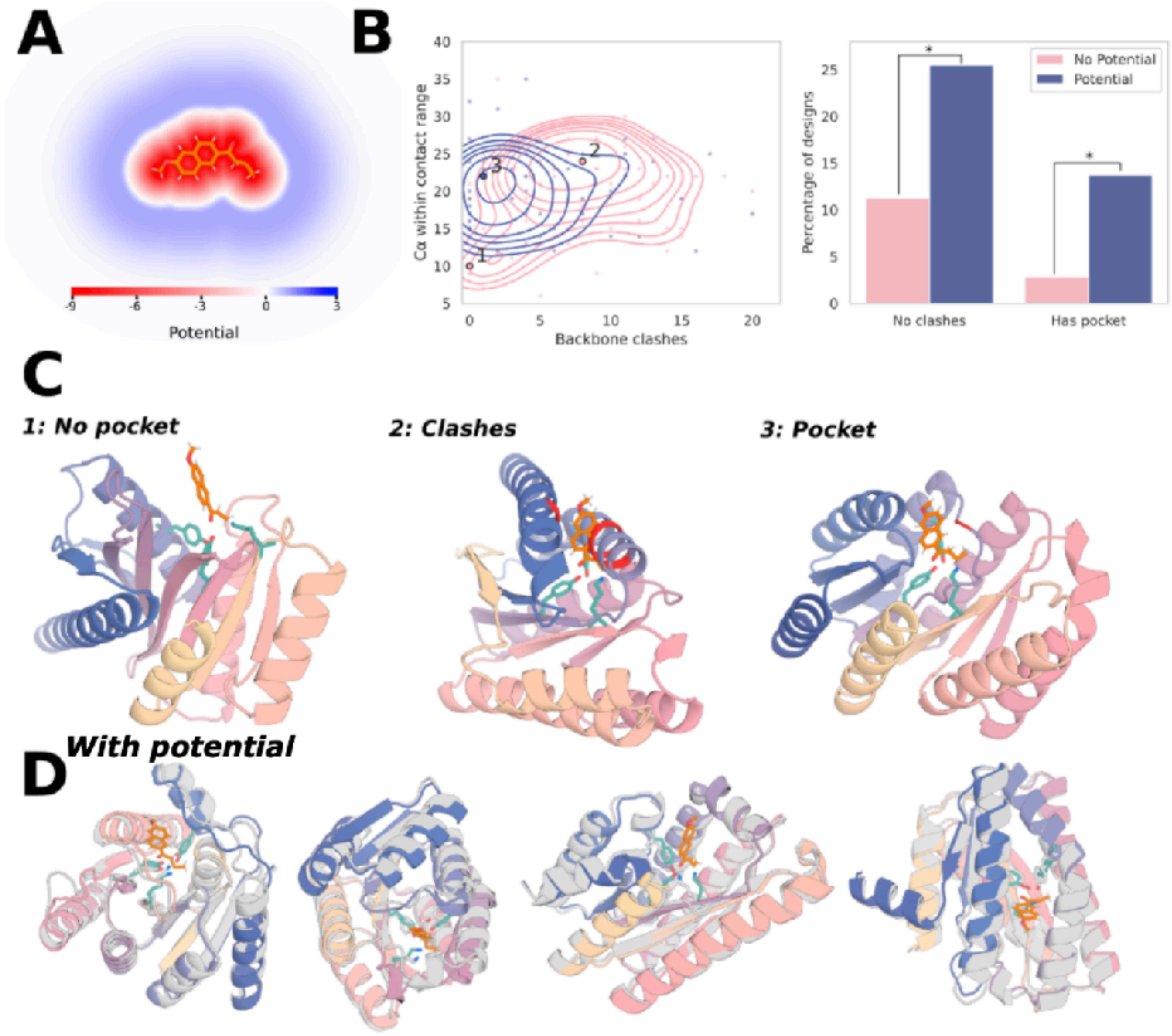
Solution: Inspired by classifier guidance, guide with potentials.

Classifier guidance:

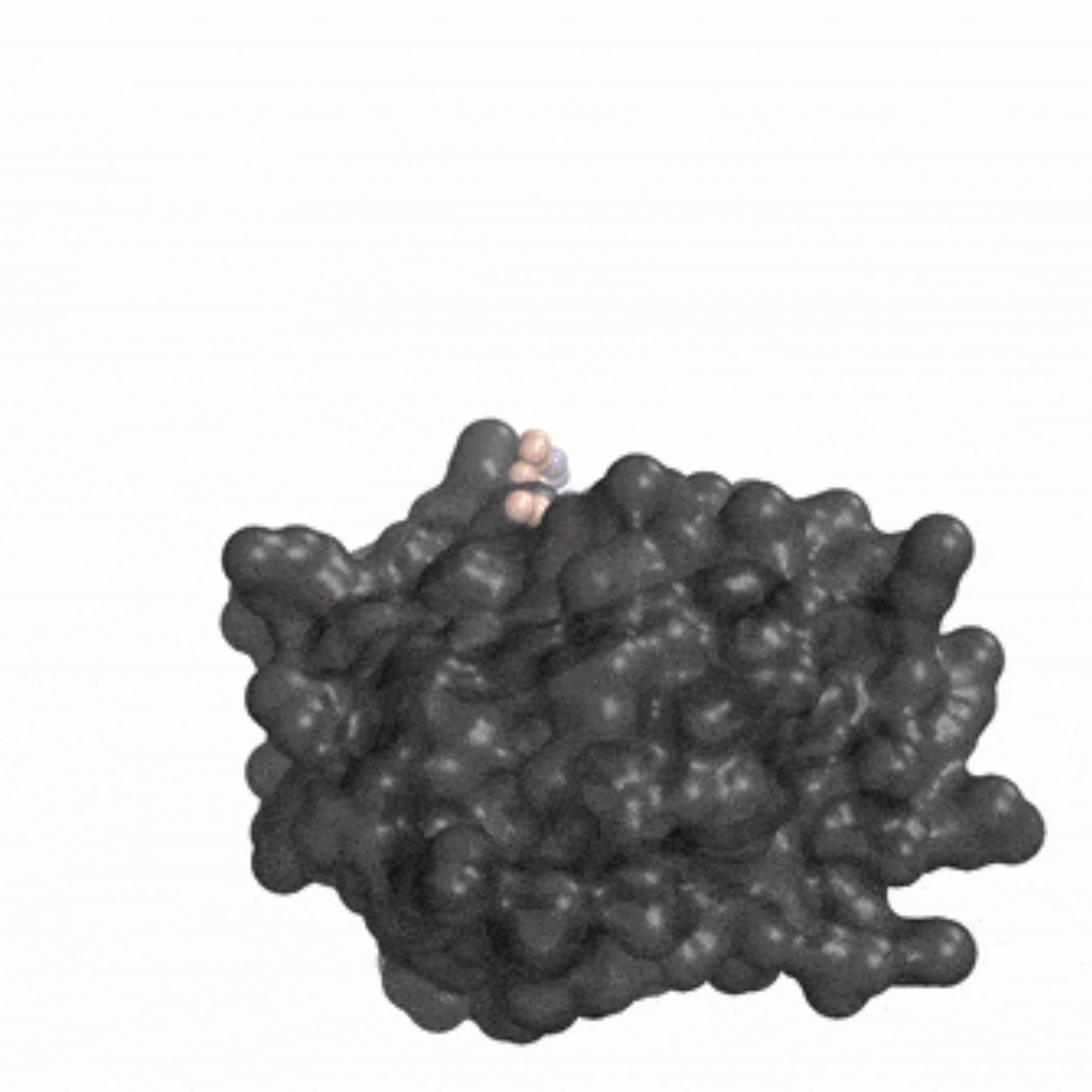
$$\nabla_{x^{(t)}} \log p(x^{(t)}) + \omega \nabla_{x^{(t)}} \log p(y = 1 \,|\, x^{(t)})$$

Potential guidance:

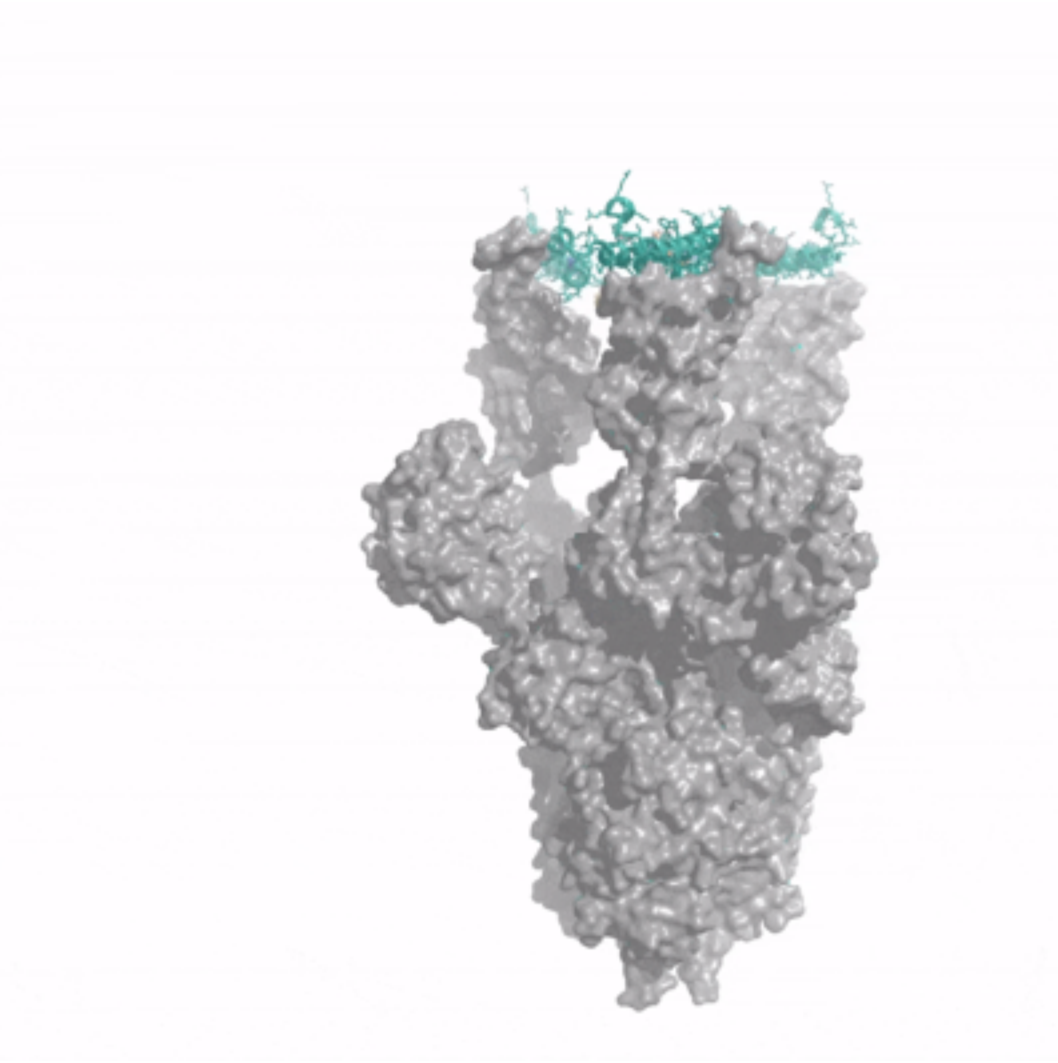
$$\nabla_{x^{(t)}} \log p(x^{(t)}) + \omega \nabla_{x^{(t)}} P(x^{(t)})$$



Binder generation



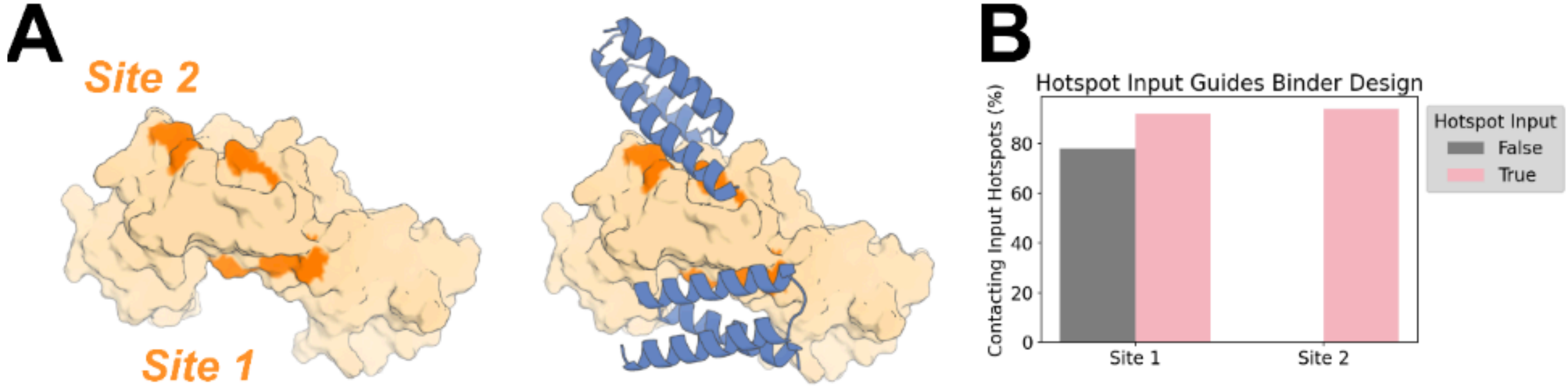
Binder generation



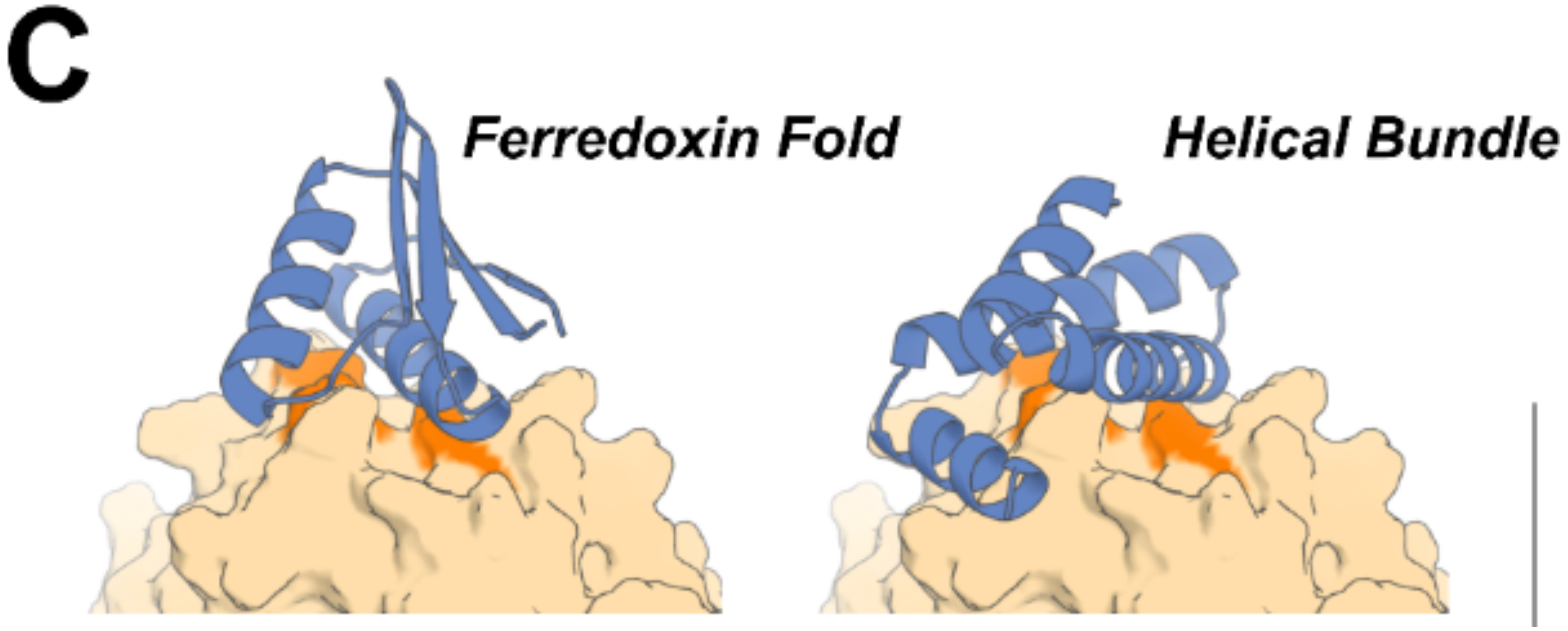
Symmetric complex
binder and scaffolding

Binder design

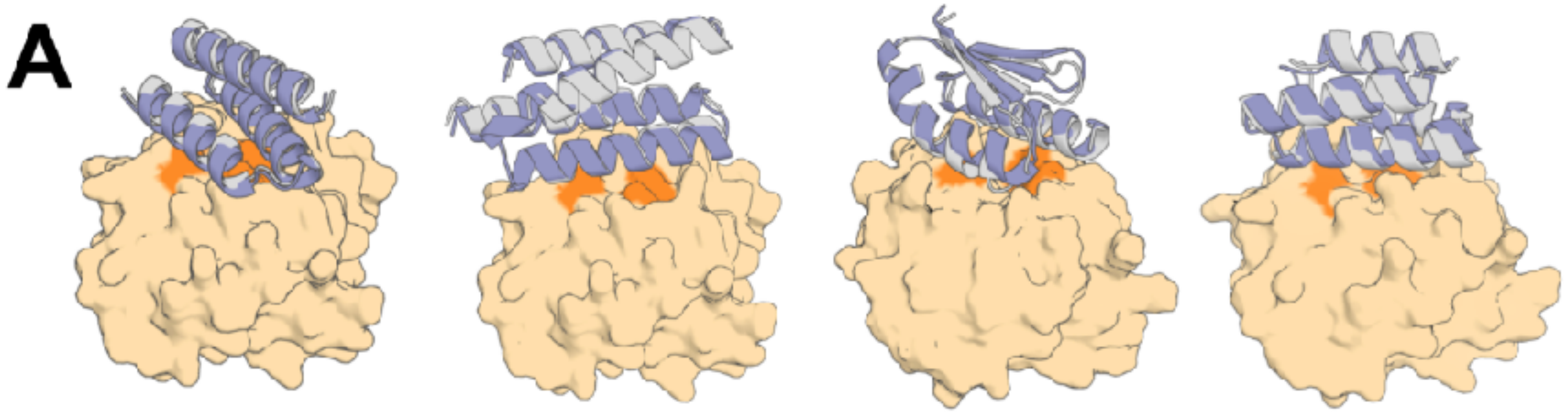
Guide binder generation towards hot spot residues.



Additionally condition the fold topology.

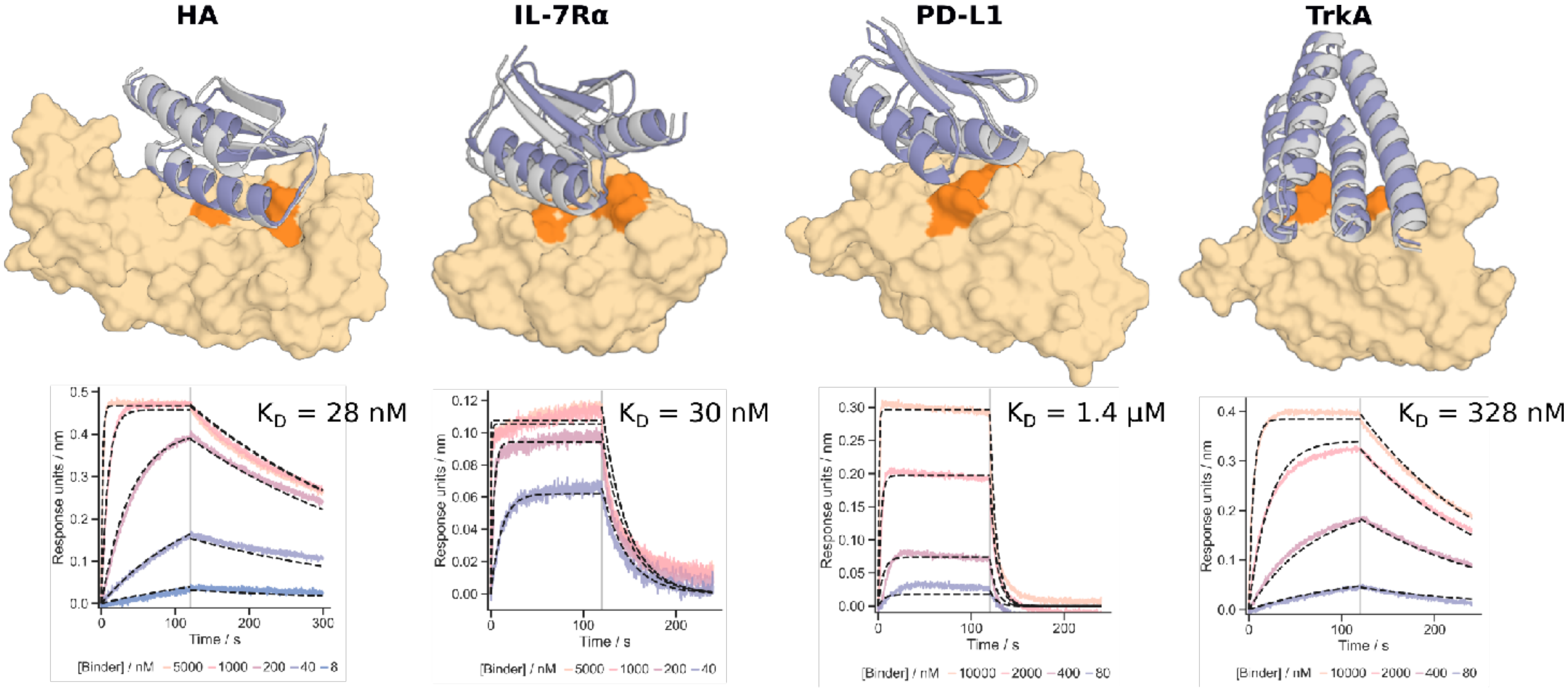


Or allow unconstrained folds.

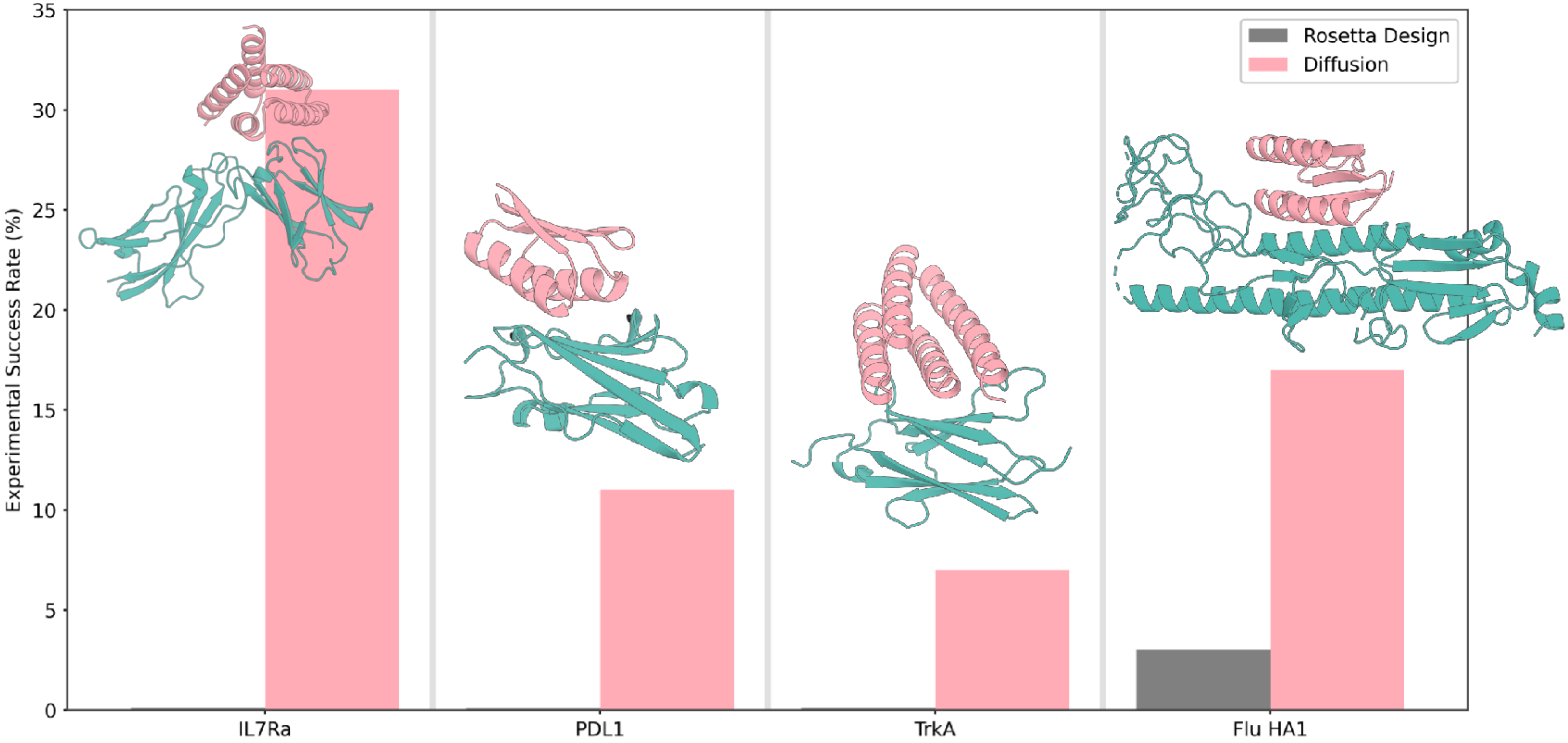


Wet-lab validation

De novo binder design

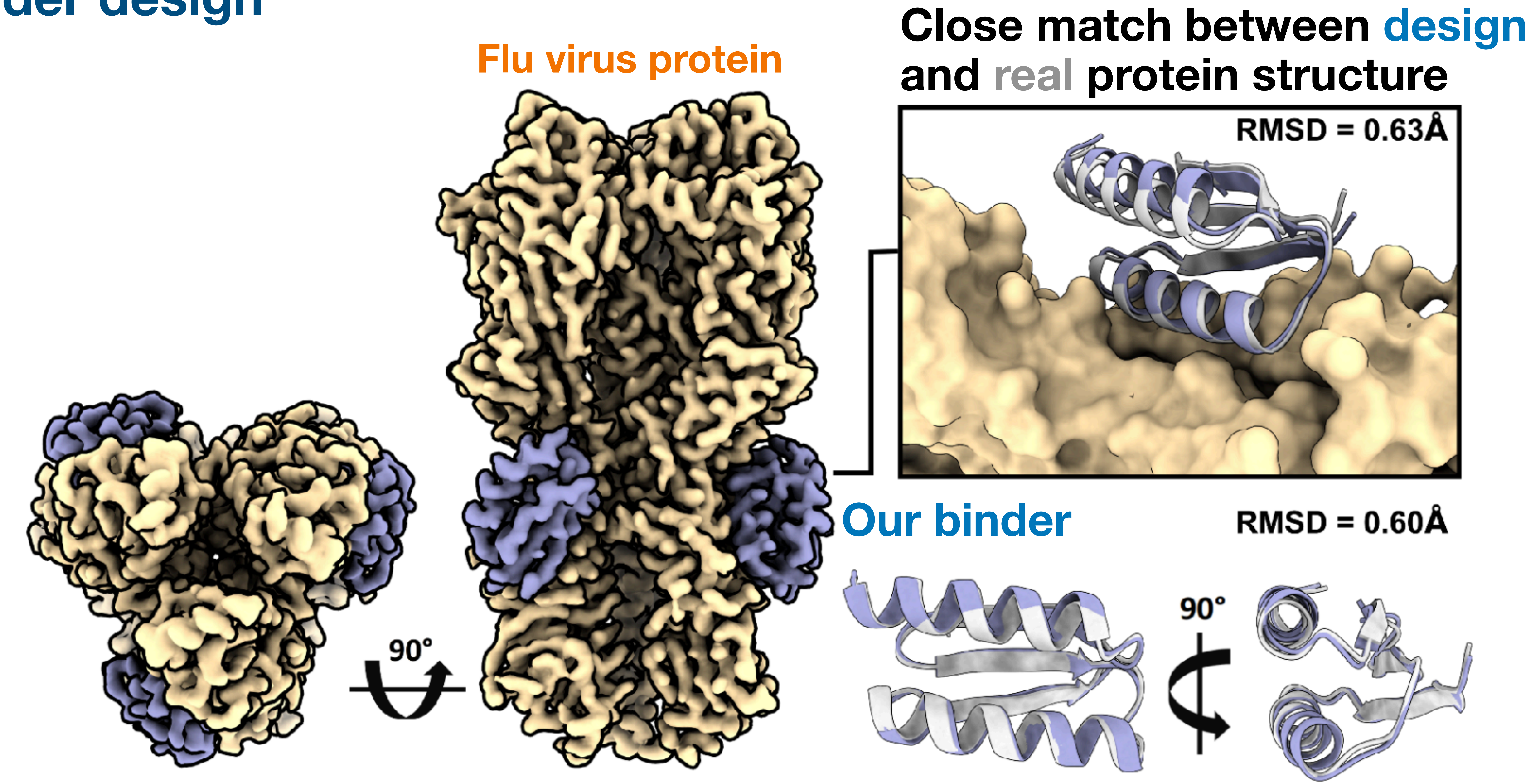


Orders of magnitude higher success than previous



Structural characterization

Binder design

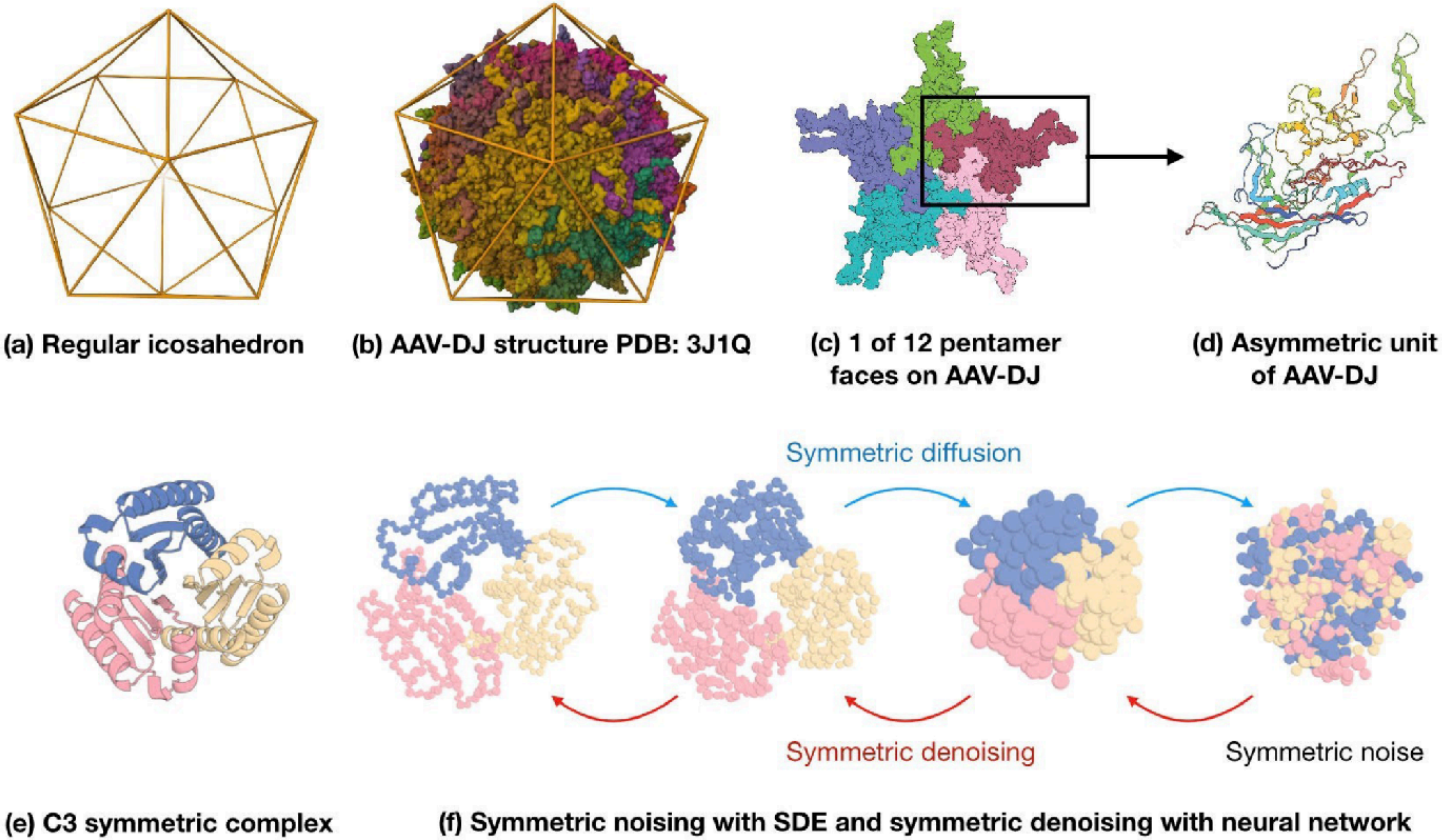


Flu virus protein

Close match between design and real protein structure

Our binder

Symmetric protein design



(a) Regular icosahedron

(b) AAV-DJ structure PDB: 3J1Q

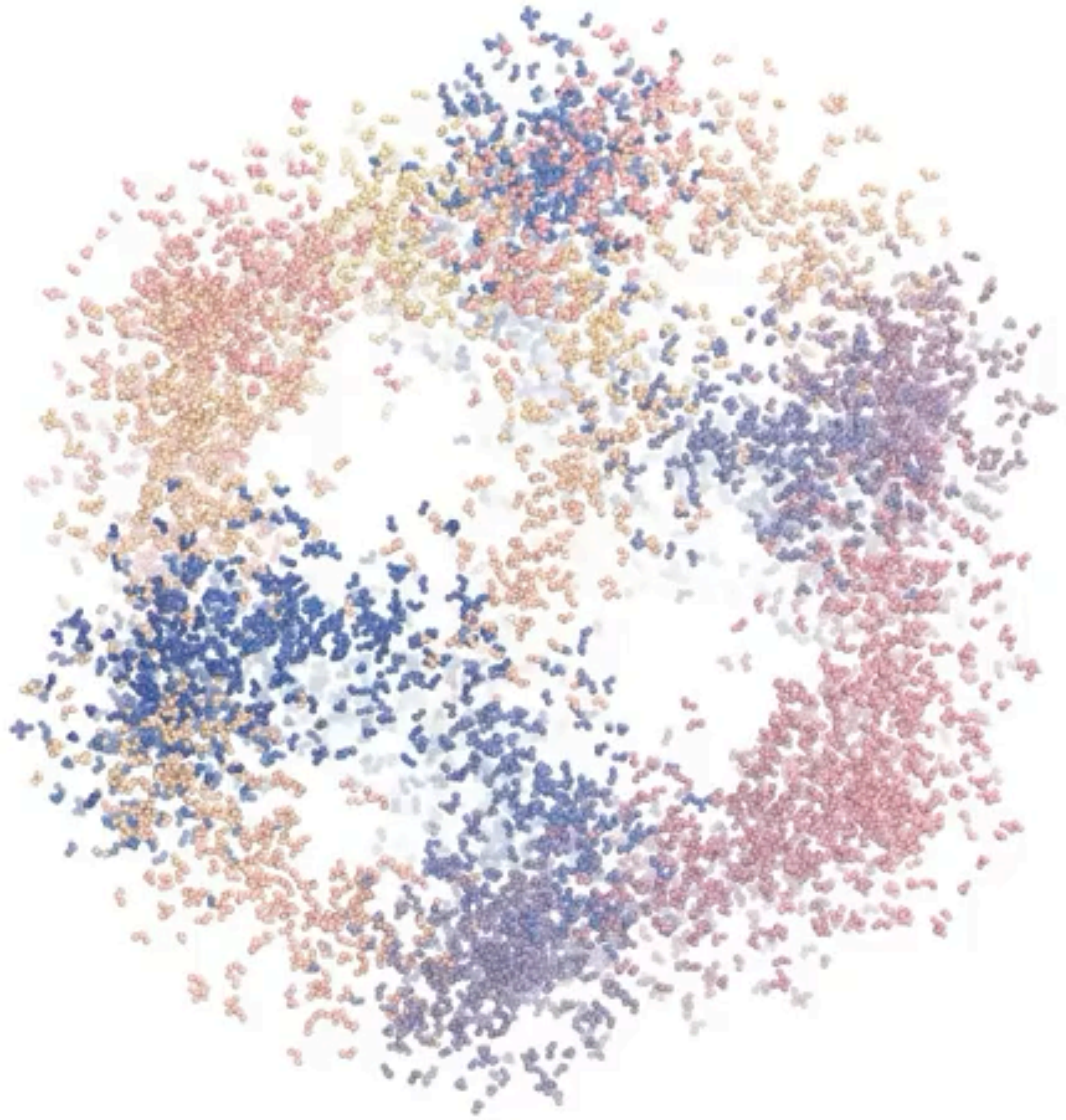
(c) 1 of 12 pentamer faces on AAV-DJ

(d) Asymmetric unit of AAV-DJ

(e) C3 symmetric complex

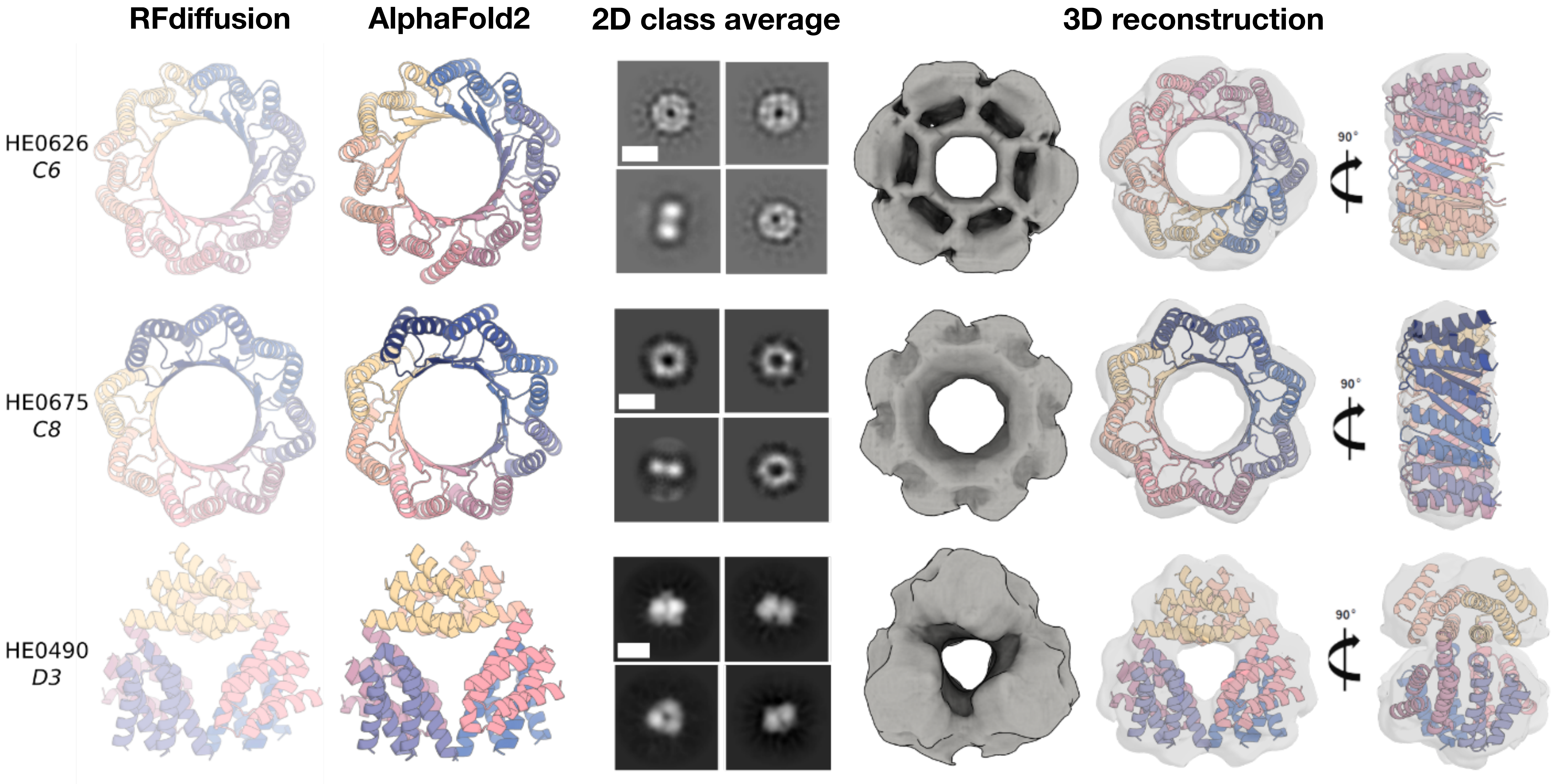
(f) Symmetric noising with SDE and symmetric denoising with neural network

Symmetric protein design



Wet-lab validation

Symmetric complex design



RFdiffusion follow-ups

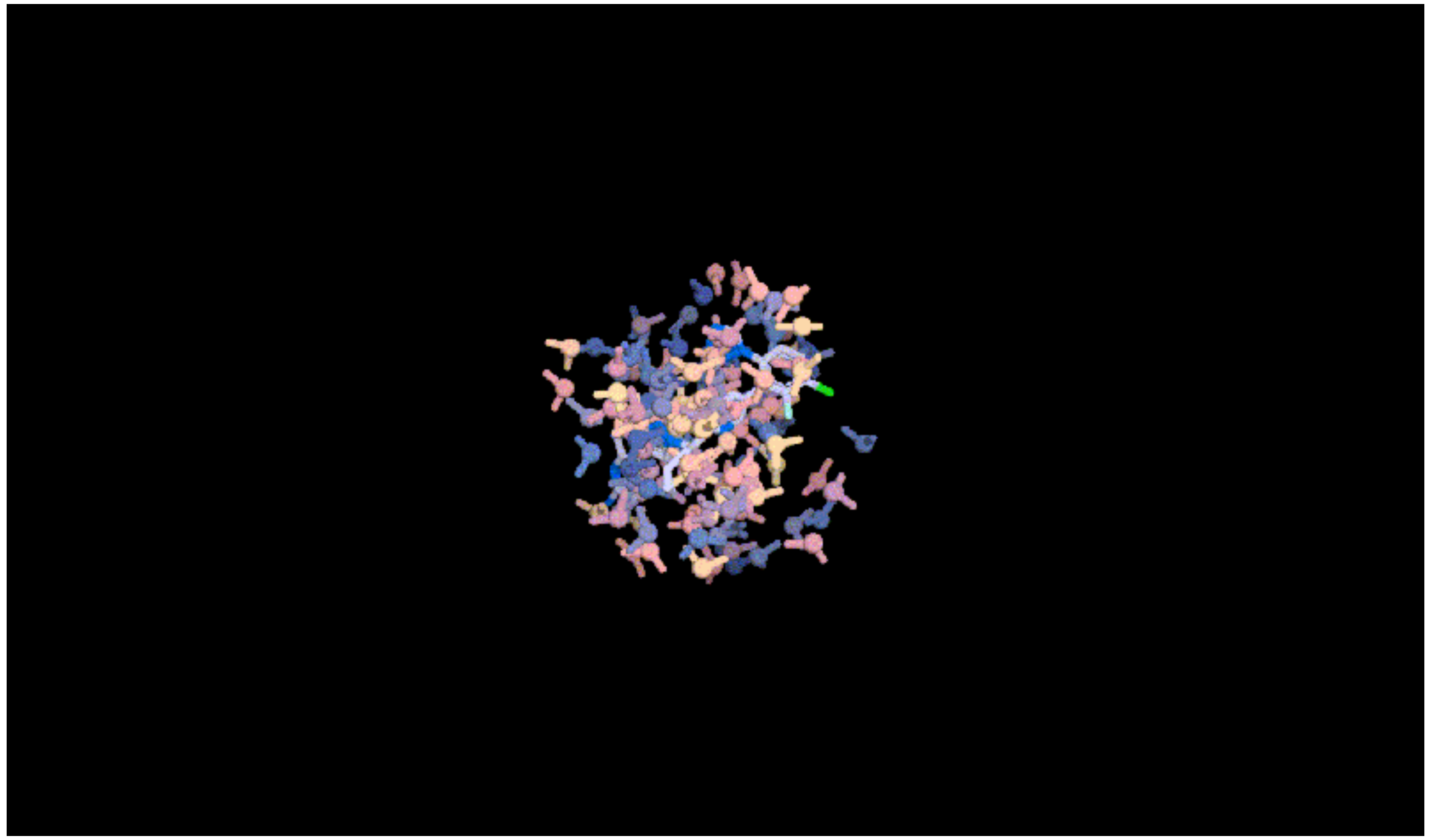
Generalized biomolecular modeling and design with RoseTTAFold All-Atom

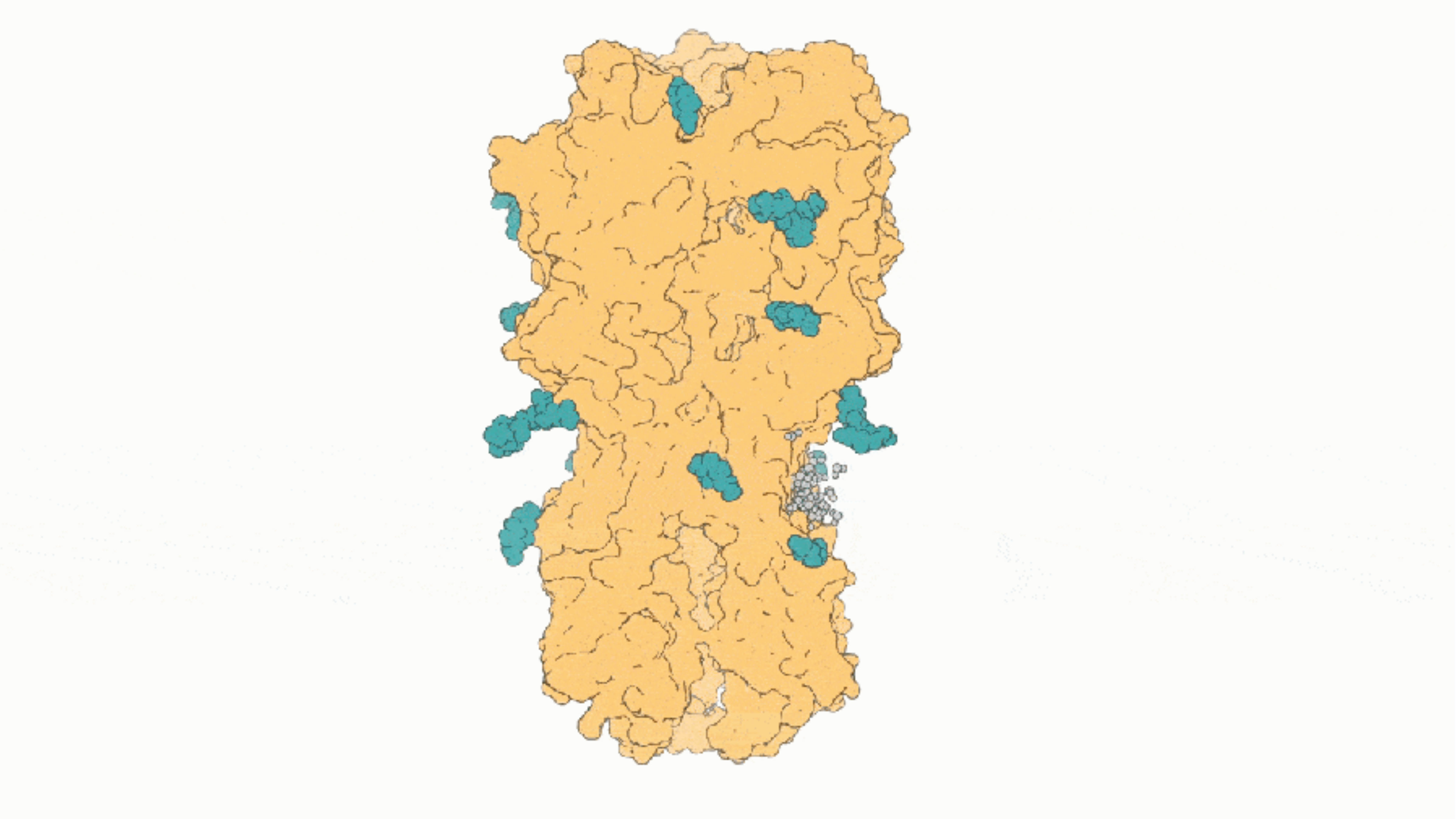
ROHITH KRISHNA, JUE WANG, WOODY AHERN, PASCAL STURMFELS, PREETHAM VENKATESH, INDREK KALVET, GYU RIE LEE, FELIX S. MOREY-BURROWS, IVAN ANISHCHENKO, [...], AND DAVID BAKER

+12 authors Authors Info & Affiliations

Atomically accurate de novo design of single-domain antibodies

Nathaniel R. Bennett[‡1,2,3], Joseph L. Watson[*‡1,2], Robert J. Ragotte[‡1,2], Andrew J. Borst[‡1,2], Déjenaé L. See[#1,2,4], Connor Weidle[#1,2], Riti Biswas[1,2,3], Ellen L. Shrock[1,2], Philip J. Y. Leung[1,2,3], Buwei Huang[1,2,4], Inna Goreshnik[1,2,5], Russell Ault[6,7], Kenneth D. Carr[2], Benedikt Singer[1,2], Cameron Criswell[1,2], Dionne Vafeados[2], Mariana Garcia Sanchez[2], Ho Min Kim[8,9], Susana Vázquez Torres[1,2,10], Sidney Chan[2], David Baker[*1,2,5]



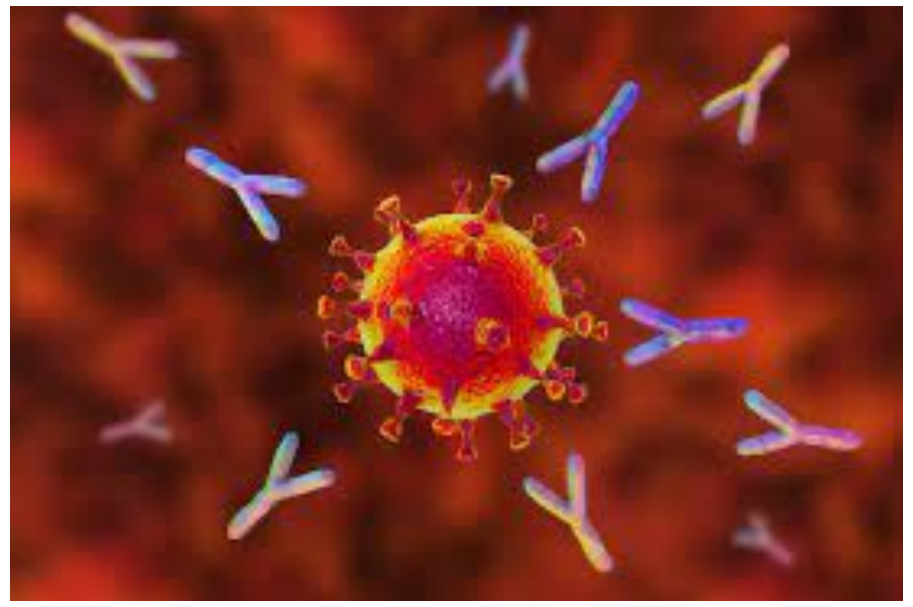


Takeaway

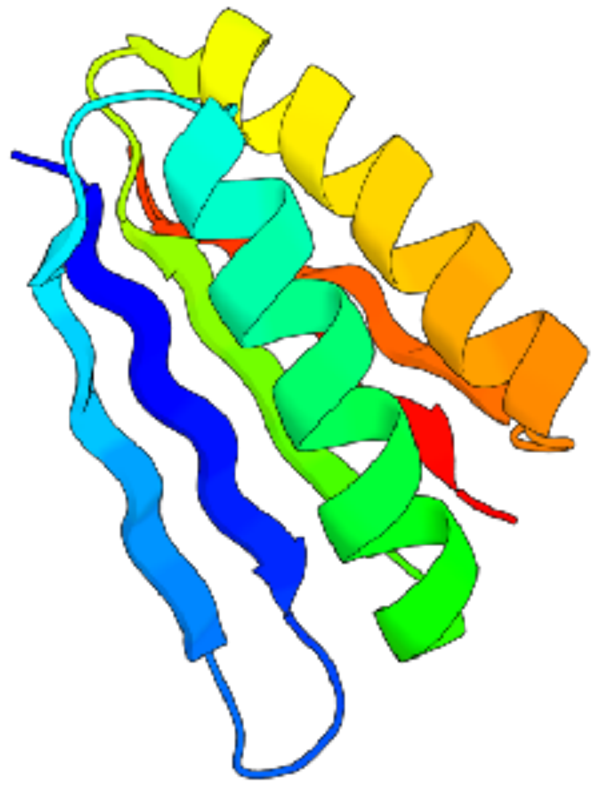
Desiderata

1. Generate **high quality** structures. ✅
2. Generate **diverse** structures. ✅
3. Generate **novel** structures. ✅
4. Generate **functional** structures. ✅

Towards co-design



RFdiffusion / FrameFlow

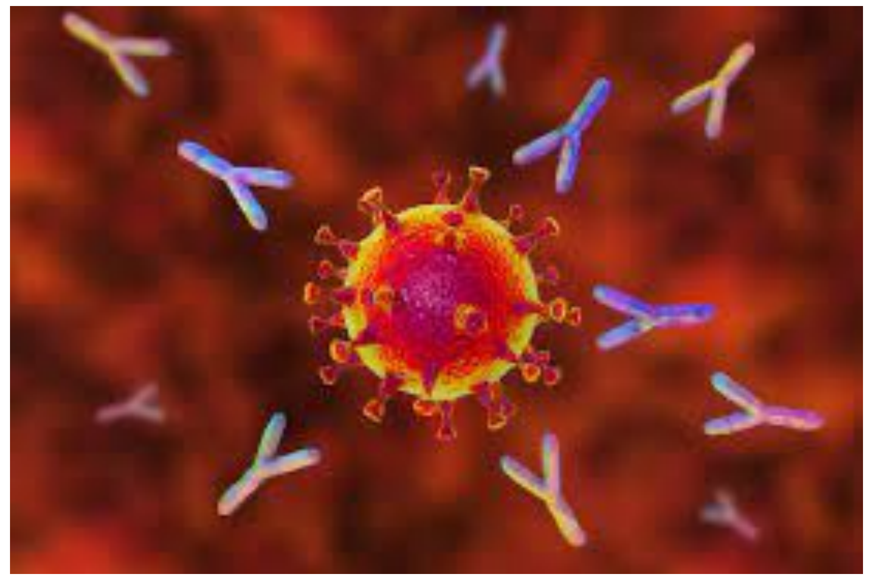


ProteinMPNN

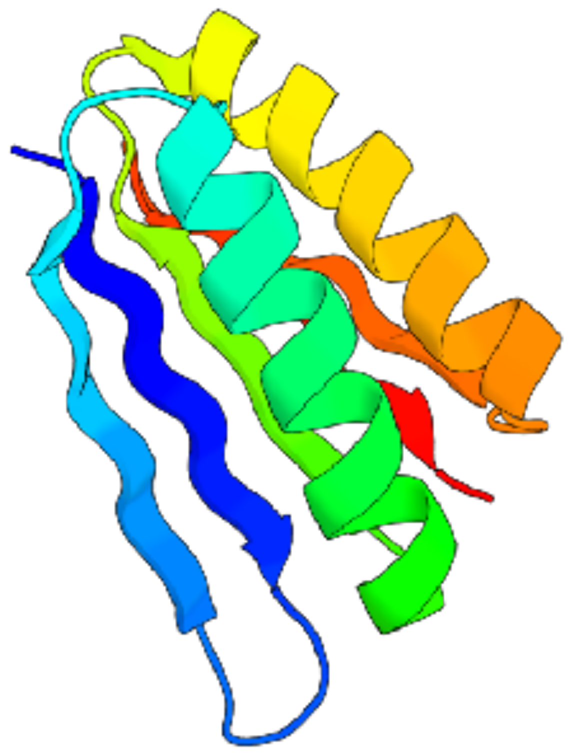
Function

Structure

Sequence



MultiFlow



Generate both sequence and structure jointly (i.e. codesign)

Function

Structure

Sequence

MultiFlow

Translations:

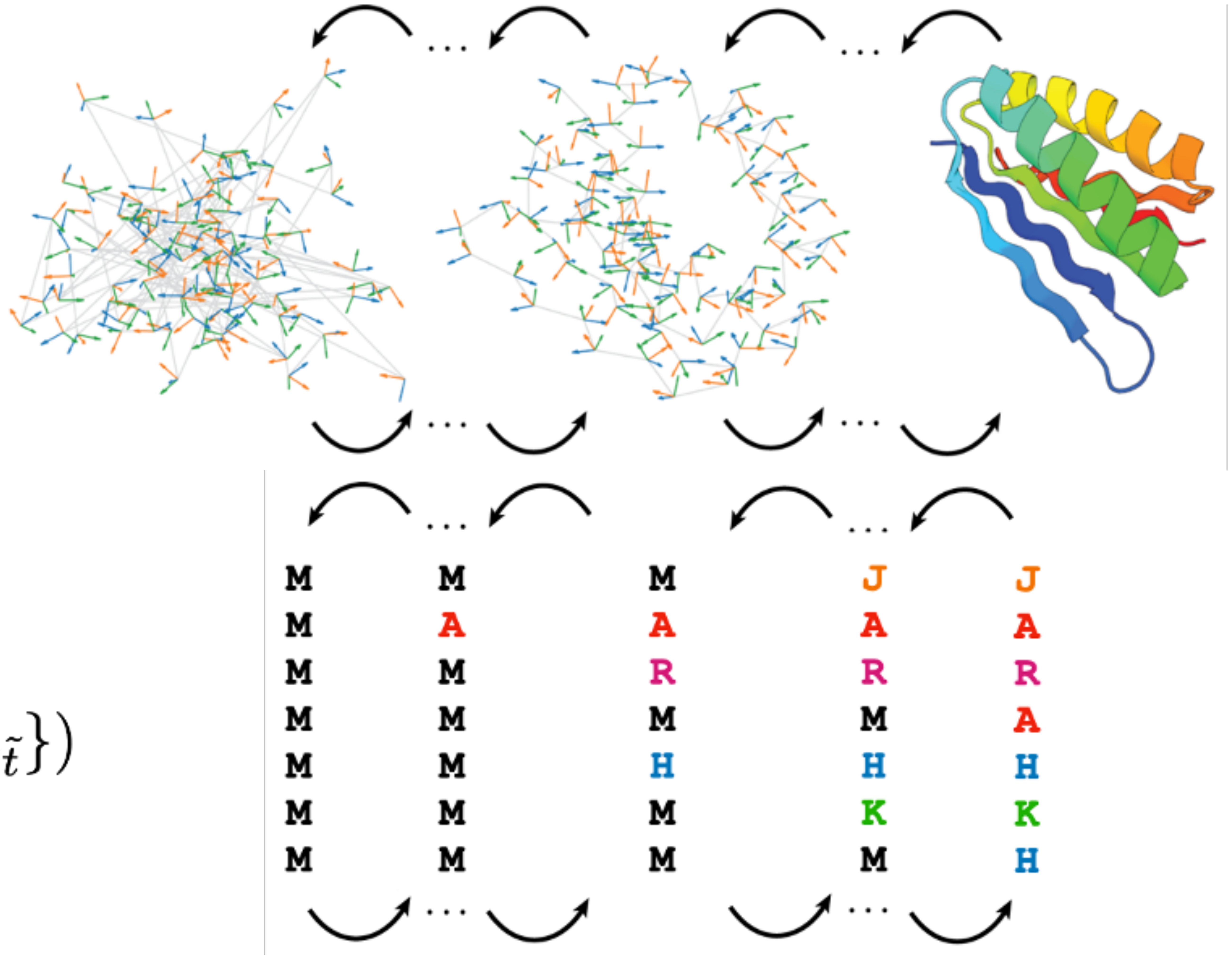
$$x_t = t x_1 + (1-t) x_0$$

Rotations:

$$r_t = \exp_{r_0}\left(t \log_{r_0}(r_1)\right)$$

Sequence:

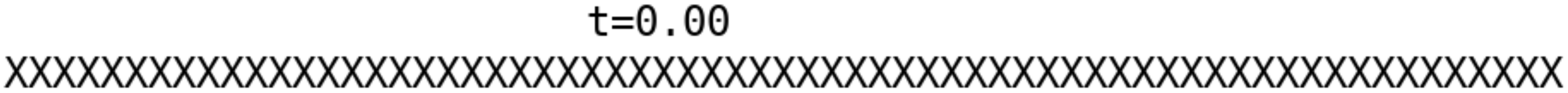
$$a_{\tilde{t}} \sim \mathrm{Cat}(\tilde{t}\delta\{a_1, a_{\tilde{t}}\} + (1-\tilde{t})\delta\{\texttt{M}, a_{\tilde{t}}\})$$



M	M	M	J	J
M	A	A	A	A
M	M	R	R	R
M	M	M	M	A
M	M	H	H	H
M	M	M	K	K
M	M	M	M	H

Our approach: discrete flow matching

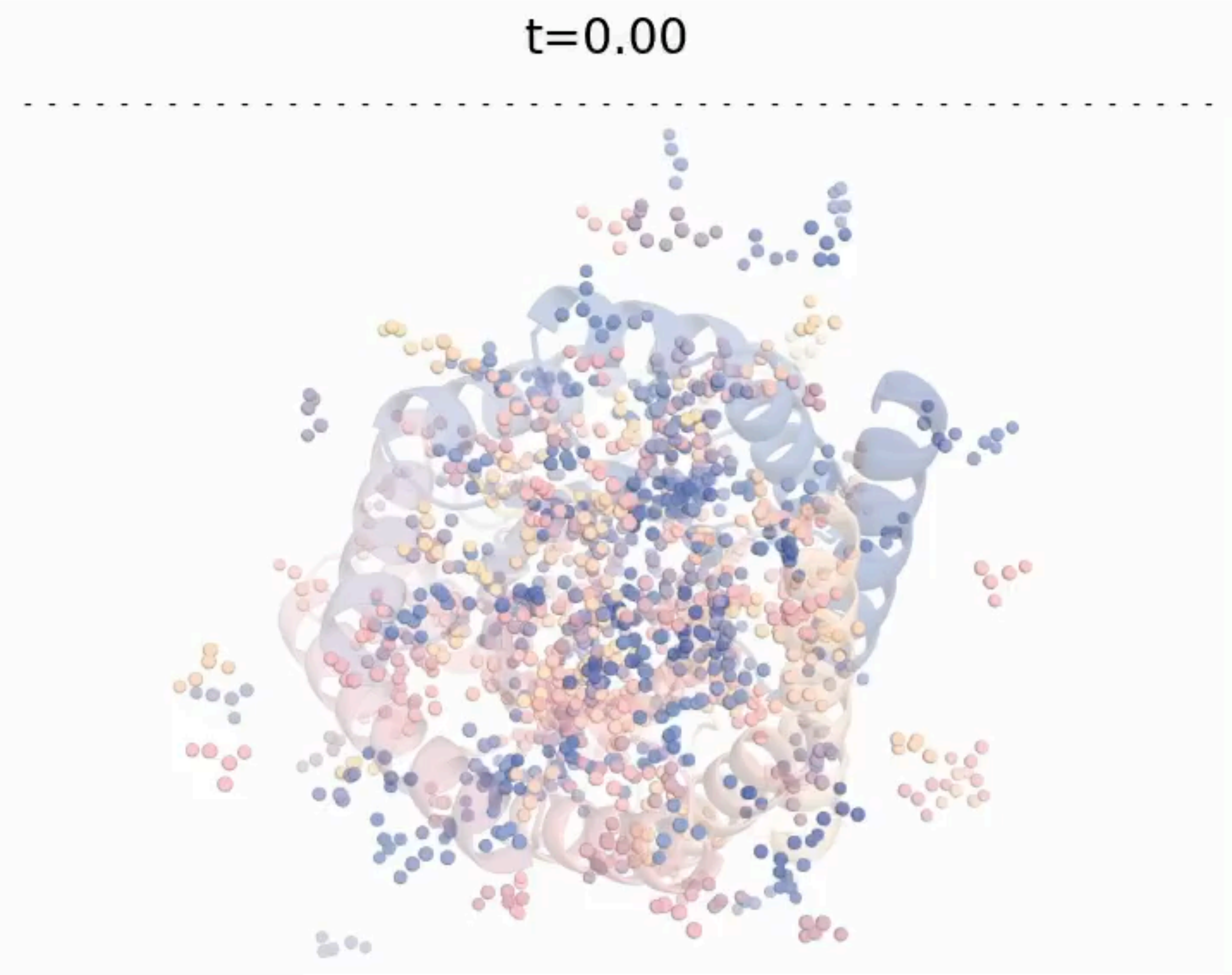
Continuous time generative model over discrete data



```
t=0.00
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
```

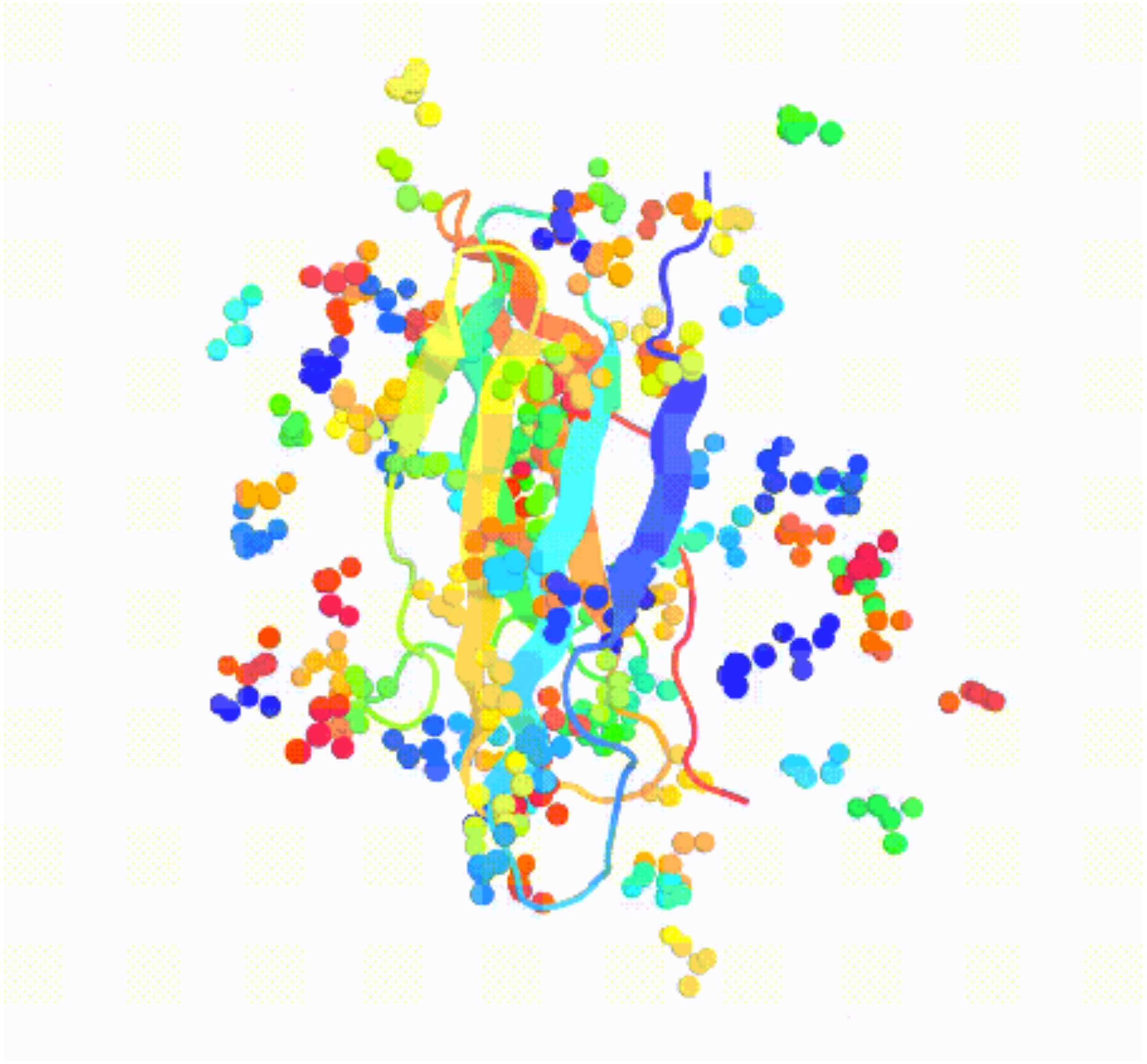
Sequence and structure co-design

MultiFlow



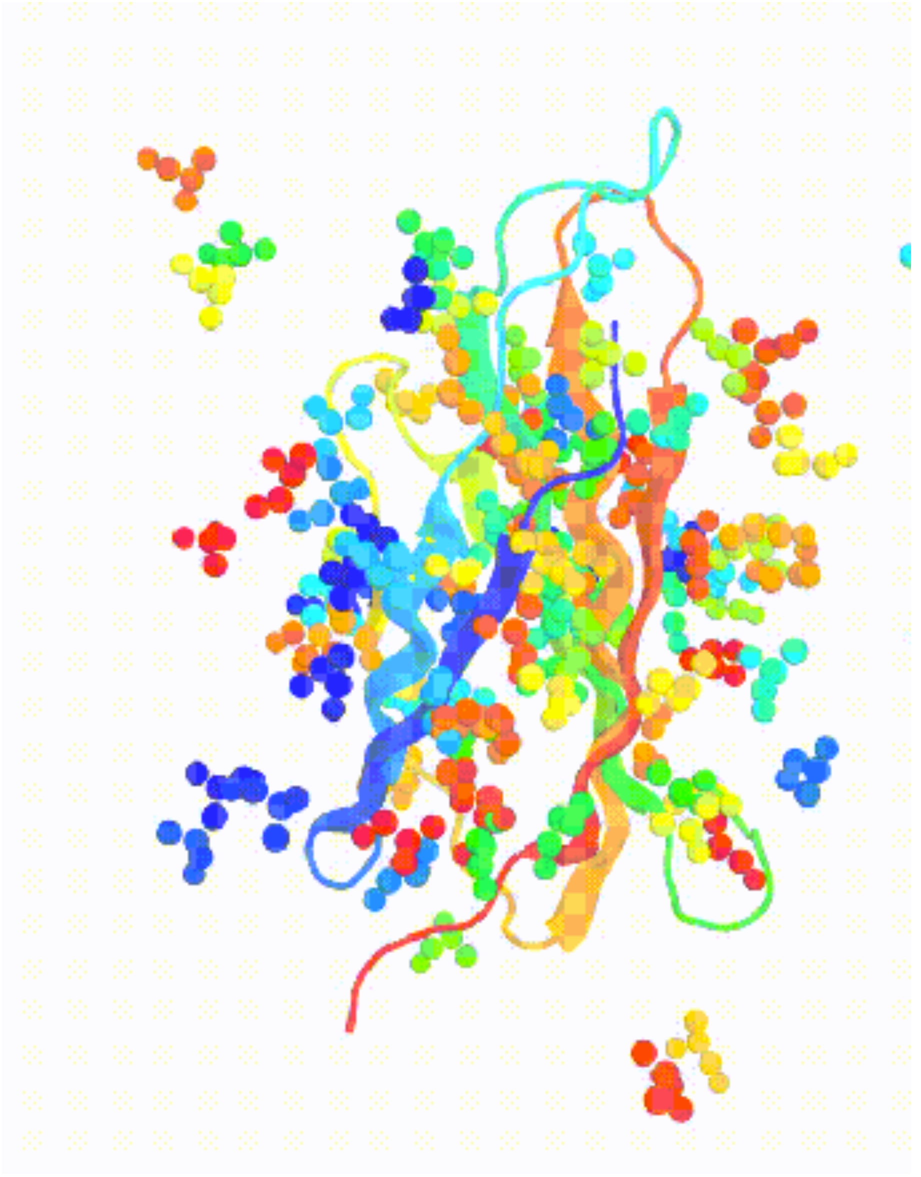
Technical Summary

Diffusion: FrameDiff



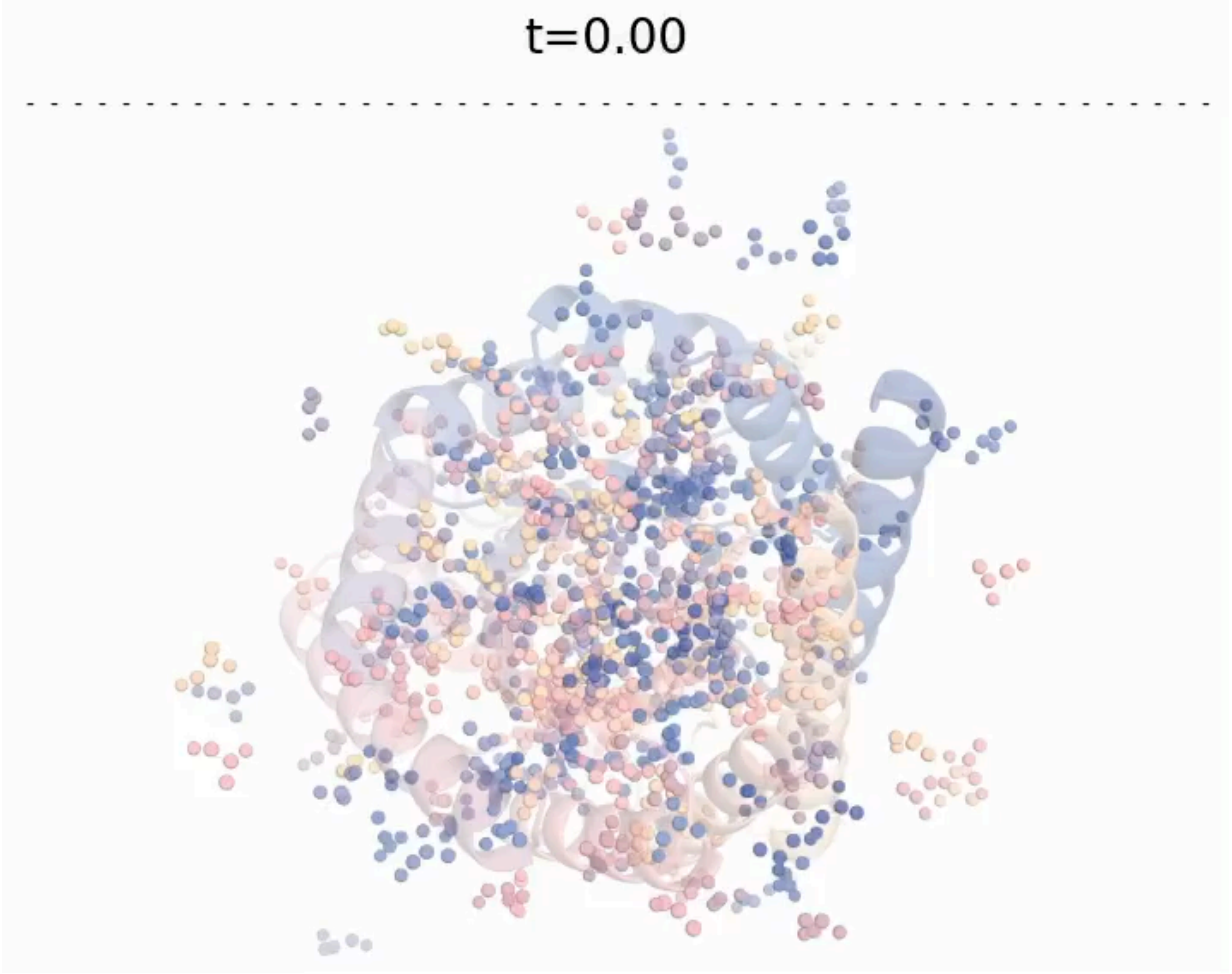
Stochastic Differential Equation (SDE)

(Riemannian) Flows: FrameFlow



Ordinary Differential equation (ODE)

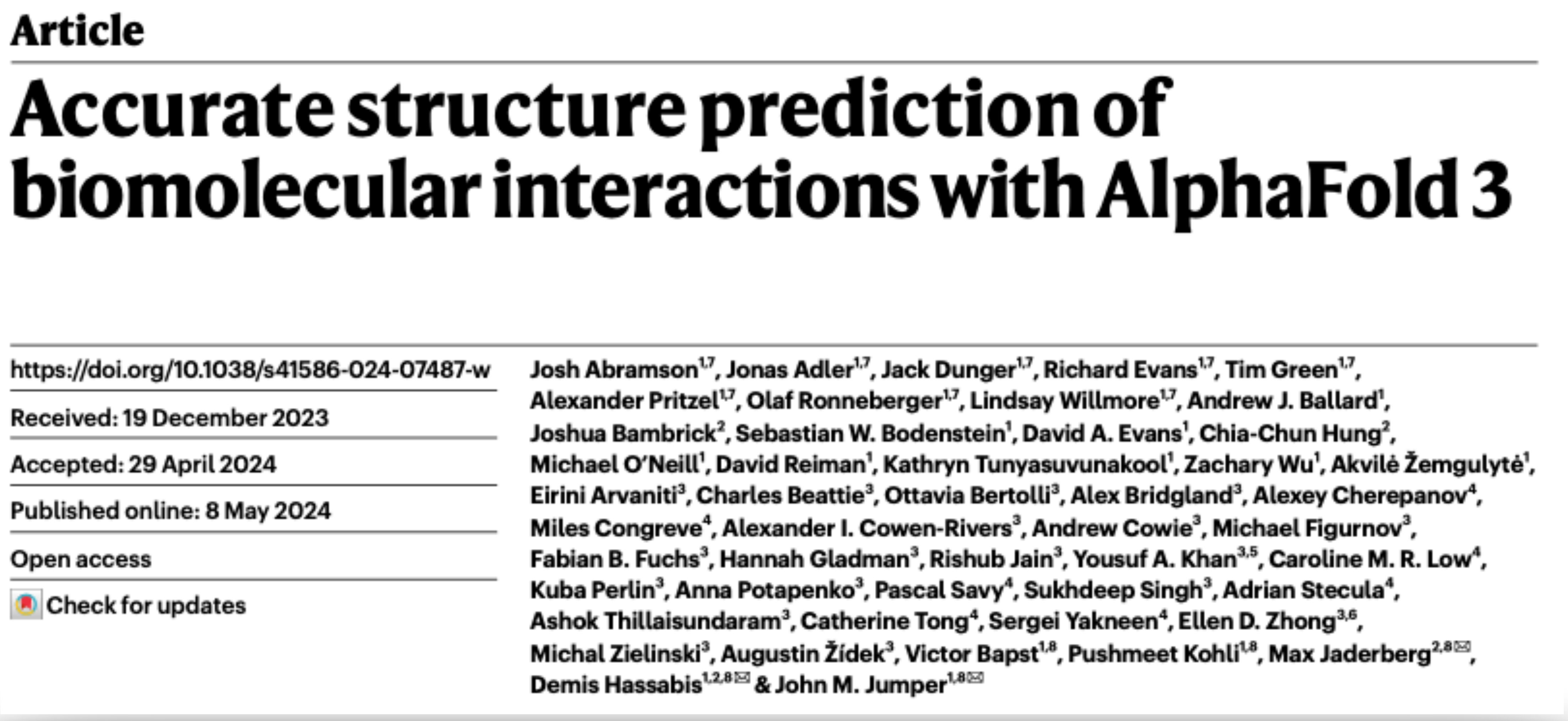
Discrete Flows: MultiFlow



Continuous Time Markov Chain (CTMC)

What’s next?

Going beyond proteins



Article

Accurate structure prediction of biomolecular interactions with AlphaFold 3

https://doi.org/10.1038/s41586-024-07487-w

Received: 19 December 2023

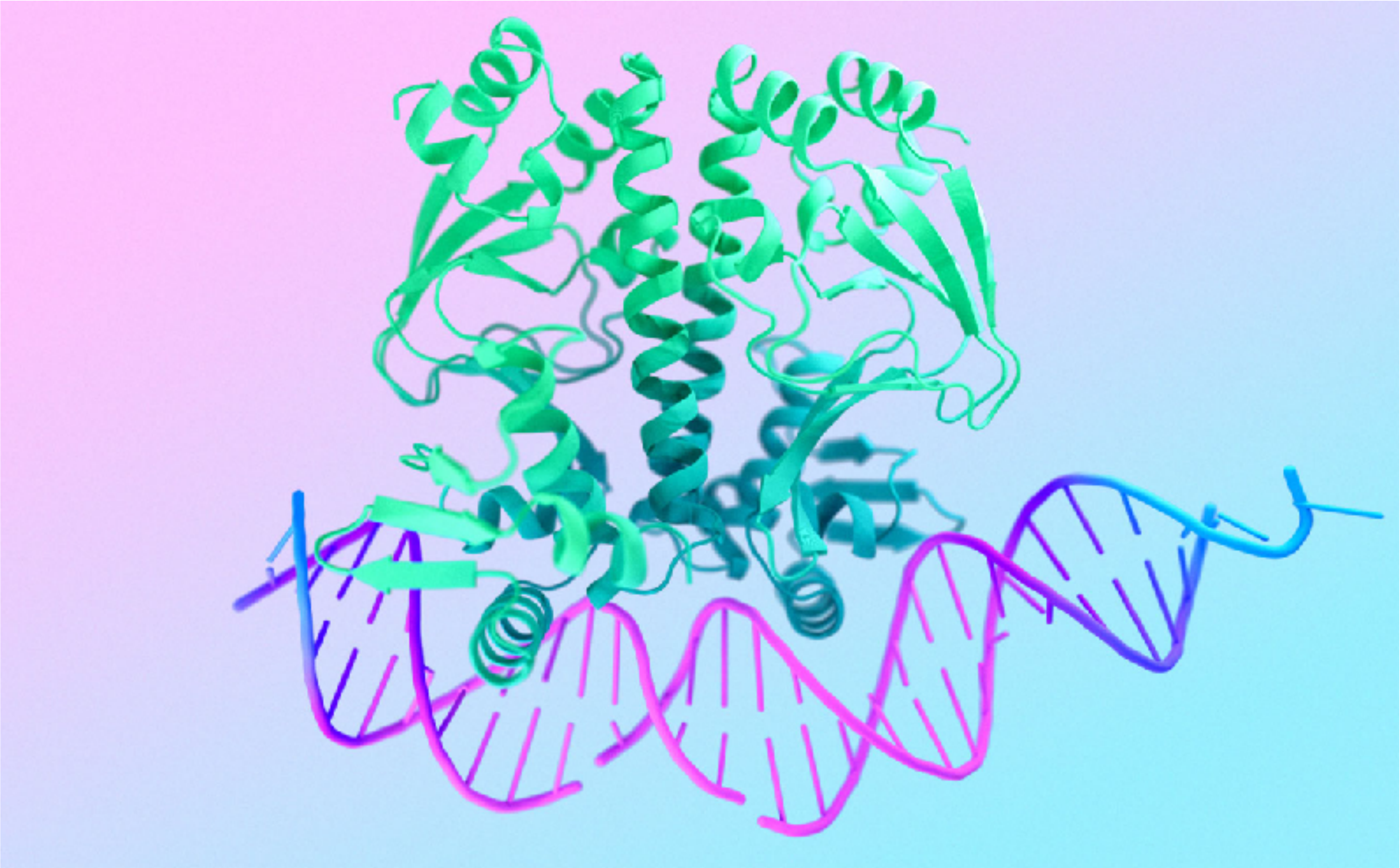
Accepted: 29 April 2024

Published online: 8 May 2024

Open access

Check for updates

Josh Abramson[1,7], Jonas Adler[1,7], Jack Dunger[1,7], Richard Evans[1,7], Tim Green[1,7], Alexander Pritzel[1,7], Olaf Ronneberger[1,7], Lindsay Willmore[1,7], Andrew J. Ballard[1], Joshua Bambrick[2], Sebastian W. Bodenstein[1], David A. Evans[1], Chia-Chun Hung[2], Michael O’Neill[1], David Reiman[1], Kathryn Tunyasuvunakool[1], Zachary Wu[1], Akvilė Žemgulytė[1], Eirini Arvaniti[3], Charles Beattie[3], Ottavia Bertolli[3], Alex Bridgland[3], Alexey Cherepanov[4], Miles Congreve[4], Alexander I. Cowen-Rivers[3], Andrew Cowie[3], Michael Figurnov[3], Fabian B. Fuchs[3], Hannah Gladman[3], Rishub Jain[3], Yousuf A. Khan[3,5], Caroline M. R. Low[4], Kuba Perlin[3], Anna Potapenko[3], Pascal Savy[4], Sukhdeep Singh[3], Adrian Stecula[4], Ashok Thillaisundaram[3], Catherine Tong[4], Sergei Yakneen[4], Ellen D. Zhong[3,6], Michal Zielinski[3], Augustin Žídek[3], Victor Bapst[1,8], Pushmeet Kohli[1,8], Max Jaderberg[2,8]✉, Demis Hassabis[1,2,8]✉ & John M. Jumper[1,8]✉

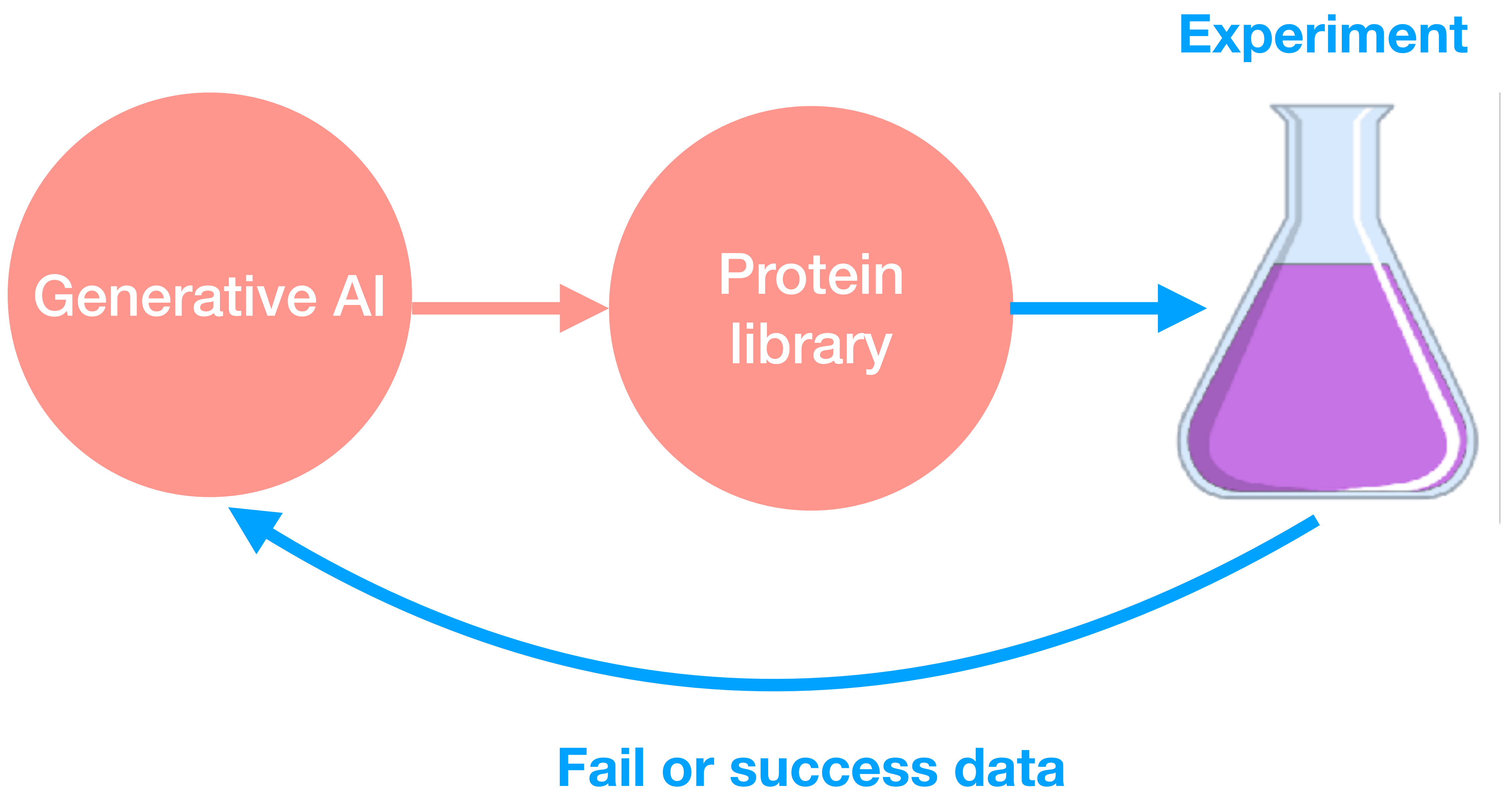


- AlphaFold3 is also a diffusion model!

What's next?

Fine-tuning / post-training

1. Generate diverse set of functional proteins
2. Learn from experiments and iteratively improve



References

[1] **Jason Yim***, Brian L. Trippe*, Valentin De Bortoli*, Emile Mathieu*, Arnaud Doucet, Regina Barzilay, Tommi S. Jaakkola. *SE (3) diffusion model with application to protein backbone generation*. International Conference of Machine Learning, July 23, 2023.

[2] **Jason Yim**, Andrew Campbell, Emile Mathieu, Andrew Y. K. Foong, Michael Gastegger, Jose Jimenez-Luna, Sarah Lewis, Victor Garcia Satorras, Bastian S. Veeling, Regina Barzilay, Frank Noe, Tommi S. Jaakkola. *Improved motif-scaffolding with SE(3) flow matching*. Transactions on Machine Learning Research, July 18, 2024.

[3] Andrew Campbell*, **Jason Yim***, Regina Barzilay, Tom Rainforth, Tommi Jaakkola. Generative Flows on Discrete State-Spaces: Enabling Multimodal Flows with Applications to Protein Co-Design. International Conference of Machine Learning, July 23, 2024.

[4] Watson, J. L.*, Juergens, D.*, Bennett, N. R.*, Trippe, B. L.*, **Yim, J**.*, Eisenach, H. E.*, ... & Baker, D. (2023). De novo design of protein structure and function with RFdiffusion. Nature, 620(7976), 1089-1100.

** Equal contribution*

Contact

- **Email**:
  - Work: jyim@mit.edu
  - Personal: jasonkyuyim@gmail.com

- **Website:** https://people.csail.mit.edu/jyim/
- **X / Twitter**: https://x.com/json_yim
- **Linkedin**: https://www.linkedin.com/in/jason-yim-92173b97/

- **Advisors:** Regina Barzilay, Tommi Jaakkola