Diffusion models

Diffusion models for image generation

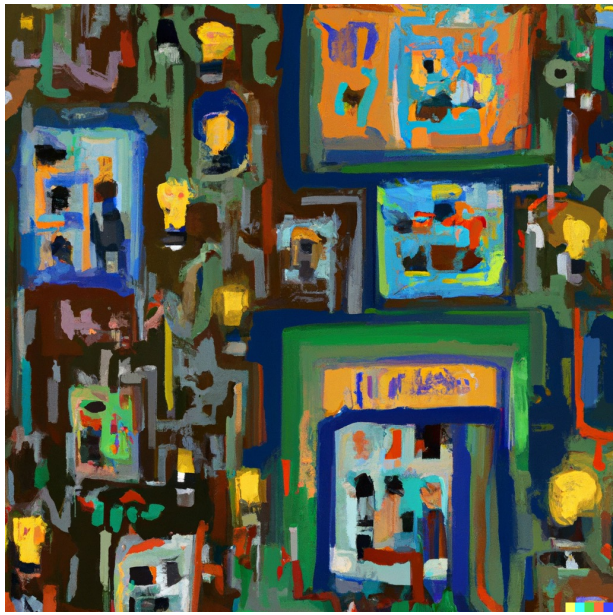
Diffusion models for image generation

Prompt: “an abstract painting of a room full of proteins learning machine learning”

Diffusion models for image generation

Prompt: “an abstract painting of a room full of proteins learning machine learning”



Generated by DALL-E

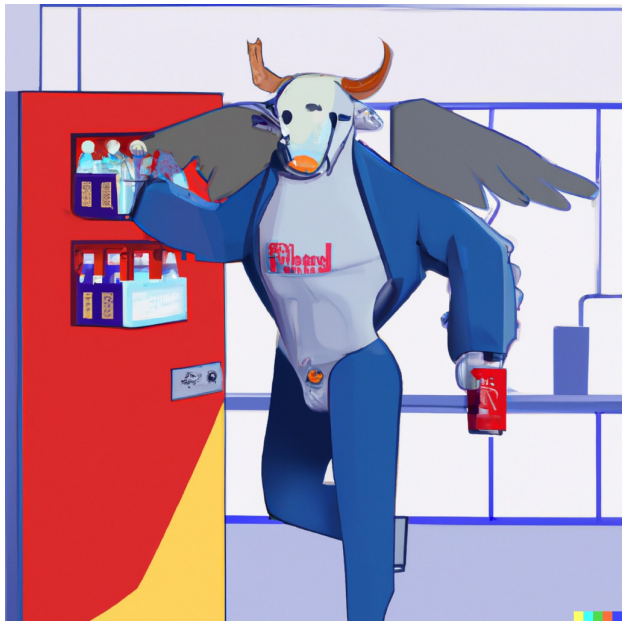
Diffusion models for image generation

Prompt: “a humanized-redbull drinking a redbull with a redbull-themed fridge full of redbull in the background”

Generated by DALL-E

Diffusion models for image generation

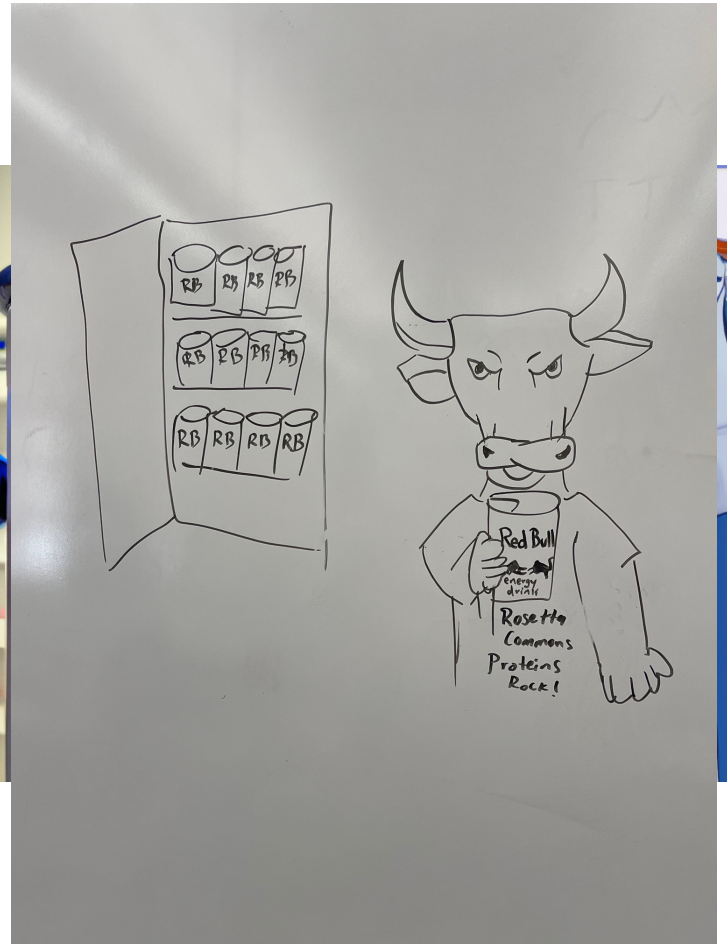
Prompt: “a humanized-redbull drinking a redbull with a redbull-themed fridge full of redbull in the background”




Generated by DALL-E

Diffusion models for image generation

Prompt: “a humanized-redbull drinking a redbull with a redbull-themed fridge full of redbull in the background”

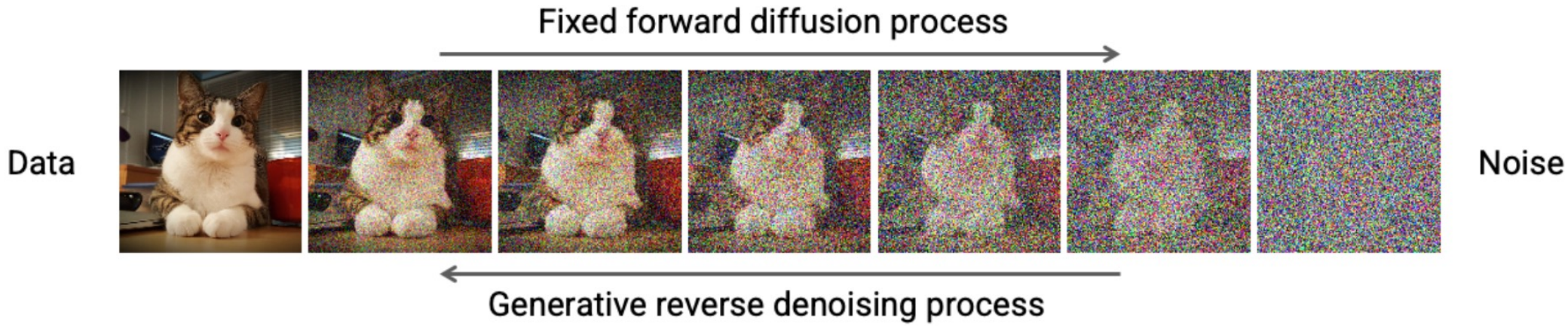


Generated by DALL-E

Diffusion models for image generation

Diffusion models iteratively add noise to a set of data and learn how to denoise by reversing this process.

The added noise is Gaussian noise.



Once a diffusion model is trained, we can endlessly generate “fake” data from the input (randomly) sampled noise.

Denoising Diffusion Probabilistic Models

Denoising Diffusion Probabilistic Models

Jonathan Ho
UC Berkeley
jonathanho@berkeley.edu

Ajay Jain
UC Berkeley
ajayj@berkeley.edu

Pieter Abbeel
UC Berkeley
pabbeel@cs.berkeley.edu

https://arxiv.org/pdf/2006.11239.pdf

Denoising Diffusion Probabilistic Models

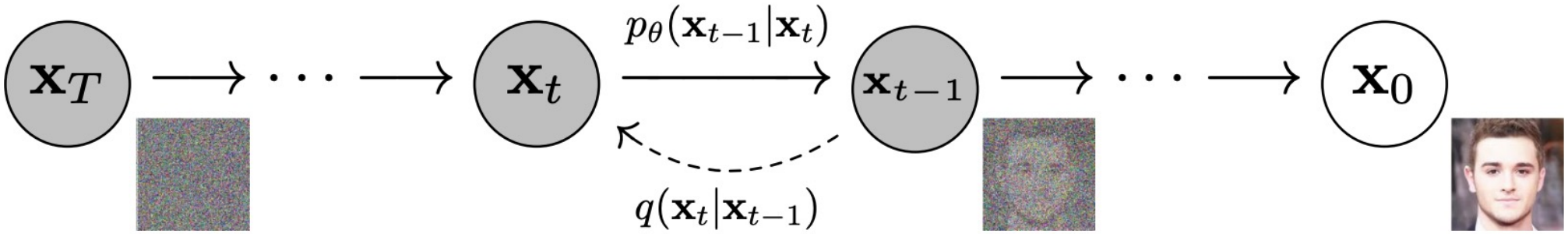


Figure 2: The directed graphical model considered in this work.

We train diffusion models to take lossy compressions of images and produce high-quality reconstruction.

What about recapitulation?

https://arxiv.org/pdf/2006.11239.pdf

Denoising Diffusion Probabilistic Models

Learning how to reverse the noise allows models to produce high-quality reconstruction of input images.

https://arxiv.org/pdf/2006.11239.pdf

Diffusion models for protein design

Introducing RFdiffusion…

Diffusion models for protein design are extremely new!

arXiv:2205.15019v1 [q-bio.QM] 26 May 2022

Protein Structure and Sequence Generation with Equivariant Denoising Diffusion Probabilistic Models

Namrata Anand
namrata.anand2@gmail.com

Tudor Achim
tachim@cs.stanford.edu

Abstract

Proteins are macromolecules that mediate a significant fraction of the cellular processes that underlie life. An important task in bioengineering is designing proteins with specific 3D structures and chemical properties which enable targeted functions. To this end, we introduce a generative model of both protein structure and sequence that can operate at significantly larger scales than previous molecular generative modeling approaches. The model is learned entirely from experimental data and conditions its generation on a compact specification of protein topology to produce a full-atom backbone configuration as well as sequence and side-chain predictions. We demonstrate the quality of the model via qualitative and quantitative analysis of its samples. Videos of sampling trajectories are available at `https://nanand2.github.io/proteins`.

1 Introduction

Proteins are large macromolecules that play fundamental roles in nearly all cellular processes. Two key scientific challenges related to these molecules are characterizing the set of all naturally-occurring proteins based on sequences collected at scale and designing new proteins whose structure and sequence achieve functional goals specified by the researcher. Recently, AlphaFold2, a purely data-driven machine learning approach, has shown great progress in the forward problem of structure prediction [23]. Similarly, machine learning approaches have come to perform well for the sequence generation inverse problem [3, 21, 20]. However, for the task of structure generation, stochastic search algorithms based on handcrafted energy functions and heuristic sampling approaches are still in wide use [25, 34, 2, 26].

Data-driven generative modeling approaches have not yet had the same impact in the protein modeling setting as they have in the image generation setting because of several key differences. First, unlike images, proteins do not have a natural representation on a discretized grid that is amenable to straightforward application of existing generative models. Interpreting the pairwise distance matrix of a protein's atoms as an image to be modeled with existing models has seen limited success because inconsistencies in the predictions lead to nontrivial errors when optimization routines are used to recover the final 3D structures [4]. Second, unlike images, proteins have no natural canonical orientation. As a result, methods that are not rotationally invariant must account for this factor of variation directly in the model weights, which reduces the effective model capacity that can be dedicated to the structural variation of interest. Finally, in protein generation, nontrivial errors in local *or* global structure lead to implausible protein structures.

Previous work has made progress on different aspects of the problem. Rotamer packing has benefited from machine learning approaches [28, 14, 1, 32, 3]. Machine learning has also made an impact on sequence design both in the case of conditioning on structural information [3, 21], and without [17, 9, 37, 38, 16, 18, 31, 20]. However, 3D molecular structure generation is a more challenging

Preprint. Under review.

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.01.518682; this version posted December 2, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Illuminating protein space with a programmable generative model

John Ingraham, Max Baranov, Zak Costello, Vincent Frappier, Ahmed Ismail, Shan Tie, Wujie Wang, Vincent Xue, Fritz Obermeyer, Andrew Beam, Gevorg Grigoryan

Generate Biomedicines

December 1, 2022

Abstract

Three billion years of evolution have produced a tremendous diversity of protein molecules, and yet the full potential of this molecular class is likely far greater. Accessing this potential has been challenging for computation and experiments because the space of possible protein molecules is much larger than the space of those likely to host function. Here we introduce Chroma, a generative model for proteins and protein complexes that can directly sample novel protein structures and sequences and that can be conditioned to steer the generative process towards desired properties and functions. To enable this, we introduce a diffusion process that respects the conformational statistics of polymer ensembles, an efficient neural architecture for molecular systems based on random graph neural networks that enables long-range reasoning with sub-quadratic scaling, equivariant layers for efficiently synthesizing 3D structures of proteins from predicted inter-residue geometries, and a general low-temperature sampling algorithm for diffusion models. We suggest that Chroma can effectively realize protein design as Bayesian inference under external constraints, which can involve symmetries, substructure, shape, semantics, and even natural language prompts. With this unified approach, we hope to accelerate the prospect of programming protein matter for human health, materials science, and synthetic biology.

Introduction

Protein molecules carry out most of the biological functions necessary for life, but inventing them is a complicated task that has taken evolution millions to billions of years. The field of computational protein design aims to shortcut this by automating the design of proteins for desired functions in a manner that is *programmable*. While there has been significant progress towards this goal over the past three decades [Kuhlman and Bradley, 2019, Huang et al., 2016], including the design of novel topologies, assemblies, binders, catalysts, and materials [Koga et al., 2012, Cao et al., 2022, Kries et al., 2013, Joh et al., 2014], most *de novo* designs have yet to approach

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.09.519842; this version posted December 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models

Joseph L. Watson[‡1,2], David Juergens[‡1,2,3], Nathaniel R. Bennett[‡1,2,3], Brian L. Trippe[‡2,4], Jason Yim[‡2,6], Helen E. Eisenach[‡1,2], Woody Ahern[‡1,2,7], 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[1,2,9], William Sheffler[1,2], Jue Wang[1,2], Preetham Venkatesh[1,2,8], Isaac Sappington[1,2,8], Susana Vázquez Torres[1,2,8], Anna Lauko[1,2,8], Valentin De Bortoli[9], Emile Mathieu[10], Regina Barzilay[6], Tommi S. Jaakkola[6], Frank DiMaio[1,2], Minkyung Baek[12], David Baker[*1,2,11]

[‡]Equal contribution
[*]To whom correspondence should be addressed

1. Department of Biochemistry, University of Washington, Seattle, WA 98105, USA
2. Institute for Protein Design, University of Washington, Seattle, WA 98105, USA
3. Graduate Program in Molecular Engineering, University of Washington, Seattle, WA 98105, USA
4. Columbia University, Department of Statistics, New York, NY 10027, USA
5. Irving Institute for Cancer Design, Columbia University, New York, NY 10027, USA
6. Massachusetts Institute of Technology, Cambridge, MA 02139, USA
7. Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA 98105, USA
8. Graduate Program in Biological Physics, Structure and Design, University of Washington, Seattle, WA 98105, USA
9. Centre National de la recherche scientifique, École Normale Supérieure rue d'Ulm, Paris 75005, France
10. Department of Engineering, University of Cambridge, Cambridge CB2 1PZ, United Kingdom
11. Howard Hughes Medical Institute, University of Washington, Seattle, WA 98105, USA
12. School of Biological Sciences, Seoul National University, Seoul 08826, Republic of Korea

Abstract

There has been considerable recent progress in designing new proteins using deep learning methods[1–9]. Despite this progress, a general deep learning framework for protein design that enables solution of a wide range of design challenges, including *de novo* binder design and design of higher order symmetric architectures, has yet to be described. Diffusion models[10,11] have had considerable success in image and language generative modeling but limited success when applied to protein modeling, likely due to the complexity of protein backbone geometry and sequence-structure relationships. Here we show that by fine tuning the RoseTTAFold structure prediction network on protein structure denoising tasks, we obtain a generative model of protein backbones that achieves outstanding performance on unconditional and topology-constrained protein monomer design, protein binder design, symmetric oligomer

Diffusion models for protein design are extremely new!

Protein Structure and Sequence Generation with Equivariant Denoising Diffusion Probabilistic Models

Namrata Anand
namrata.anand2@gmail.com

Tudor Achim
tachim@cs.stanford.edu

Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models

Joseph L. Watson[‡1,2], David Juergens[‡1,2,3], Nathaniel R. Bennett[‡1,2,3], Brian L. Trippe[‡2,4], Jason Yim[‡2,6], Helen E. Eisenach[‡1,2], Woody Ahern[‡1,2,7], 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[1,2,9], William Sheffler[1,2], Jue Wang[1,2], Preetham Venkatesh[1,2,8], Isaac Sappington[1,2,8], Susana Vázquez Torres[1,2,8], Anna Lauko[1,2,8], Valentin De Bortoli[9], Emile Mathieu[10], Regina Barzilay[6], Tommi S. Jaakkola[6], Frank DiMaio[1,2], Minkyung Baek[12], David Baker*[1,2,11]

Illuminating protein space
with a programmable generative model

John Ingraham, Max Baranov, Zak Costello, Vincent Frappier,
Ahmed Ismail, Shan Tie, Wujie Wang, Vincent Xue, Fritz Obermeyer,
Andrew Beam, Gevorg Grigoryan

Generate Biomedicines

Anand/Achim Diffusion Model for Protein Generation

Protein Structure and Sequence Generation with Equivariant Denoising Diffusion Probabilistic Models

Namrata Anand
namrata.anand2@gmail.com

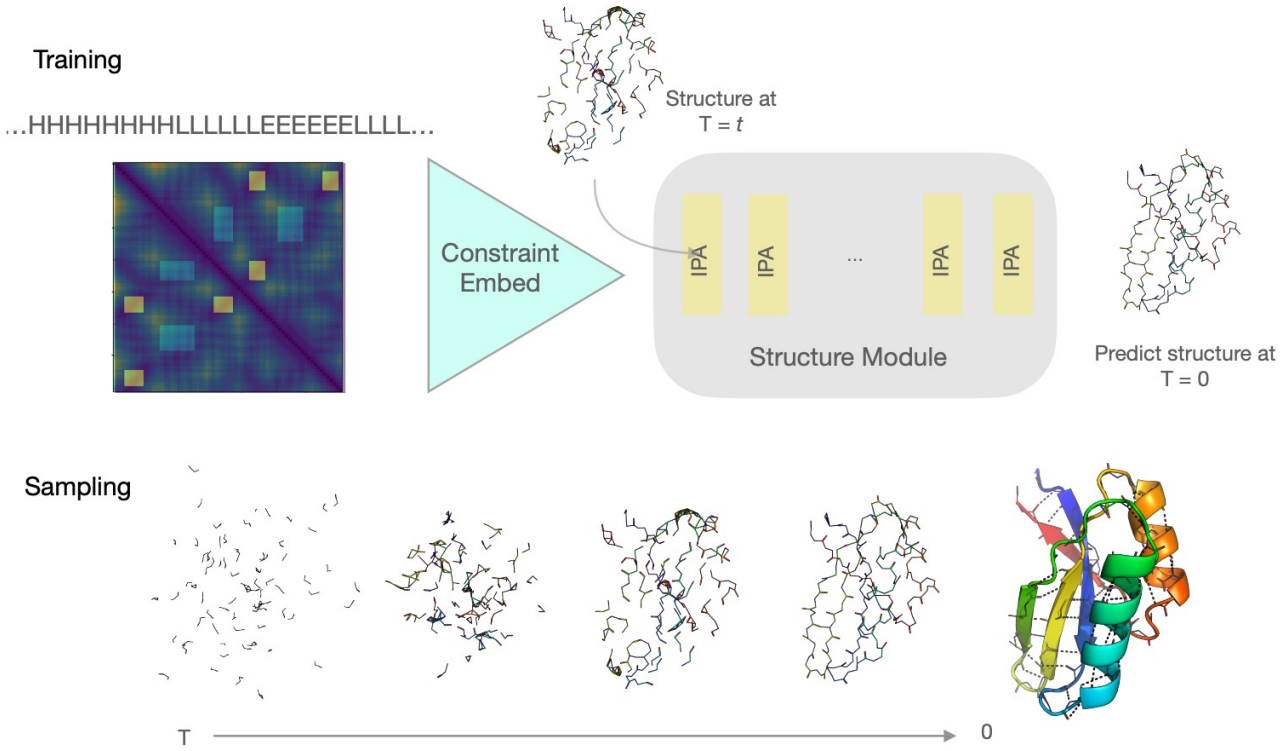
Tudor Achim
tachim@cs.stanford.edu

Anand/Achim Diffusion Model for Protein Generation

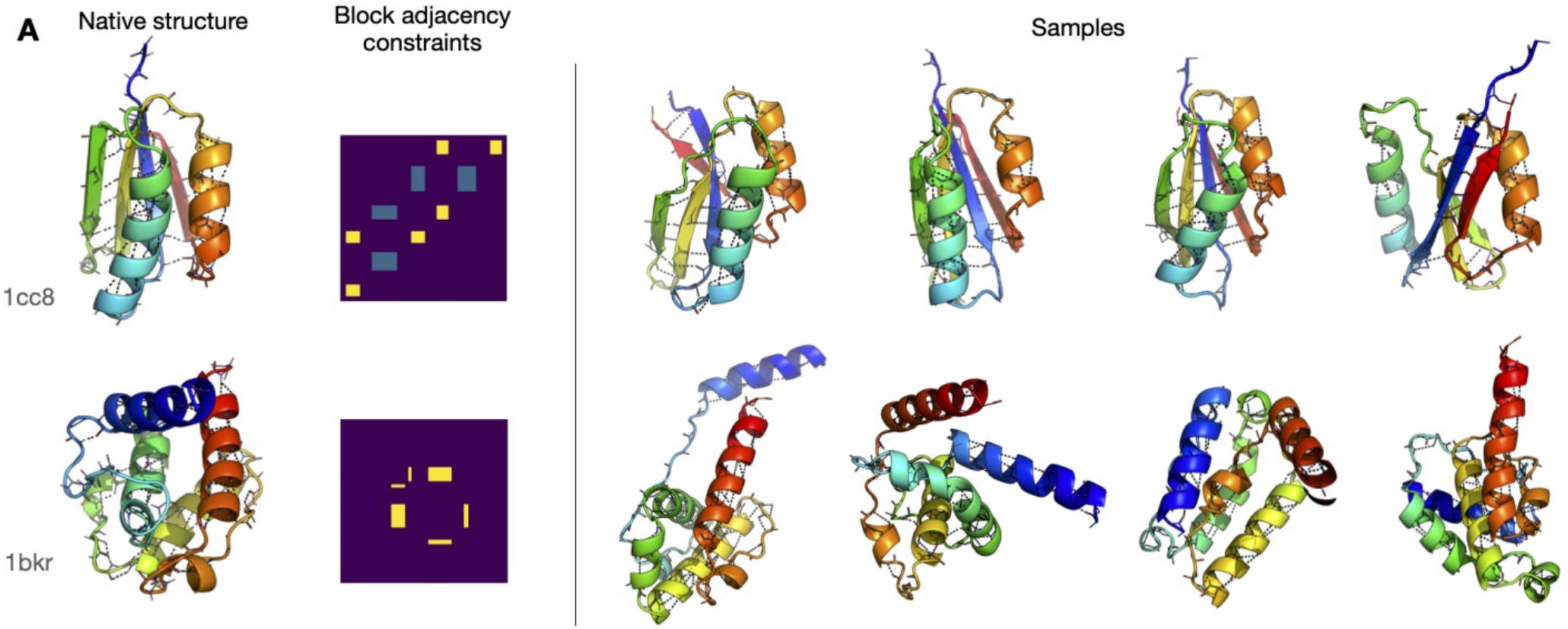
Protein Structure and Sequence Generation with Equivariant Denoising Diffusion Probabilistic Models

Namrata Anand
namrata.anand2@gmail.com

Tudor Achim
tachim@cs.stanford.edu



Anand/Achim Diffusion Model for Protein Generation



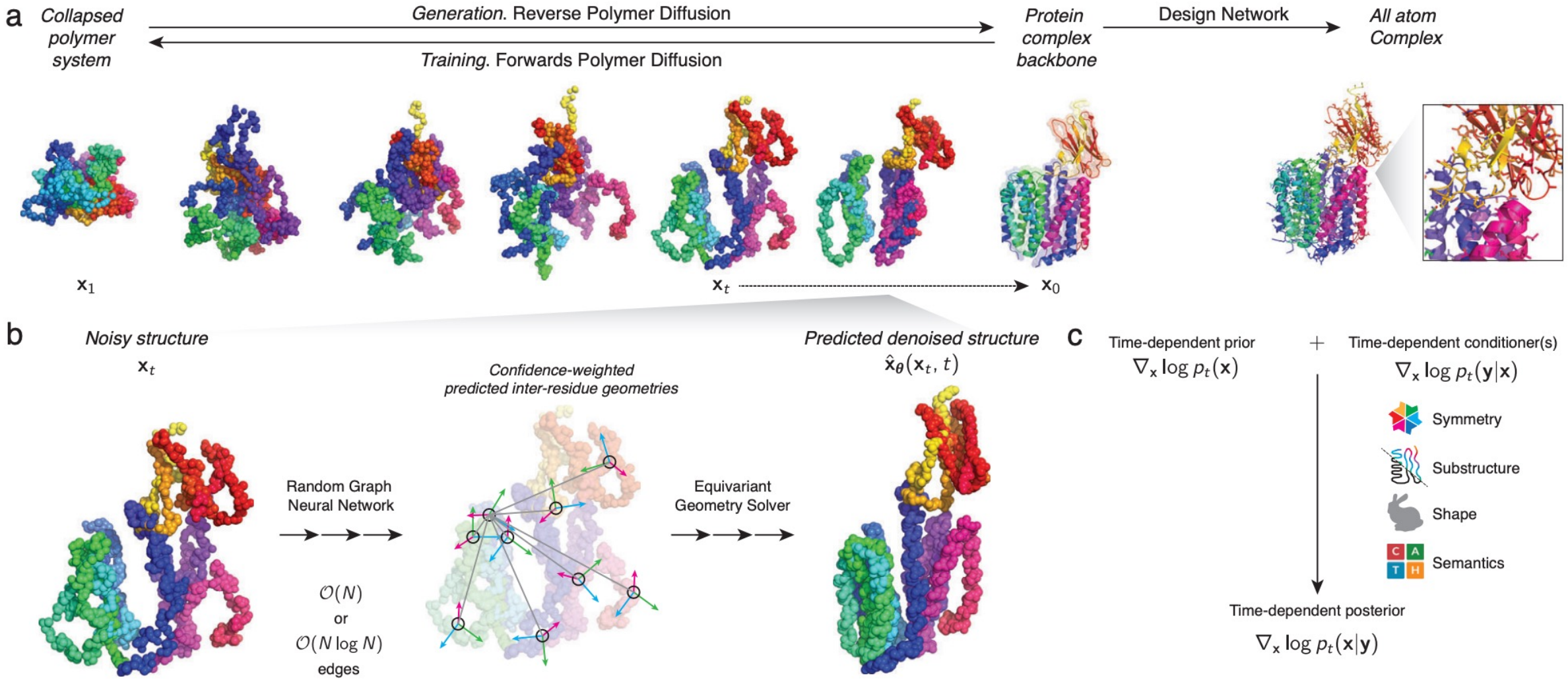
Generate's Diffusion Model for Protein Generation

Illuminating protein space
with a programmable generative model

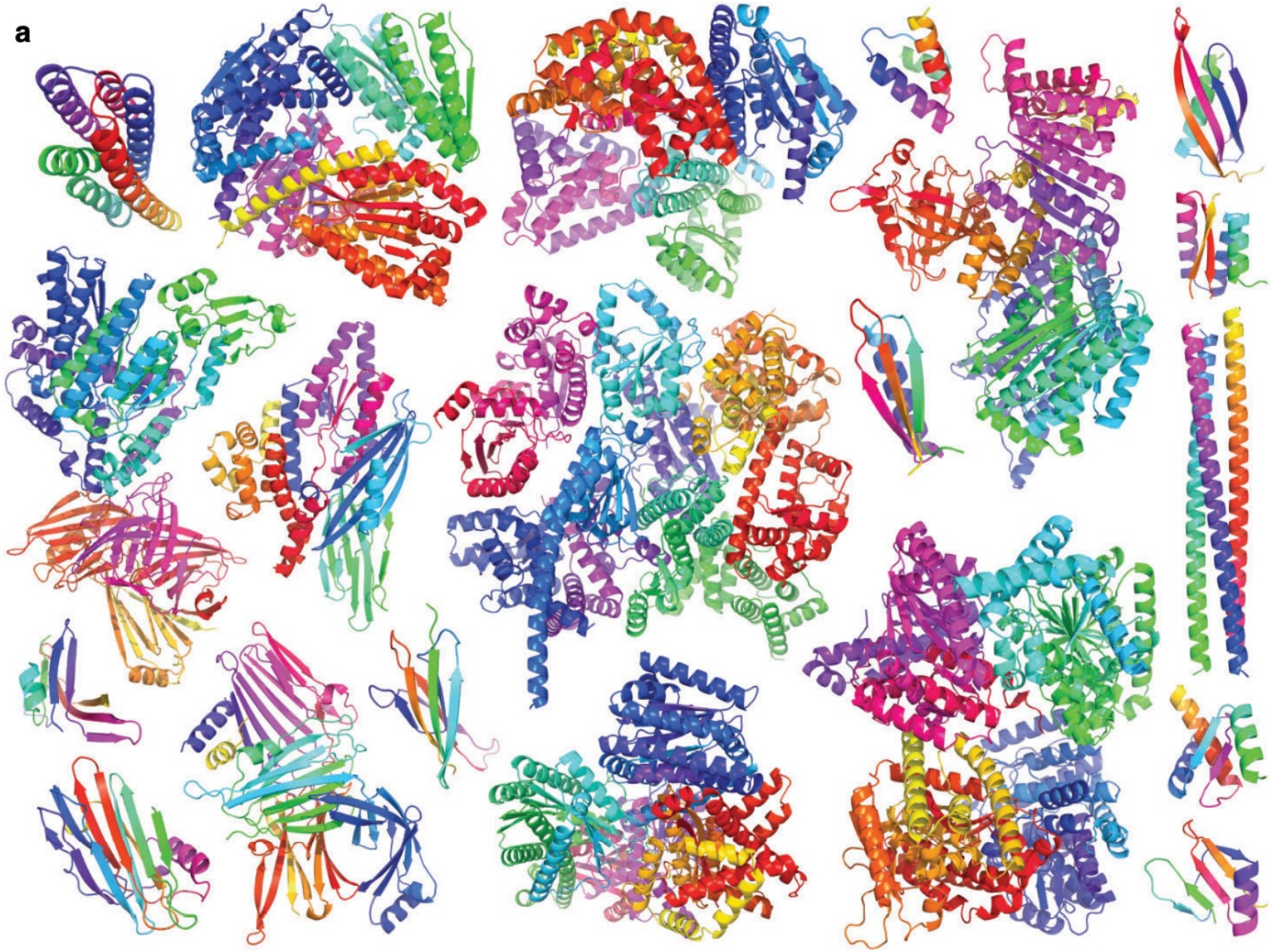
John Ingraham, Max Baranov, Zak Costello, Vincent Frappier,
Ahmed Ismail, Shan Tie, Wujie Wang, Vincent Xue, Fritz Obermeyer,
Andrew Beam, Gevorg Grigoryan

Generate Biomedicines

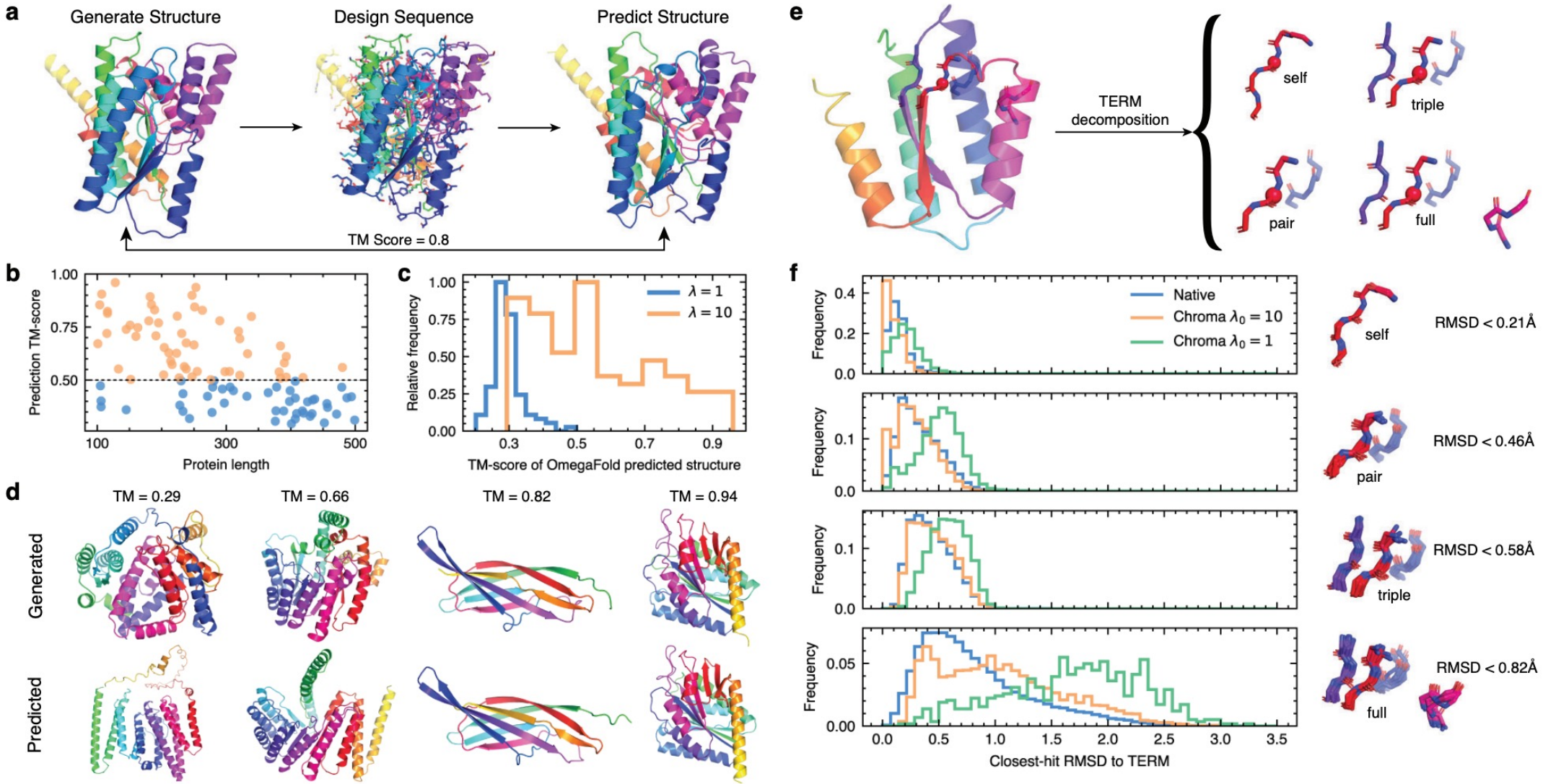
Generate's Diffusion Model for Protein Generation



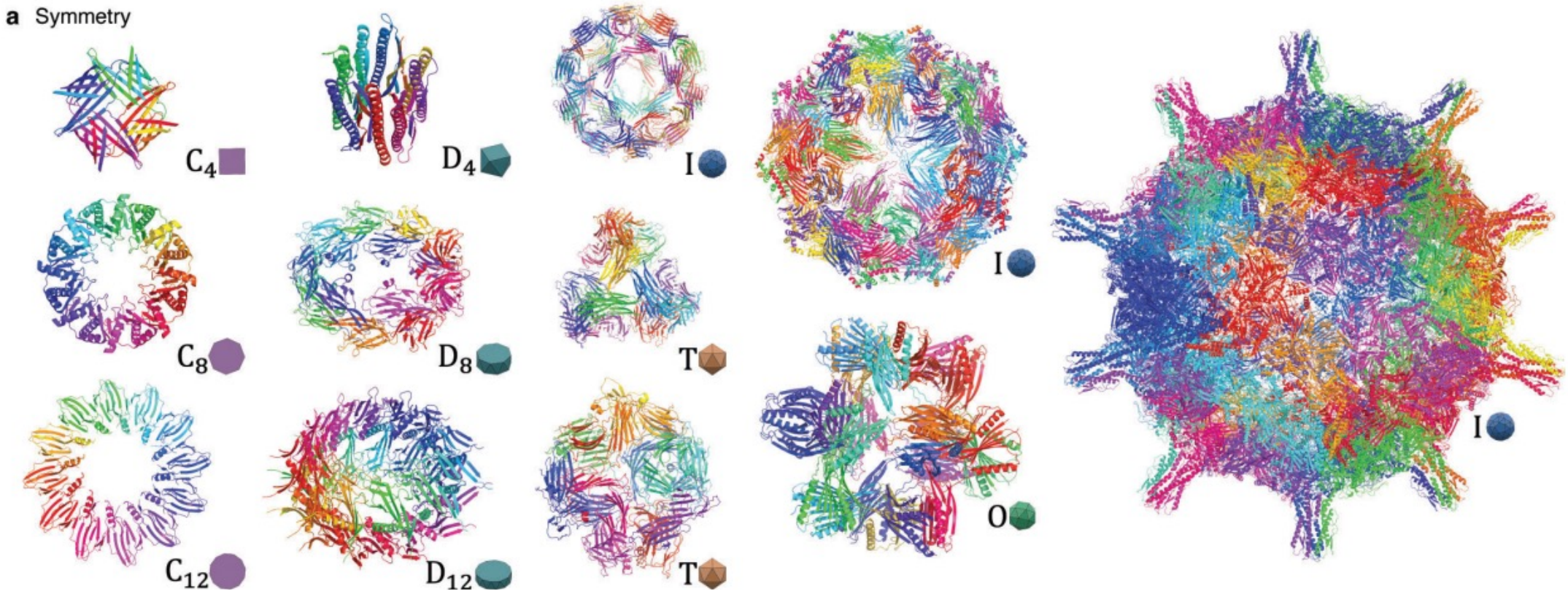
Generate's Diffusion Model for Protein Generation



Generate's Diffusion Model for Protein Generation

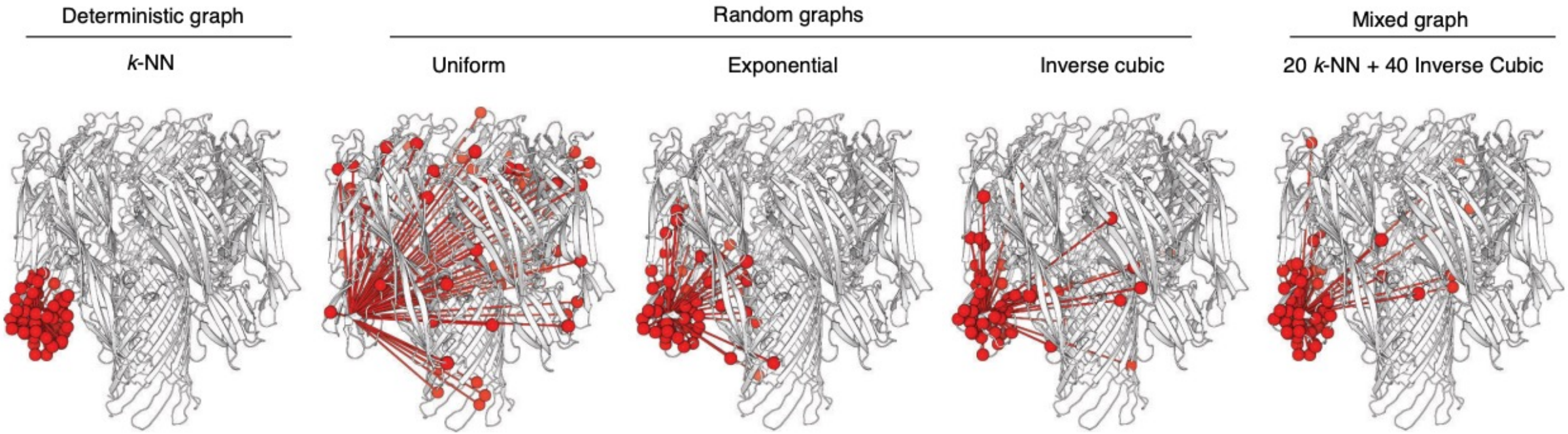


Generate's Diffusion Model for Protein Generation

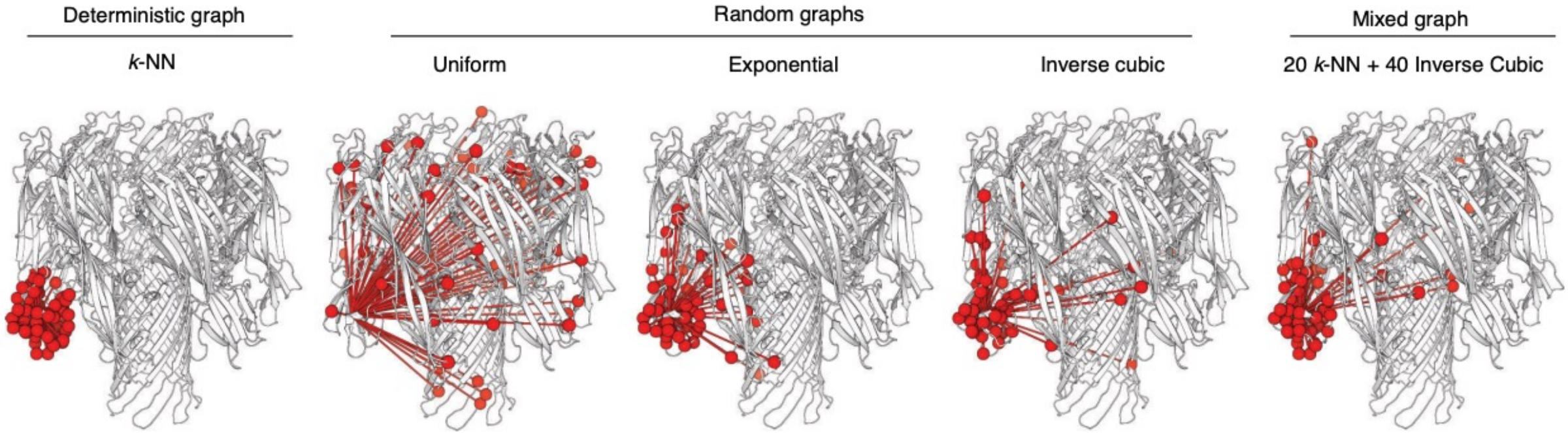


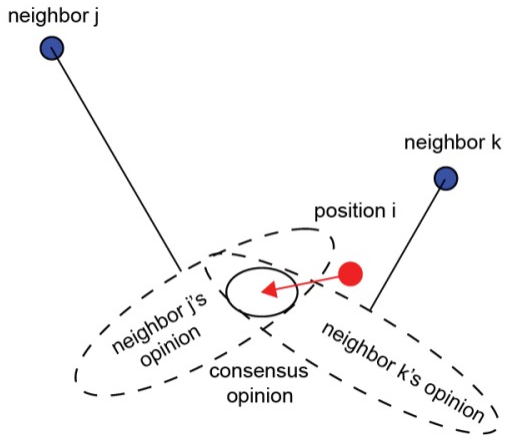
Generate's Diffusion Model for Protein Generation

Generate's Diffusion Model for Protein Generation

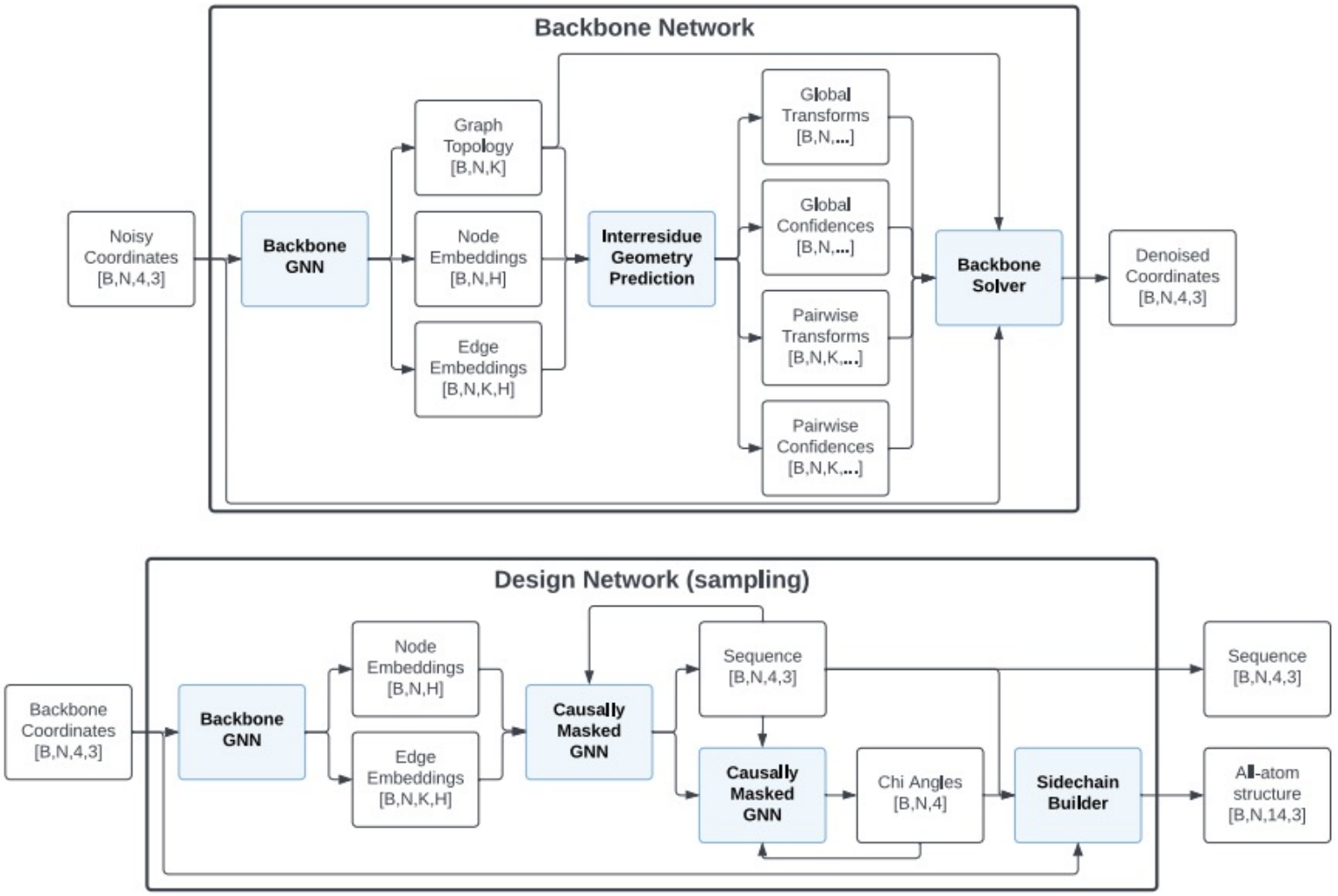


Generate's Diffusion Model for Protein Generation

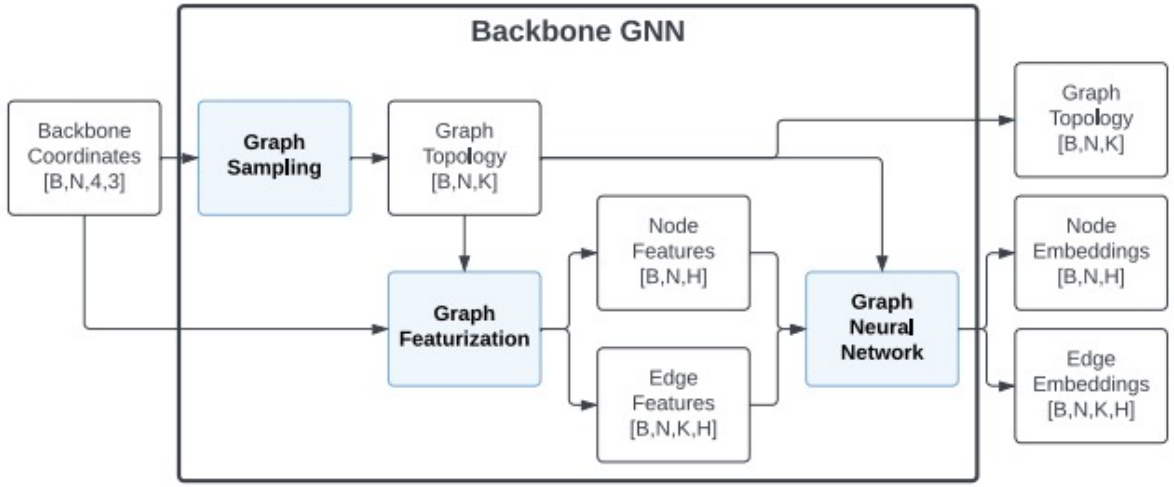


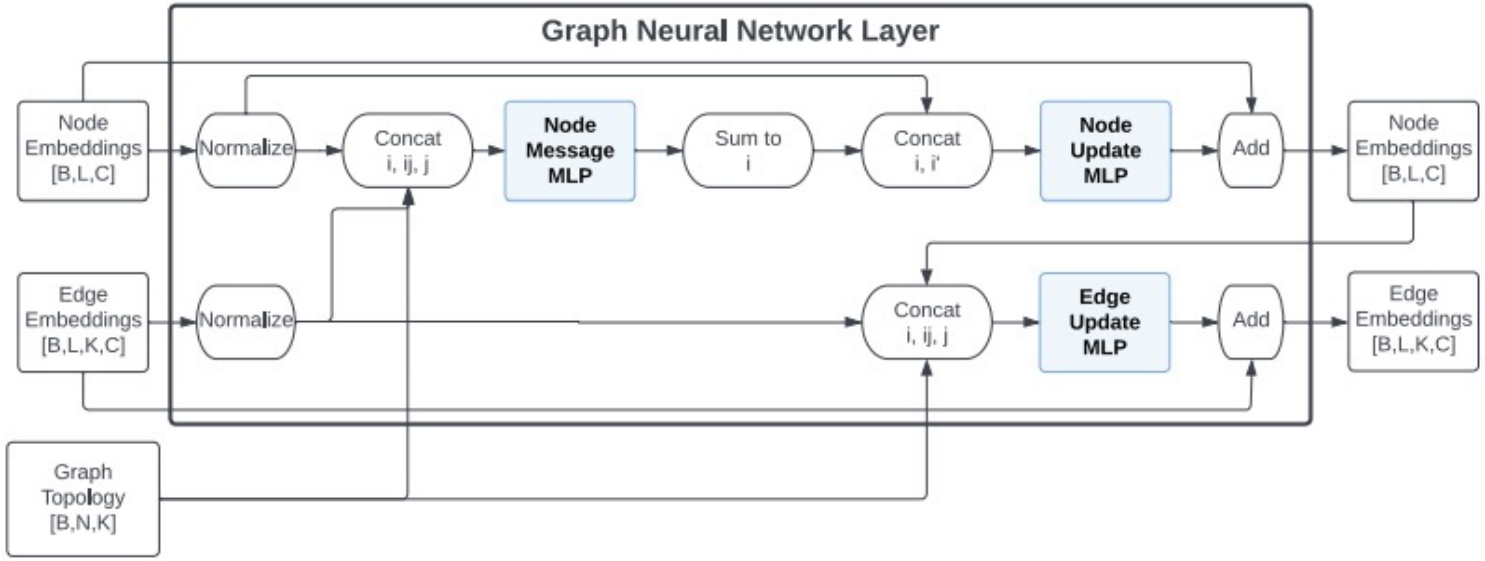


Generate's Diffusion Model for Protein Generation



Generate's Diffusion Model for Protein Generation

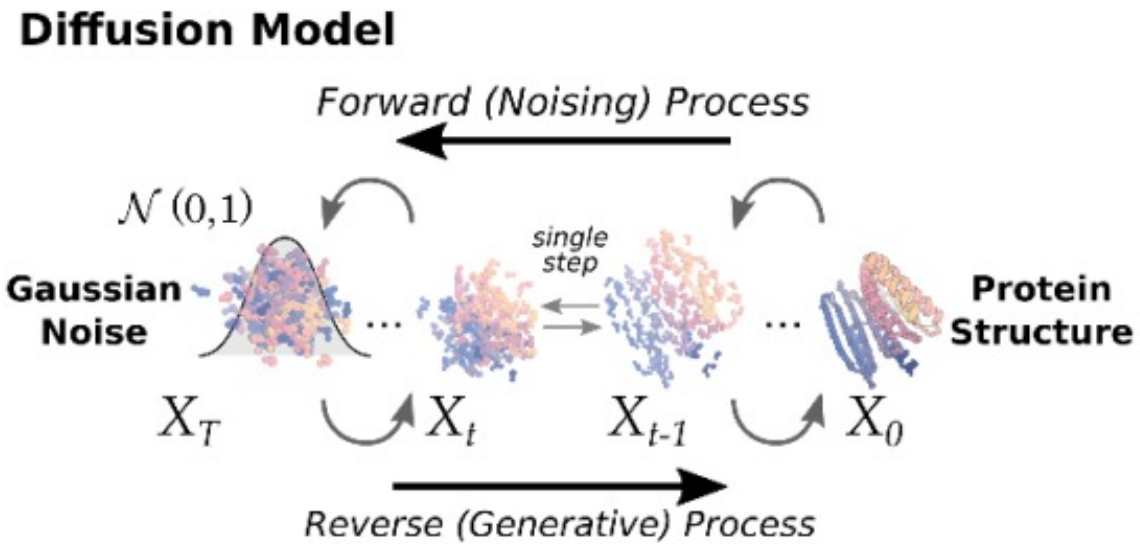




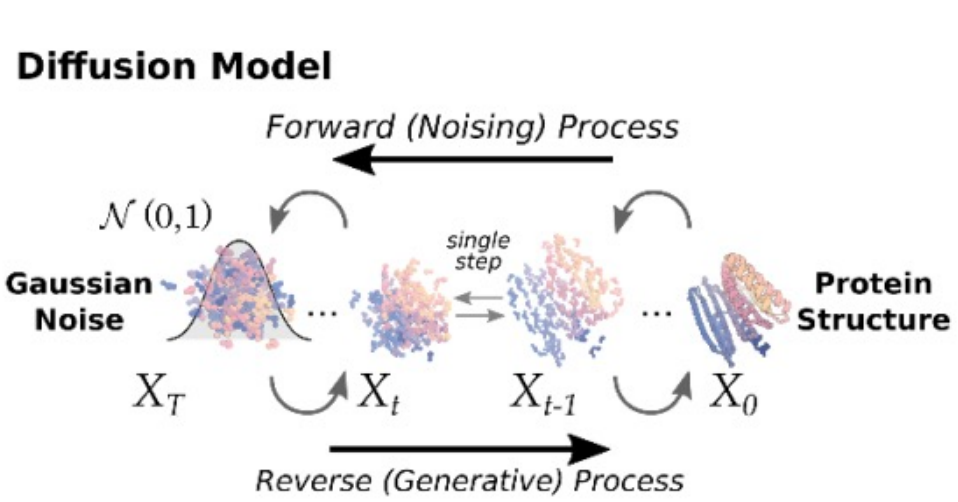
Baker Lab's Diffusion Model for Protein Generation

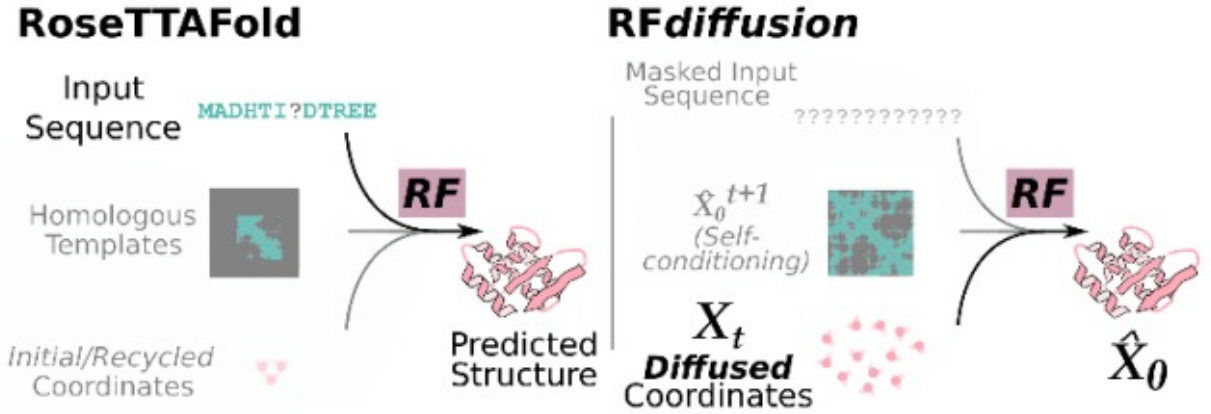
Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models

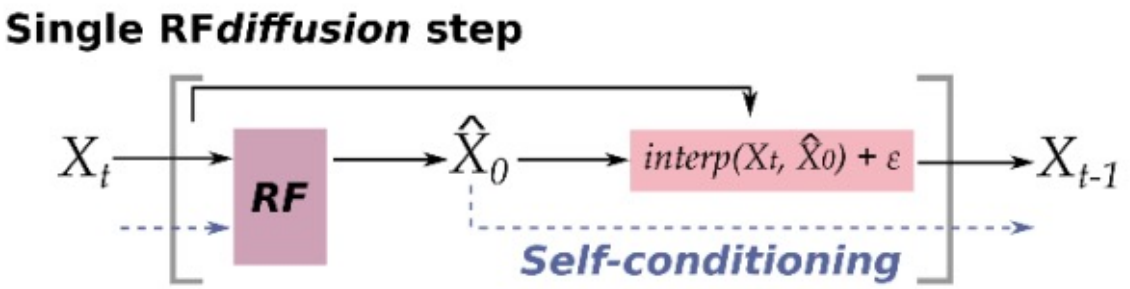
Joseph L. Watson[‡1,2], David Juergens[‡1,2,3], Nathaniel R. Bennett[‡1,2,3], Brian L. Trippe[‡2,4], Jason Yim[‡2,6], Helen E. Eisenach[‡1,2], Woody Ahern[‡1,2,7], Andrew J. Borst[1,2], Robert J. Ragotte[1,2], Lukas F. Milles[1,2], Basile I. M. Wicky[1,2], Nikita Hanikel[1,2], Samuel J. Pellock[1,2],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



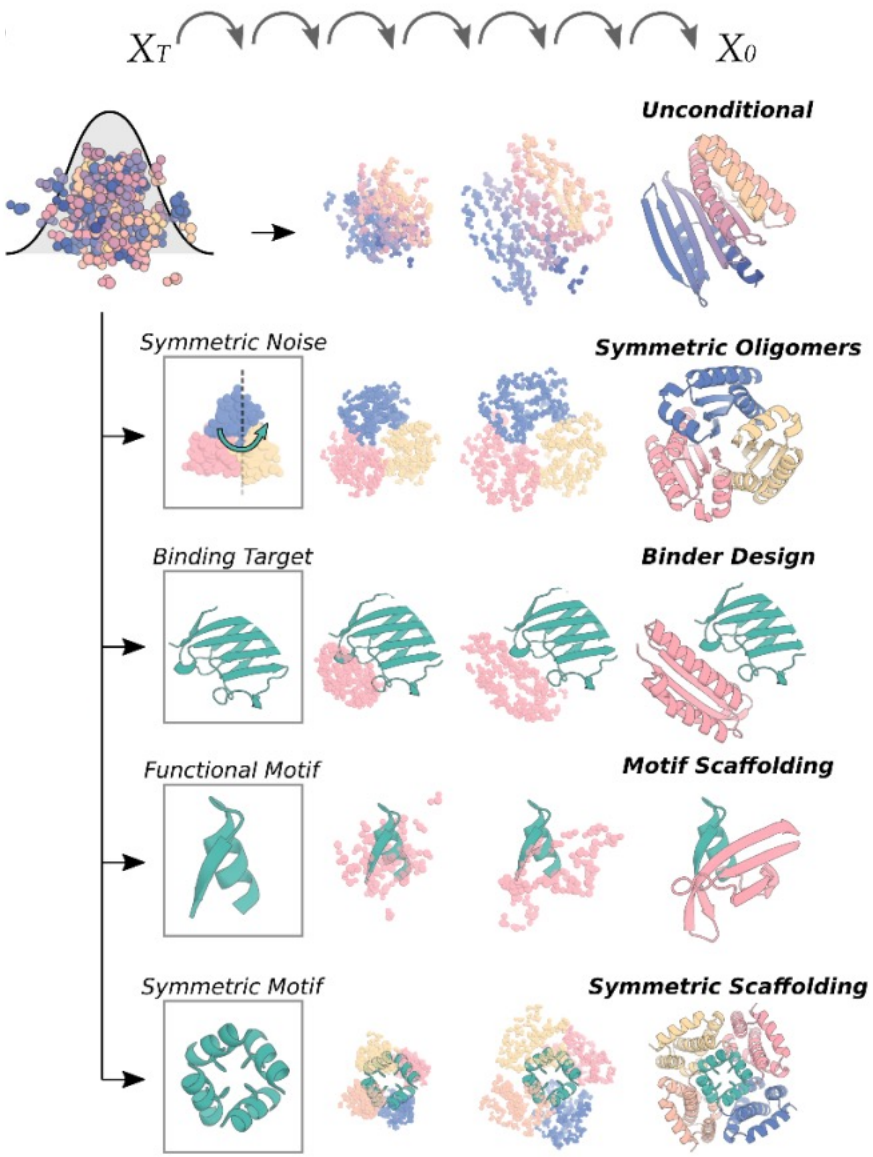
Baker Lab's Diffusion Model for Protein Generation



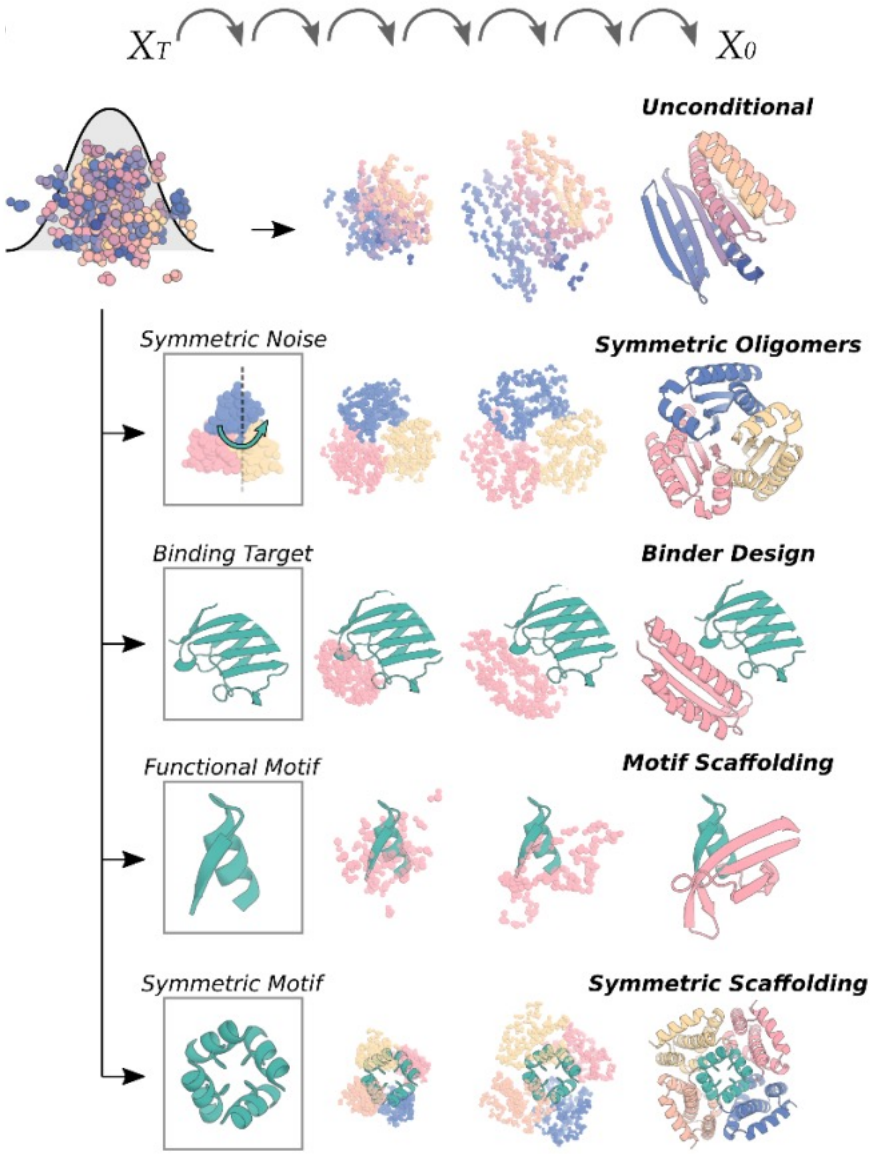


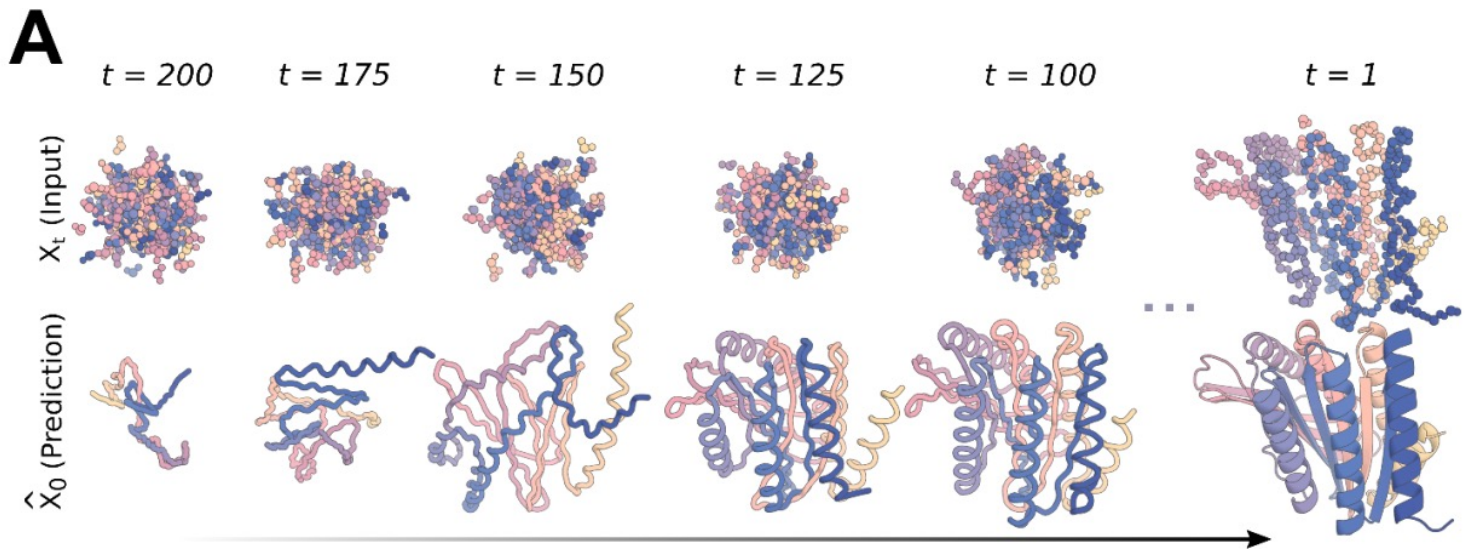


Baker Lab's Diffusion Model for Protein Generation



Baker Lab's Diffusion Model for Protein Generation





What design problems can be explored with diffusion?

There are 4 particularly interesting design cases:

- Unconditional generation (design a backbone with no constraint)
- Binder generation (design a protein that binds target)
- Scaffold generation (design a protein that scaffolds a motif)
- Partial generation (re-generate structure from given structure)

Monomeric, homo-oligomeric, hetero-oligomeric specifications*

RFDiffusion Demonstration

RFDiffusion Demonstration

search *“rfdiffusion github”* or go to

https://github.com/RosettaCommons/RFdiffusion

Slide Credit

Deniz Akpinaroglu