BIOCHEMISTRY & MOLECULAR BIOLOGY

Seminar Series

Scalable Emulation of Protein Equilibrium Ensembles *with* Generative Deep Learning

Andrew Y. K. Foong, Ph.D.

October 14th, 2025

MAYO CLINIC

Radiation Oncology
AI & Data Analytics
AIDA

About me

Senior Associate Consultant · Radiation Oncology

Senior Researcher · Microsoft Research

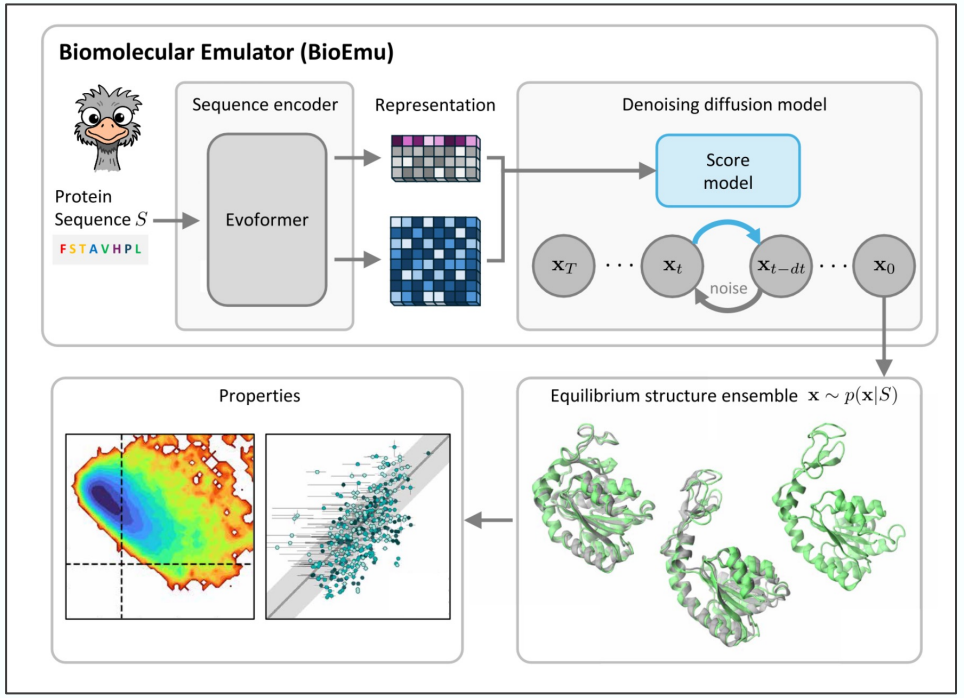
Ph.D. in Machine Learning · Cambridge University

Research Scientist Intern · Google DeepMind

More info: https://andrewfoongyk.github.io

BioEmu—a Biomolecular Emulator

- BioEmu is a deep learning model that takes protein sequences as input, and outputs a *distribution* over 3D protein structures.
- The distributions are trained to emulate physical conformational variability:
  - Predicts *multiple* conformations
  - Predicts *thermodynamics*: the relative probabilities of each conformation.



BioEmu—a Biomolecular Emulator

PROTEIN SIMULATIONS

Scalable emulation of protein equilibrium ensembles with generative deep learning

Sarah Lewis[1]†, Tim Hempel[1]†, José Jiménez-Luna[1]†, Michael Gastegger[1]†, Yu Xie[1]†, Andrew Y. K. Foong[1]†, Victor García Satorras[1]†, Osama Abdin[1]†, Bastiaan S. Veeling[1]†, Iryna Zaporozhets[1,2], Yaoyi Chen[1,2], Soojung Yang[1], Adam E. Foster[1], Arne Schneuing[1], Jigyasa Nigam[1], Federico Barbero[1], Vincent Stimper[1], Andrew Campbell[1], Jason Yim[1], Marten Lienen[1], Yu Shi[1], Shuxin Zheng[1], Hannes Schulz[1], Usman Munir[1], Roberto Sordillo[1], Ryota Tomioka[1], Cecilia Clementi[1,2,3], Frank Noé[1,2,3]*

Following the sequence and structure revolutions, predicting functionally relevant protein structure changes at scale remains an outstanding challenge. We introduce BioEmu, a deep learning system that emulates protein equilibrium ensembles by generating thousands of statistically independent structures per hour on a single graphics processing unit (GPU). BioEmu integrates more than 200 milliseconds of molecular dynamics (MD) simulations, static structures, and experimental protein stabilities using new training algorithms. It captures diverse functional motions—including cryptic pocket formation, local unfolding, and domain rearrangements—and predicts relative free energies with 1 kilocalorie per mole accuracy compared with millisecond-scale MD and experimental data. BioEmu provides mechanistic insights by jointly modeling structural ensembles and thermodynamic properties. This approach amortizes the cost of MD and experimental data generation, demonstrating a scalable path toward understanding and designing protein function.

Sarah Lewis, Tim Hempel, José Jiménez-Luna, Michael Gastegger, Yu Xie, Andrew Y. K. Foong, Victor García Satorras, Osama Abdin, Bastiaan S. Veeling *et al.*, Scalable emulation of protein equilibrium ensembles with generative deep learning. *Science* 389, eadv9817 (2025).
DOI:10.1126/science.adv9817

Today's talk

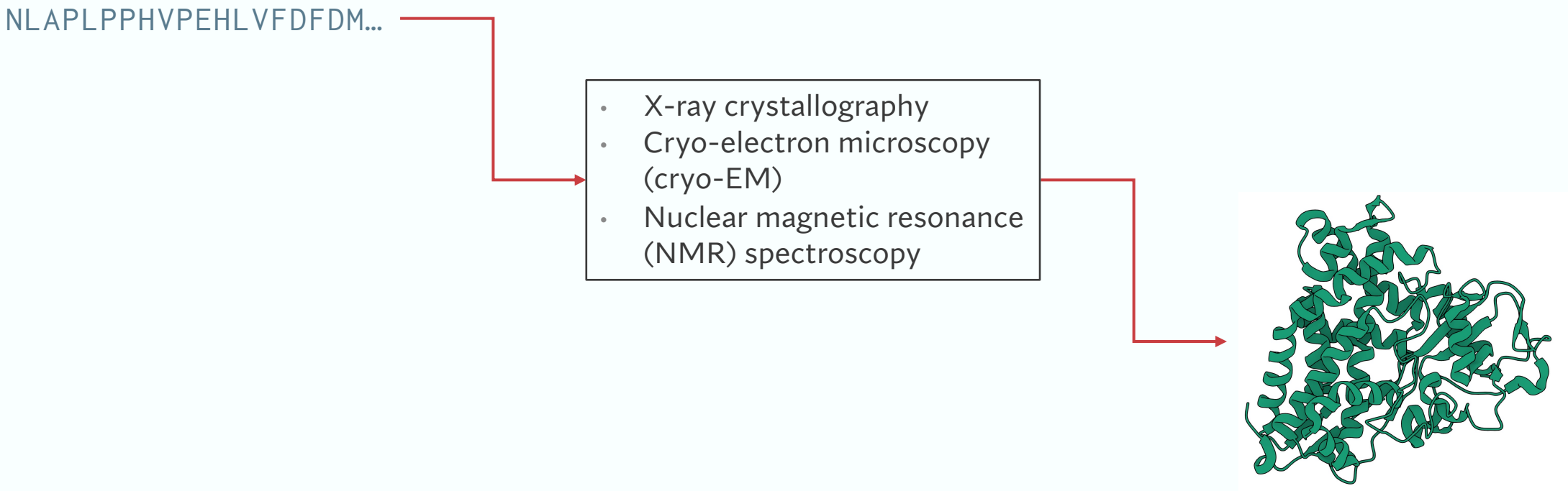
1. Background
   - *AlphaFold and the deep learning revolution*
2. AlphaFold
   - *How does it work?*
3. Structure is not enough
   - *Equilibrium distributions*
4. BioEmu
   - *Training data*
   - *Diffusion model*
   - *Capabilities*
5. Q&A

Background

AlphaFold and the Deep Learning Revolution

The structure prediction problem

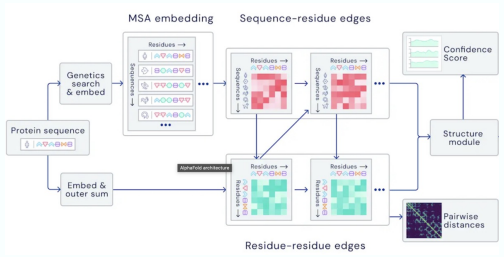
- Protein *function* is mediated through structure.
- Protein *structure* is determined by its amino acid sequence.
- PROBLEM: Given an amino acid sequence, *what is the three-dimensional structure that the protein folds into?*



Deep learning for protein structure

- Google DeepMind cast protein structure prediction as *supervised learning*.
- Training data obtained from the Protein DataBank (PDB).
- KEY BREAKTHROUGH: *deep learning*.

NLAPLPPHVPEHLVFDFDM…

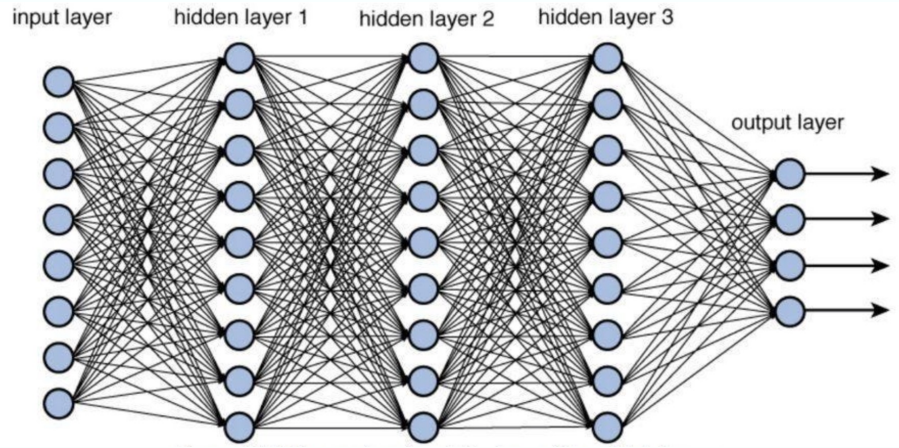


AlphaFold2
Deep neural network

Jumper, J., Evans, R., Pritzel, A. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021). https://doi.org/10.1038/s41586-021-03819-2

Deep learning in a nutshell

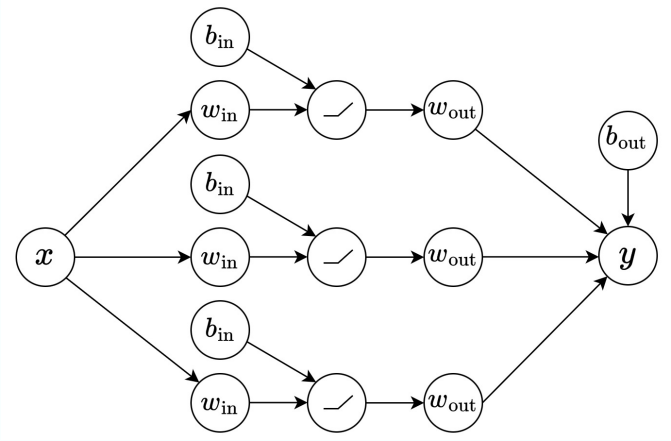
- Deep learning = use of *artificial neural networks*.
- Mathematical functions with millions of numbers: “parameters/weights”.
- The numbers determine how the neural network behaves.
  1. Start by choosing parameters randomly (garbage predictions).
  2. Optimizer automatically adjusts parameters to fit example data.
  3. Apply the function to new data (great predictions, hopefully).



Deep learning in a nutshell

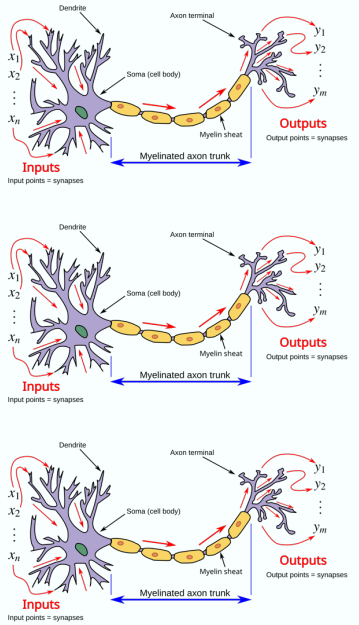
- Neural networks composed of hundreds of thousands of *artificial neurons.*
- Each neuron performs linear regression.
- Layers of these neurons are stacked interleaved with non-linear *activation functions.*

TOY EXAMPLE NETWORK (*vastly simplified*)



$\approx$

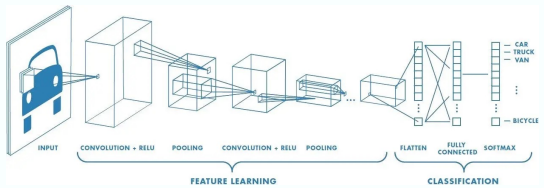
STIMULUS



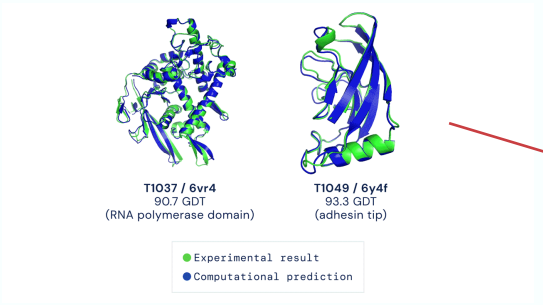
RESPONSE

Deep learning timeline

AlexNet:
Image classification



AlphaFold:
Protein structure prediction



ChatGPT:
AI chatbot

2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023 2024 2025

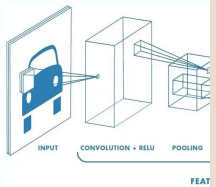
AlphaGo:
Playing board games

DALL-E:
Text-to-image generation

Deep learning timeline

AlphaFold:
Protein structure

AlexNet:
Image c



2012

23
2024
2025

Same deep learning methodology underlies all of these applications

Nobel prize for AlphaFold

DEMIS HASSABIS
Google DeepMind

JOHN JUMPER
Google DeepMind

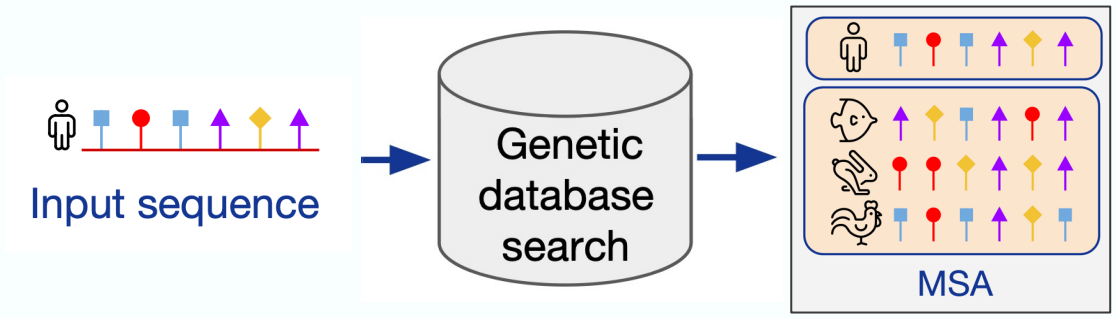
2024 NOBEL PRIZE IN CHEMISTRY

AlphaFold

How does it work?

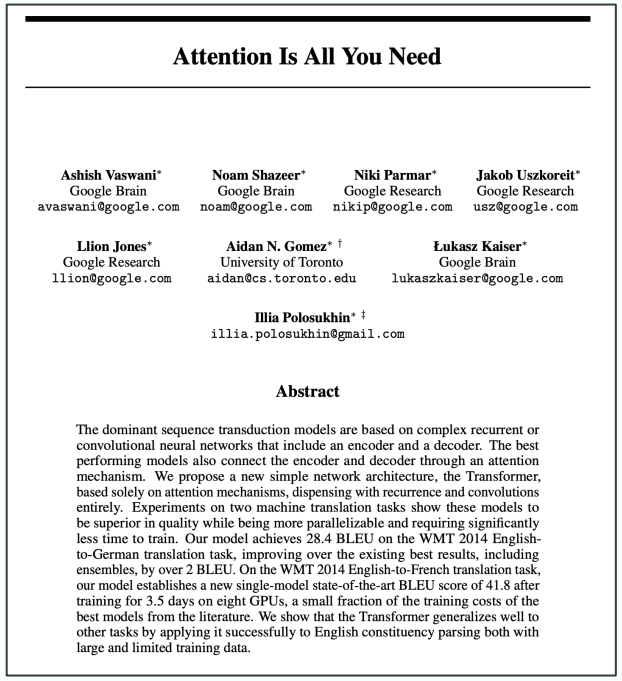
Multiple sequence alignment

- MSAs are a way of associating amino-acid sequences of related proteins in different species.
- Provides information about *structure*:
  - Maintaining consistent structure is critical to protein function.
  - Amino acids critical to that structure are *conserved* across evolution.
  - Those that are not *vary* across evolution.
- AlphaFold exploits this: the input to the network is an MSA.



Evoformer

- The MSAs are processed by the *Evoformer* neural network module.
- A variant of the *transformer* neural network architecture (e.g., ChatGPT).



Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

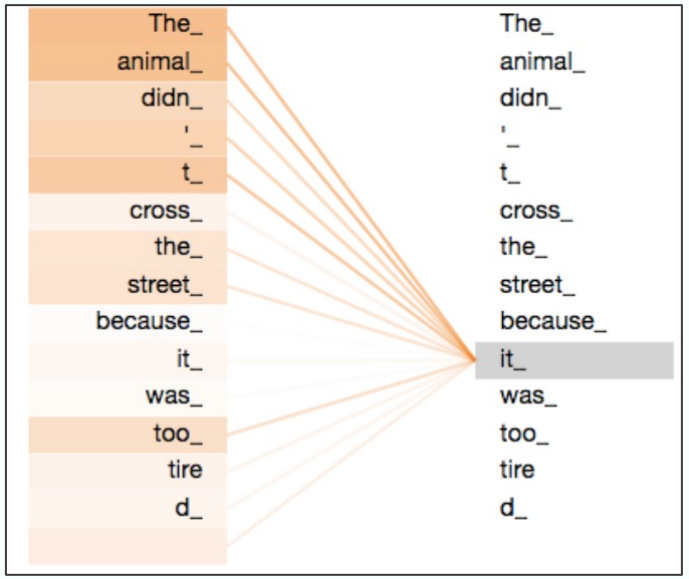
Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

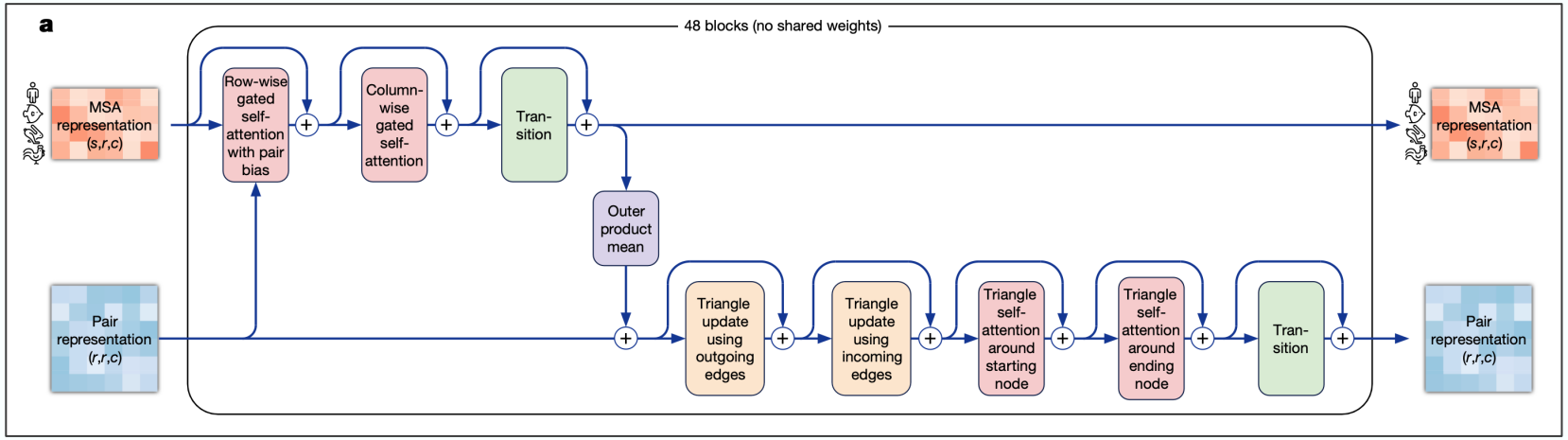
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 6000–6010.



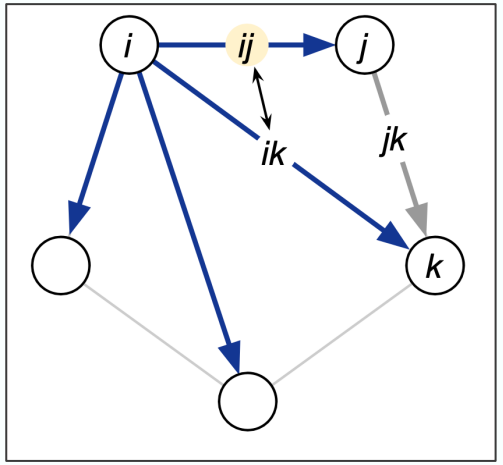
Self-attention in language generation

Evoformer

- Instead of attending to words, the network attends to *amino acids* in the MSA.
- The most important ingredient to AlphaFold's success.
- Specially-designed attention variants:
  - *Triangle self-attention*—moving from pairwise communication ('A talks to B') to group dynamics ('A and B both talk to C, so what does that mean about A–C?')
  - *Row-wise attention*—captures *within-sequence* context.
  - *Column-wise attention*—captures *across-sequence* context.



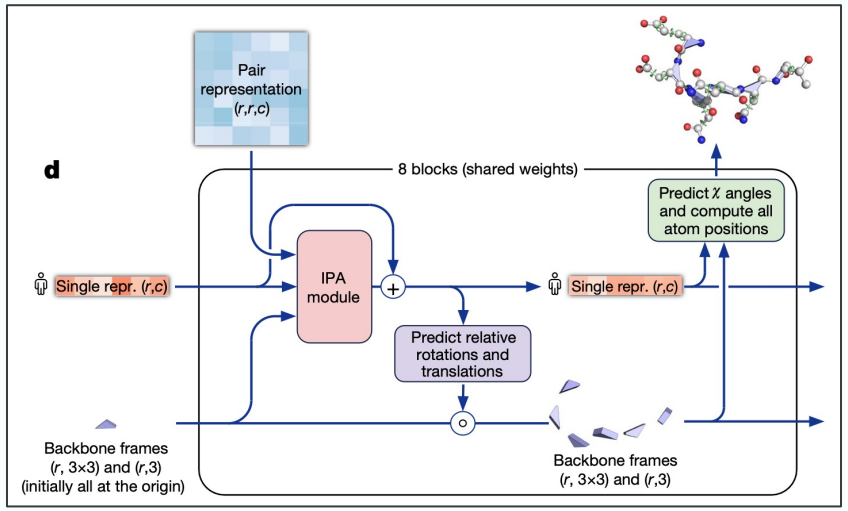
Evoformer architecture



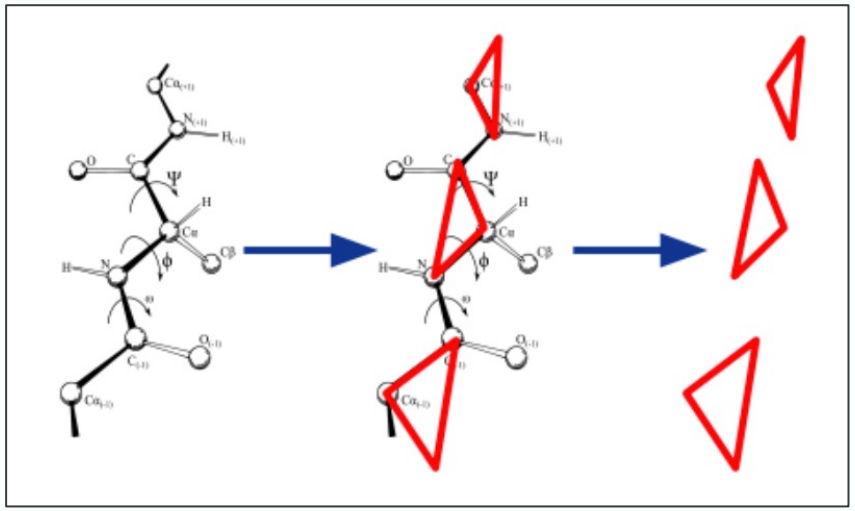
Triangle self-attention

Structure module

- The structure module takes the processed embeddings from the Evoformer and translates them into *3D geometry*.
- Each amino acid is represented by a *backbone-frame*:
  - Position in 3D space: (x, y, z).
  - Orientation: (roll, pitch, yaw).



AlphaFold structure module



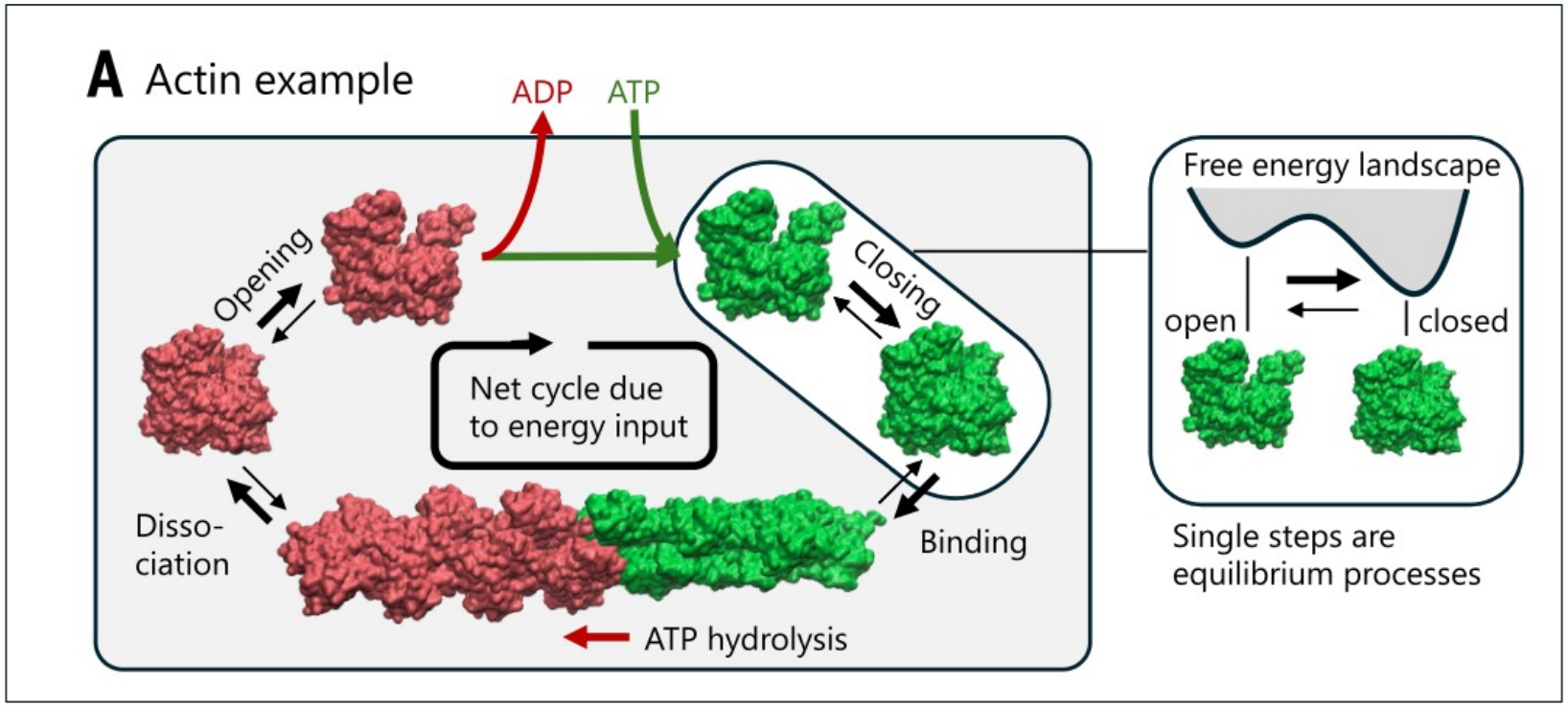
Backbone frame representation of amino acids

Structure is not enough

Equilibrium distributions

Protein equilibrium distributions

- Knowing the folded state of a protein is not everything.
- Protein *function* is often mediated by *transitions between multiple conformations.*



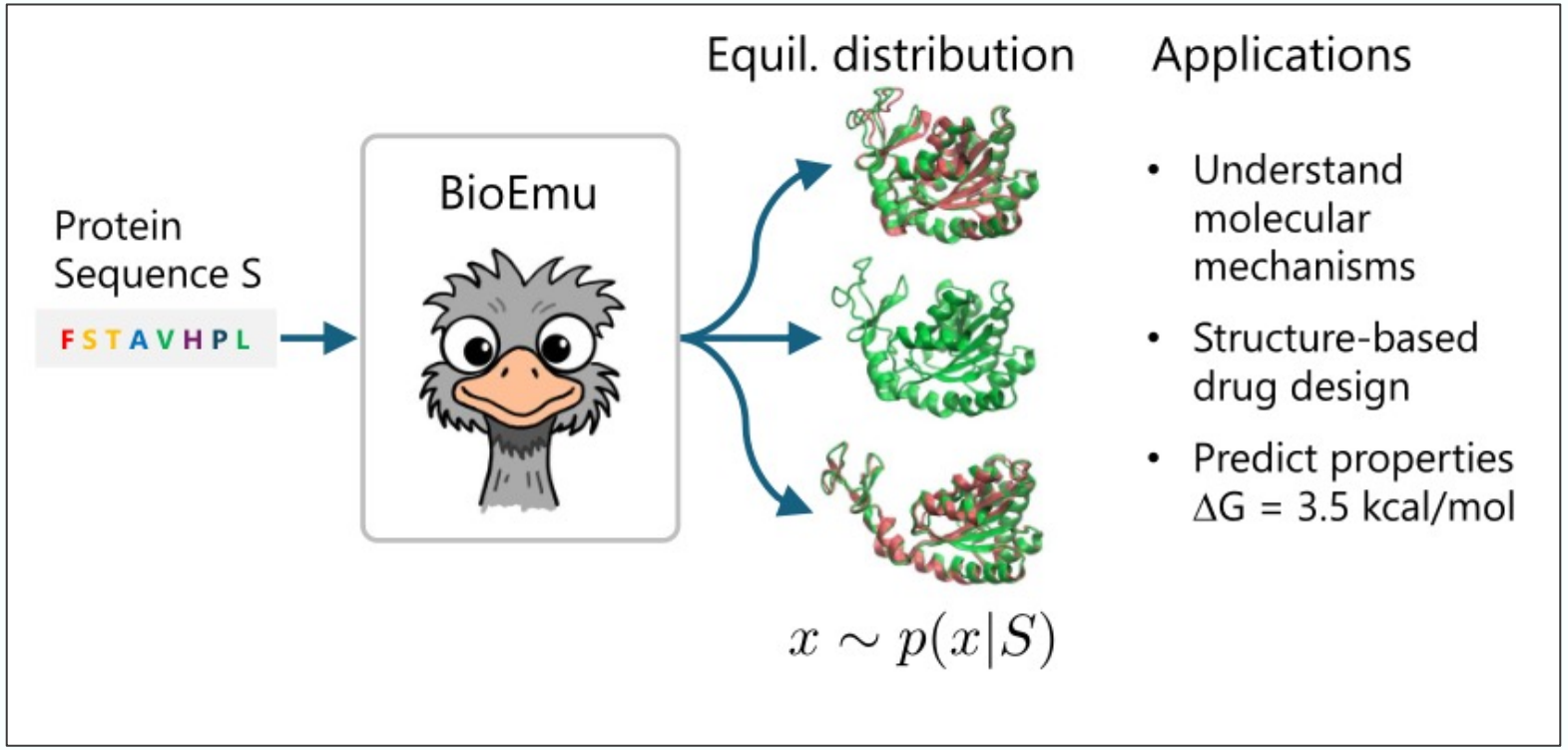
Conformational change drives function

Protein equilibrium distributions

- BOLTZMANN DISTRIBUTION: $p(x) = \exp(-U(x)/k_{\mathrm{B}}T) / Z$.
  - x is the 3D conformation of the protein.
  - $p(x)$ is its probability density.
  - $U(x)$ is the potential energy of a conformation.
  - k_{B} is Boltzmann's constant.
  - T is the temperature.
  - Z is a normalizing constant.
- *If we can sample from the Boltzmann distribution, we can describe every possible state of a protein and its relative probability.*

Going beyond AlphaFold

- BioEmu goal: approximate sampling from Boltzmann distribution *without* expensive molecular dynamics simulations.



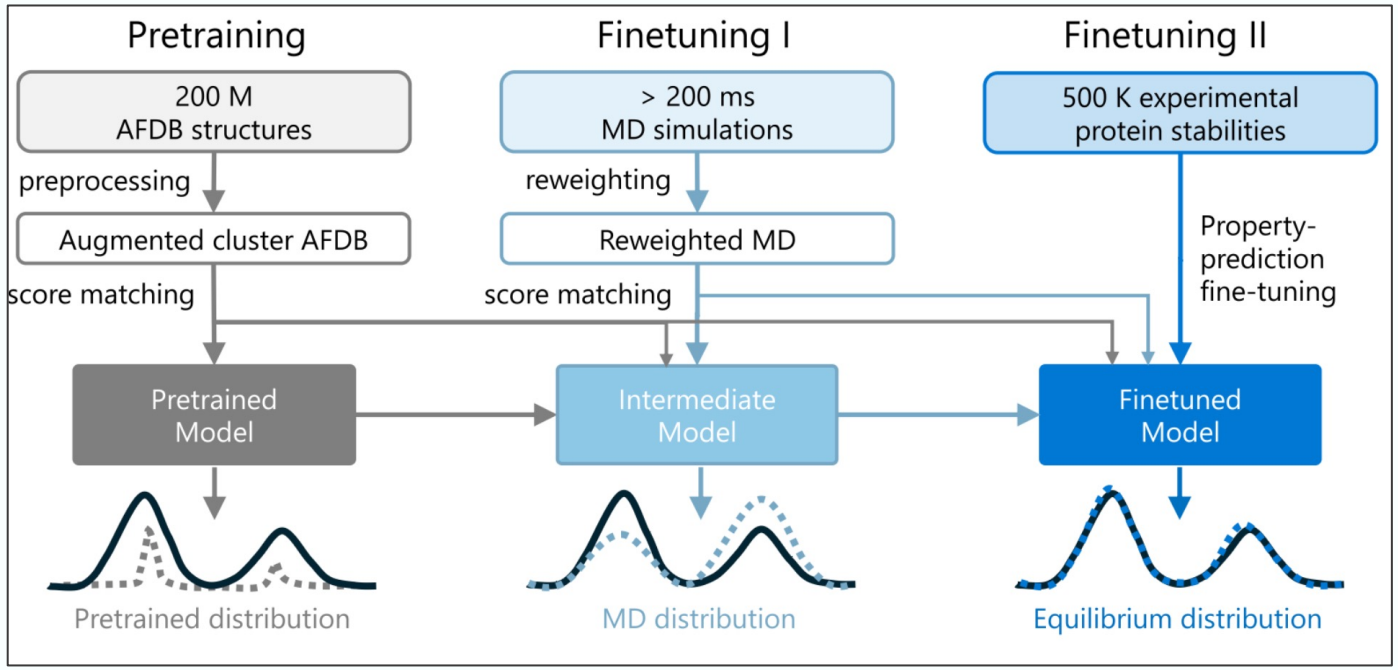
BioEmu in a nutshell

BioEmu

Training data

Sources of data

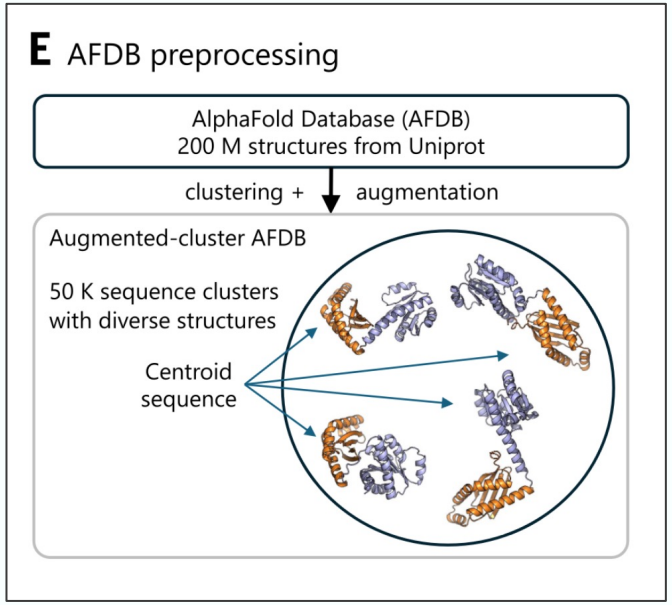
- First challenge: *data for equilibrium distributions is hard to come by.*
- There is no "PDB" for distributional data.
- Our approach:



Three-stage approach for BioEmu training data

Sources of data

- Pretraining on diversified AlphaFold *predictions*.
  - Download 200 million protein structure predictions from AlphaFold Database.
  - Cluster by sequence similarity.
  - Group *diverse structures* with *similar sequences* together.
- CONSEQUENCE: BioEmu is forced to learn *conformational variability*.



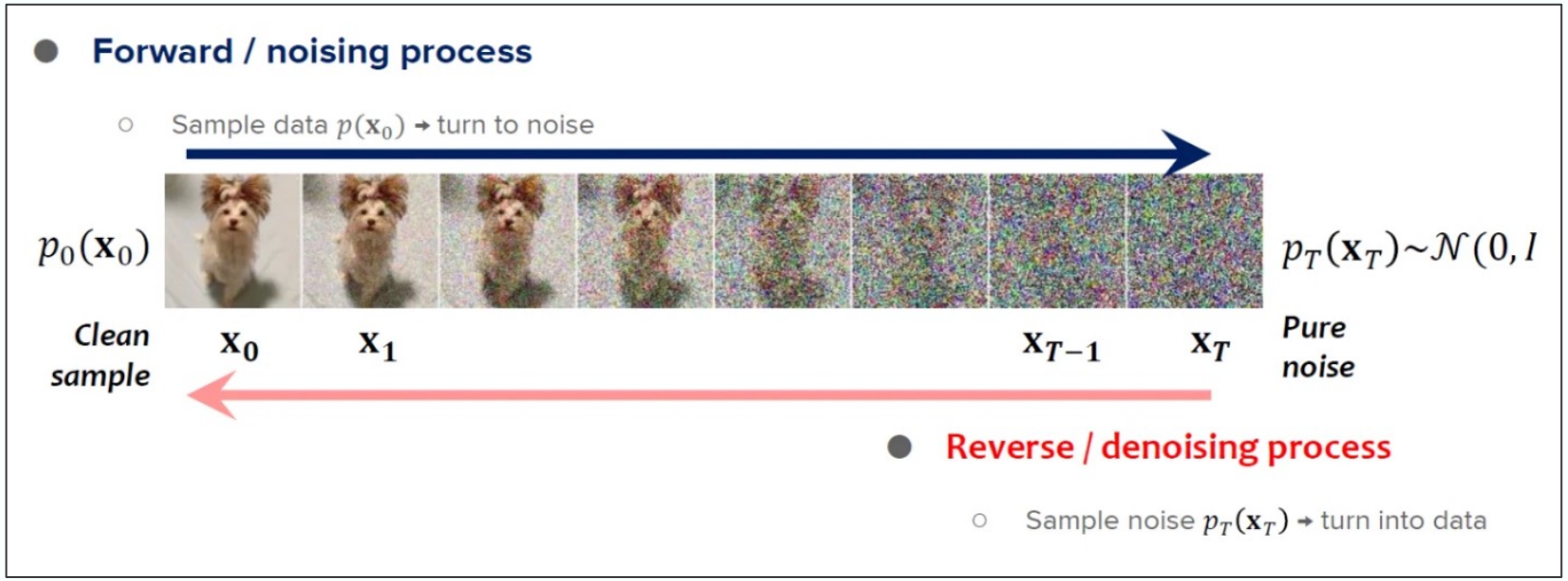
Pretraining with diversified AFDB structure clusters

BioEmu

Diffusion model

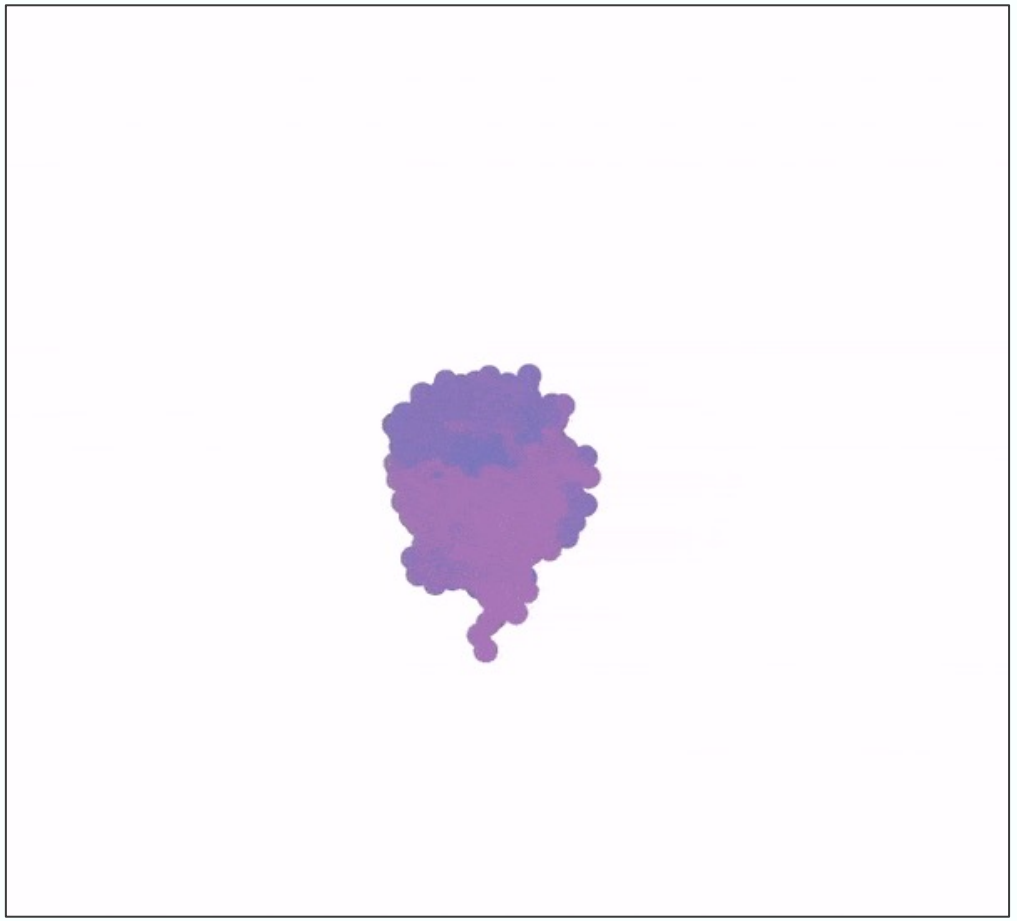
Diffusion models

- AlphaFold has *one output* structure for *one input sequence*.
- DIFFUSION MODELS output *distributions over data*:
  - Random noise input is converted to structures.
  - Variability in the noise translates to variability in the structures.
- Ideal candidate for a Boltzmann distribution emulator.



Diffusion models learn how to map random noise to data

Diffusion models

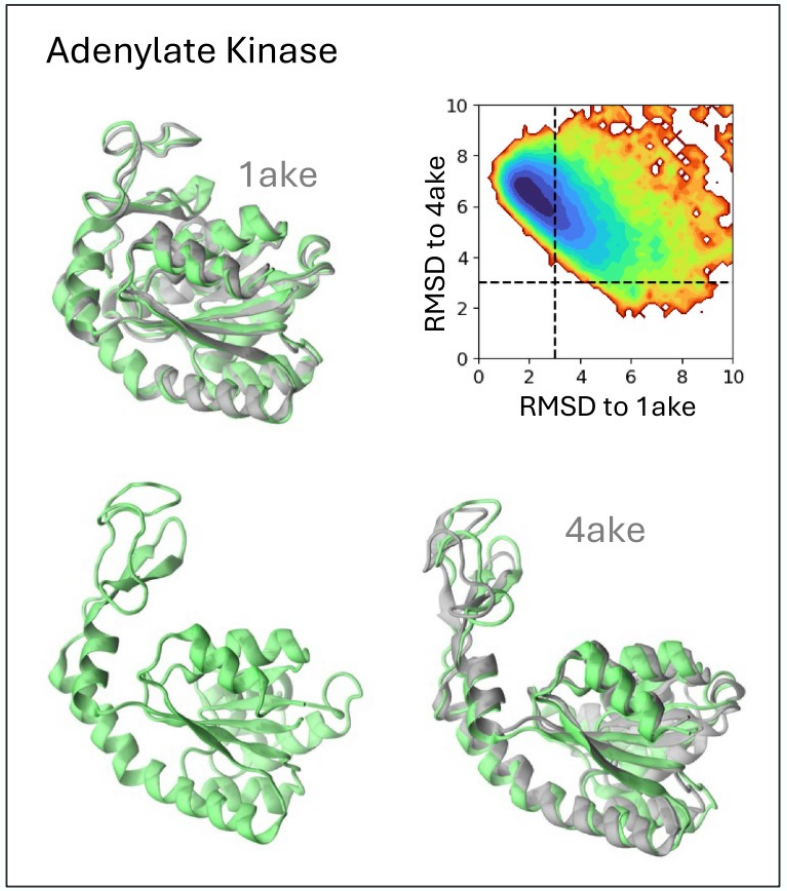


EXAMPLE: diffusion model generating protein structure

BioEmu

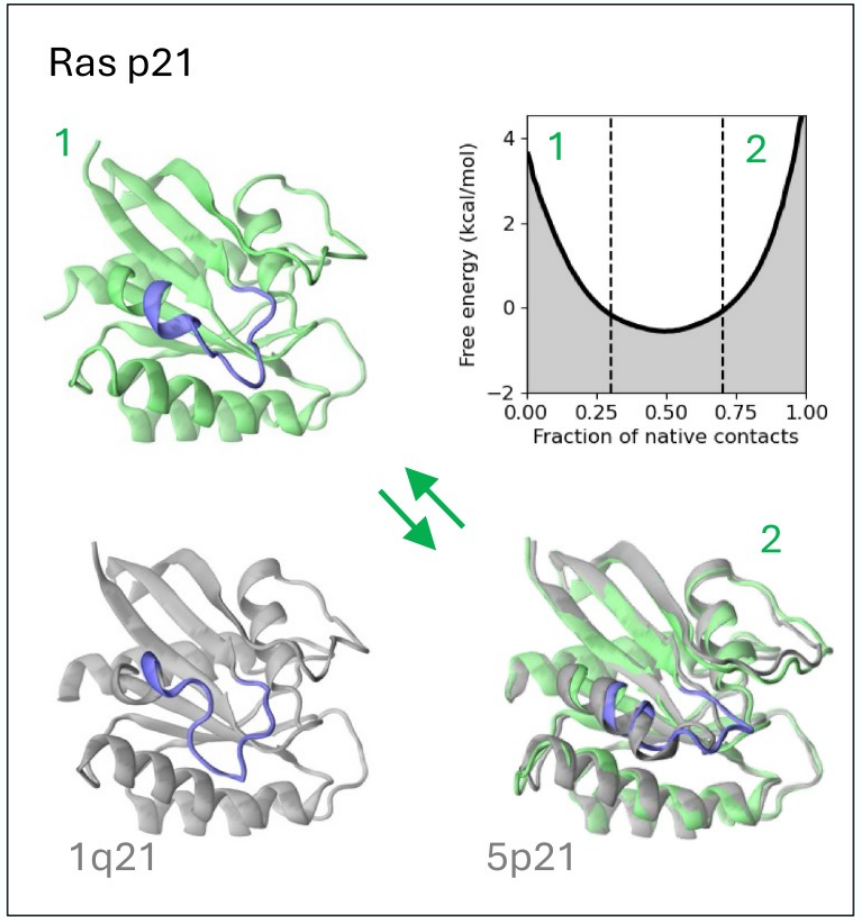
Capabilities on unseen proteins

Domain motions



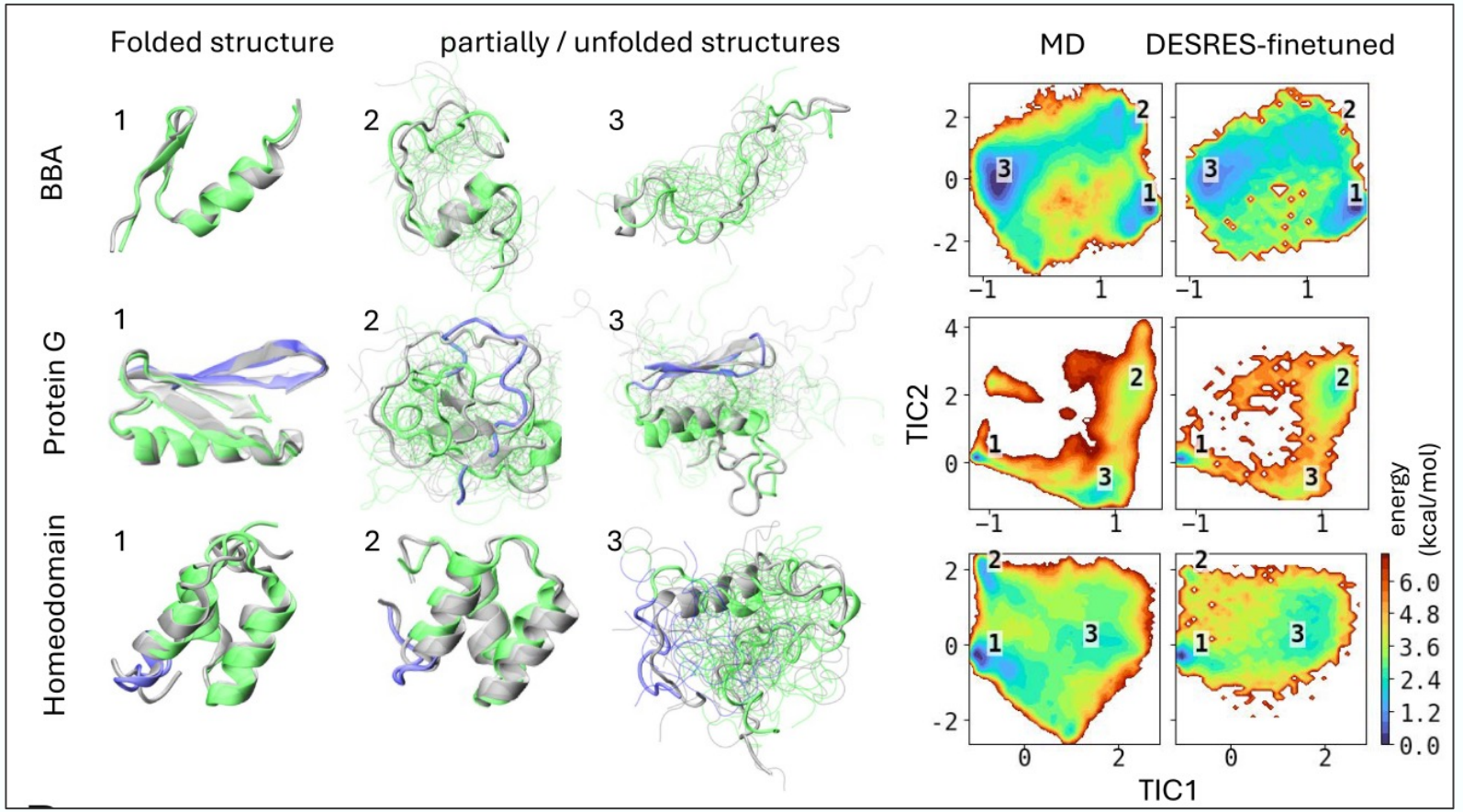
BioEmu predicts large-scale domain motions

Local unfolding



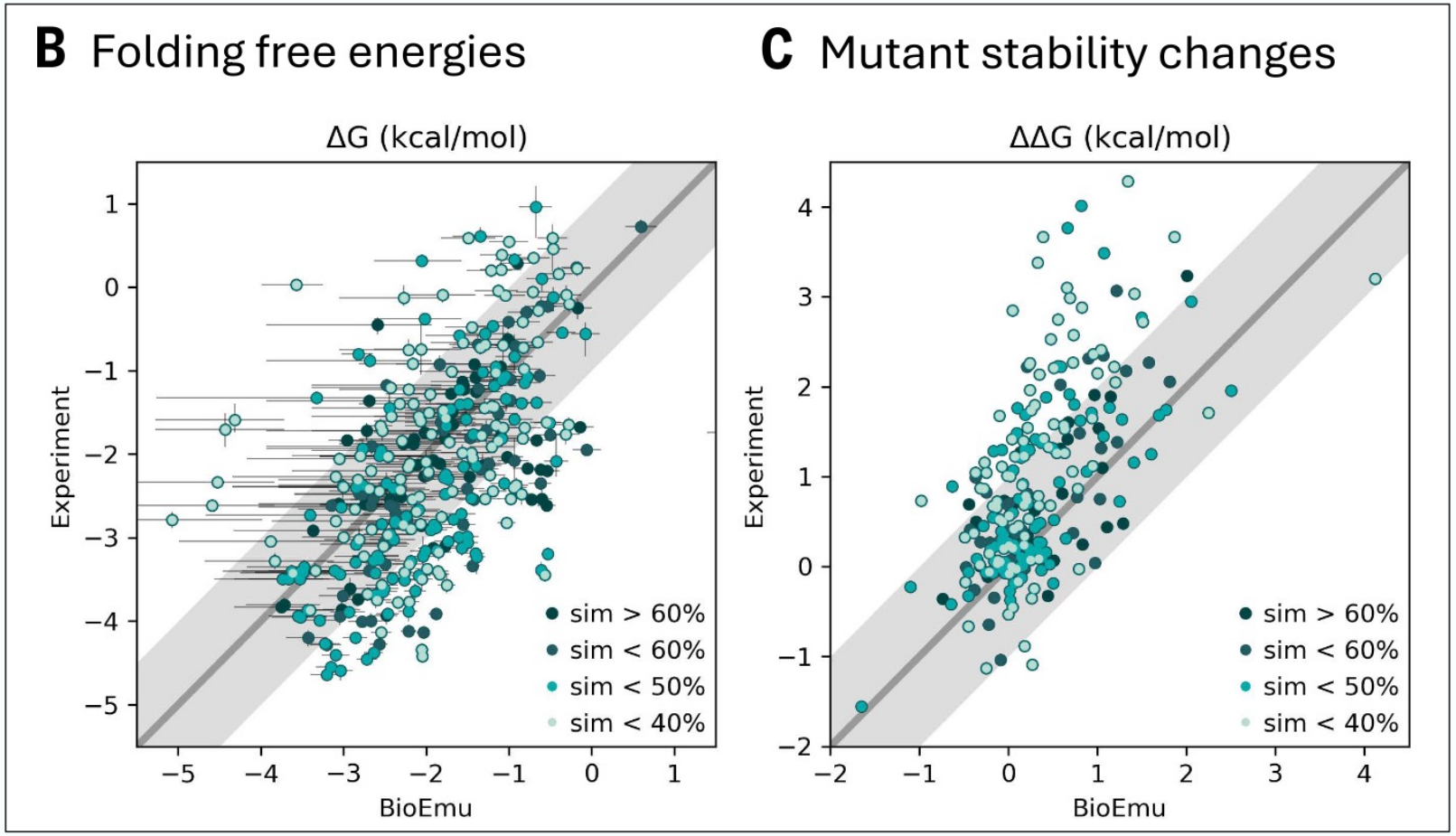
BioEmu predicts local unfolding transitions (*alpha helix destabilizes*)

Comparison with molecular dynamics



Sampling from BioEmu yields similar distributions of structures as lengthy MD simulations

Free-energy predictions

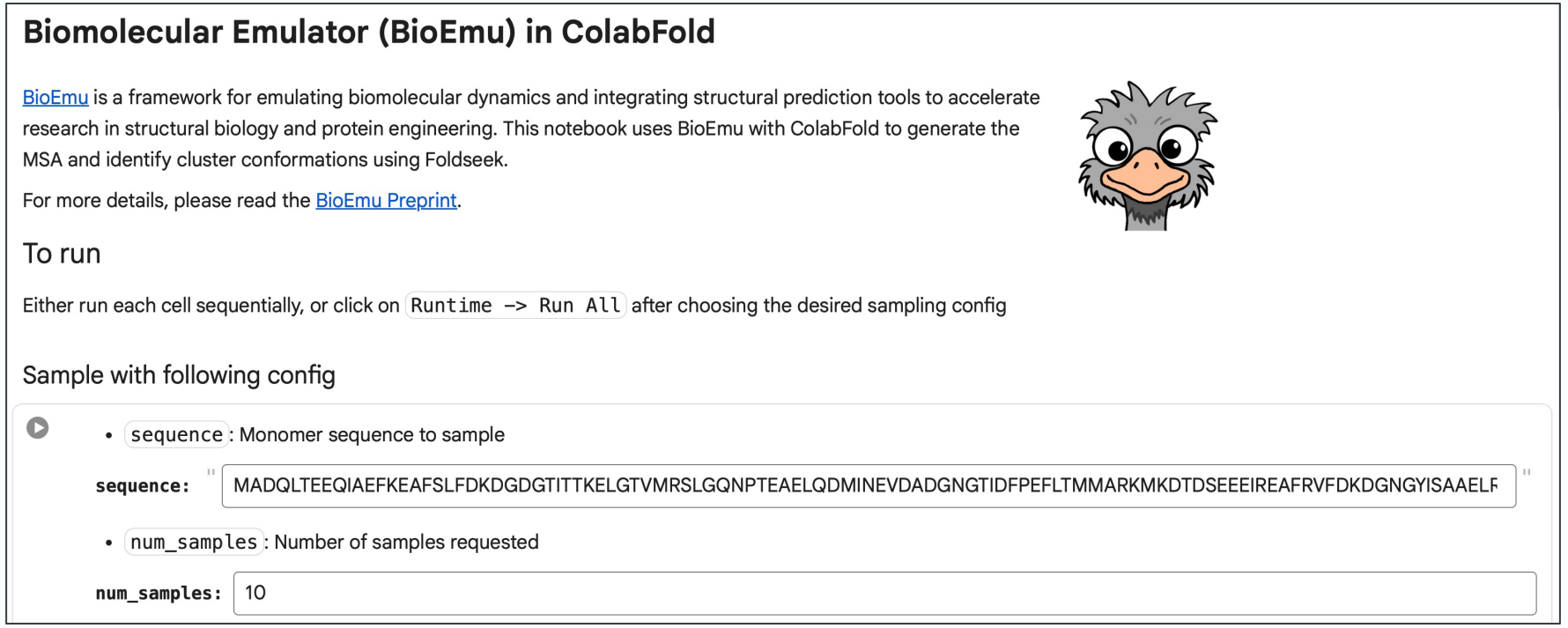


BioEmu samples reflect relative stability of protein mutants

Using BioEmu on ColabFold

Try BioEmu for free today:

- Runs in browser; no software installation required.
- Just input a sequence and download the sampled structures.



https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/BioEmu.ipynb

Thanks for listening!

Questions?