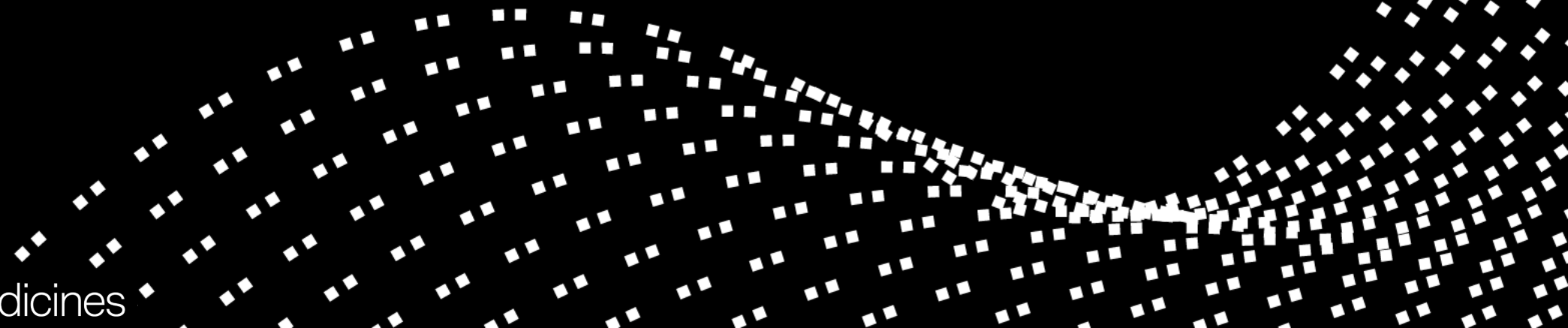
Generative Modeling for Science

(and especially biomolecules)

John Ingraham
Berkeley CS294, Deep Unsupervised Learning
April 18, 2024



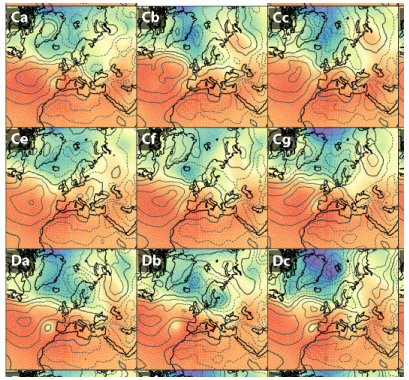
Generate : Biomedicines

Uncertainty propagation in complex physical systems

Generative emulation of weather forecast ensembles with diffusion models

Lizao Li, Robert Carver†, Ignacio Lopez-Gomez†, Fei Sha*, John Anderson

Uncertainty quantification is crucial to decision-making. A prominent example is probabilistic forecasting in numerical weather prediction. The dominant approach to representing uncertainty in weather forecasting is to generate an ensemble of forecasts by running physics-based simulations under different conditions, which is a computationally costly process. We propose to amortize the computational cost by emulating these forecasts with deep generative diffusion models learned from historical data. The learned models are highly scalable with respect to high-performance computing accelerators and can sample thousands of realistic weather forecasts at low cost. When designed to emulate operational ensemble forecasts, the generated ones are similar to physics-based ensembles in statistical properties and predictive skill. When designed to correct biases present in the operational forecasting system, the generated ensembles show improved probabilistic forecast metrics. They are more reliable and forecast probabilities of extreme weather events more accurately. While we focus on weather forecasting, this methodology may enable creating large climate projection ensembles for climate risk assessment.

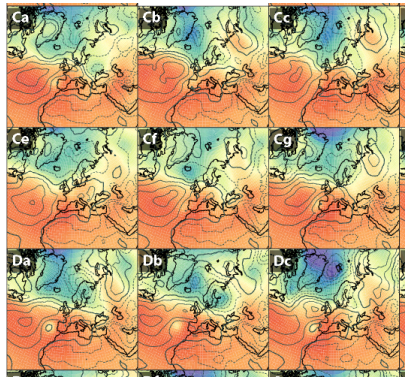


Uncertainty propagation in complex physical systems

Generative emulation of weather forecast ensembles with diffusion models

Lizao Li, Robert Carver†, Ignacio Lopez-Gomez†, Fei Sha*, John Anderson

Uncertainty quantification is crucial to decision-making. A prominent example is probabilistic forecasting in numerical weather prediction. The dominant approach to representing uncertainty in weather forecasting is to generate an ensemble of forecasts by running physics-based simulations under different conditions, which is a computationally costly process. We propose to amortize the computational cost by emulating these forecasts with deep generative diffusion models learned from historical data. The learned models are highly scalable with respect to high-performance computing accelerators and can sample thousands of realistic weather forecasts at low cost. When designed to emulate operational ensemble forecasts, the generated ones are similar to physics-based ensembles in statistical properties and predictive skill. When designed to correct biases present in the operational forecasting system, the generated ensembles show improved probabilistic forecast metrics. They are more reliable and forecast probabilities of extreme weather events more accurately. While we focus on weather forecasting, this methodology may enable creating large climate projection ensembles for climate risk assessment.

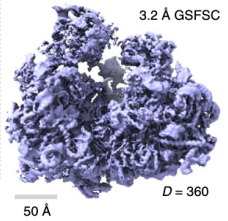


Scientific imaging & inverse problems

CryoDRGN: reconstruction of heterogeneous cryo-EM structures using neural networks

Ellen D. Zhong[1,2], Tristan Bepler[1,2], Bonnie Berger[2,3] ✉ and Joseph H. Davis[1,4] ✉

Cryo-electron microscopy (cryo-EM) single-particle analysis has proven powerful in dete
romolecules. However, many imaged protein complexes exhibit conformational and comp
major challenge to existing three-dimensional reconstruction methods. Here, we present
ages the representation power of deep neural networks to directly reconstruct continuous
map per-particle heterogeneity of single-particle cryo-EM datasets. Using cryoDRGN, w
in high-resolution datasets of the 80S ribosome and the RAG complex, revealed a new st
ribosome, and visualized large-scale continuous motions of a spliceosome complex. Cry
visualize a dataset's distribution of per-particle variability, generate density maps for e
subsets for use with other tools and generate trajectories to visualize molecular motions
freely available at http://cryodrgn.csail.mit.edu.



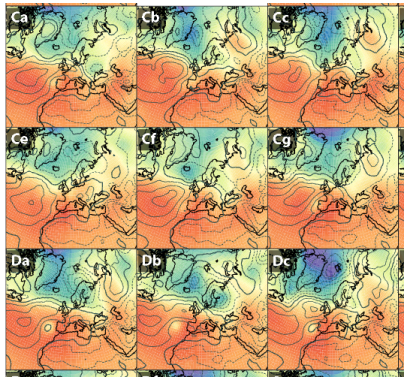
NeRF for molecules!

Uncertainty propagation in complex physical systems

Generative emulation of weather forecast ensembles with diffusion models

Lizao Li, Robert Carver†, Ignacio Lopez-Gomez†, Fei Sha*, John Anderson

Uncertainty quantification is crucial to decision-making. A prominent example is probabilistic forecasting in numerical weather prediction. The dominant approach to representing uncertainty in weather forecasting is to generate an ensemble of forecasts by running physics-based simulations under different conditions, which is a computationally costly process. We propose to amortize the computational cost by emulating these forecasts with deep generative diffusion models learned from historical data. The learned models are highly scalable with respect to high-performance computing accelerators and can sample thousands of realistic weather forecasts at low cost. When designed to emulate operational ensemble forecasts, the generated ones are similar to physics-based ensembles in statistical properties and predictive skill. When designed to correct biases present in the operational forecasting system, the generated ensembles show improved probabilistic forecast metrics. They are more reliable and forecast probabilities of extreme weather events more accurately. While we focus on weather forecasting, this methodology may enable creating large climate projection ensembles for climate risk assessment.

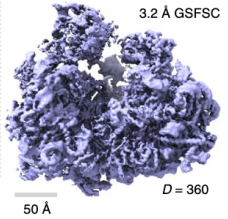


Scientific imaging & inverse problems

CryoDRGN: reconstruction of heterogeneous cryo-EM structures using neural networks

Ellen D. Zhong[1,2], Tristan Bepler[1,2], Bonnie Berger[2,3]✉ and Joseph H. Davis[1,4]✉

Cryo-electron microscopy (cryo-EM) single-particle analysis has proven powerful in dete
romolecules. However, many imaged protein complexes exhibit conformational and comp
major challenge to existing three-dimensional reconstruction methods. Here, we present
ages the representation power of deep neural networks to directly reconstruct continuous
map per-particle heterogeneity of single-particle cryo-EM datasets. Using cryoDRGN, w
in high-resolution datasets of the 80S ribosome and the RAG complex, revealed a new st
ribosome, and visualized large-scale continuous motions of a spliceosome complex. Cry
visualize a dataset's distribution of per-particle variability, generate density maps for e
subsets for use with other tools and generate trajectories to visualize molecular motions
freely available at http://cryodrgn.csail.mit.edu.



NeRF for molecules!

Inference machinery for simulations

The frontier of simulation-based inference

Kyle Cranmer[a,b,1], Johann Brehmer[a,b], and Gilles Louppe[c]

[a]Center for Cosmology and Particle Physics, New York University, New York, NY 10003; [b]Center for Data Science, New Yor
and [c]Montefiore Institute, University of Liège, B-4000 Liège, Belgium

Edited by Jitendra Malik, University of California, Berkeley, CA, and approved April 10, 2020 (received for review Novemb

Many domains of science have developed complex simulations to describe phenomena of interest. While these simulations provide high-fidelity models, they are poorly suited for inference and lead to challenging inverse problems. We review the rapidly developing field of simulation-based inference and identify the forces giving additional momentum to the field. Finally, we describe how the frontier is expanding so that a broad audience can appreciate the profound influence these developments may have on science.

the simulator—is being recognized
sample efficiency of various inferen
of research has stopped treating the
focused on integrations that allow
into the internal details of the simu

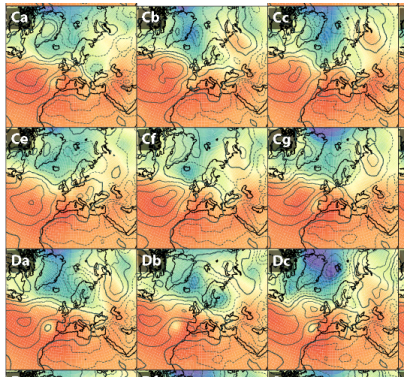
Amidst this ongoing revolution,
based inference is changing rapid
provide the reader with a high-leve

Uncertainty propagation in complex physical systems

Generative emulation of weather forecast ensembles with diffusion models

Lizao Li, Robert Carver†, Ignacio Lopez-Gomez†, Fei Sha*, John Anderson

Uncertainty quantification is crucial to decision-making. A prominent example is probabilistic forecasting in numerical weather prediction. The dominant approach to representing uncertainty in weather forecasting is to generate an ensemble of forecasts by running physics-based simulations under different conditions, which is a computationally costly process. We propose to amortize the computational cost by emulating these forecasts with deep generative diffusion models learned from historical data. The learned models are highly scalable with respect to high-performance computing accelerators and can sample thousands of realistic weather forecasts at low cost. When designed to emulate operational ensemble forecasts, the generated ones are similar to physics-based ensembles in statistical properties and predictive skill. When designed to correct biases present in the operational forecasting system, the generated ensembles show improved probabilistic forecast metrics. They are more reliable and forecast probabilities of extreme weather events more accurately. While we focus on weather forecasting, this methodology may enable creating large climate projection ensembles for climate risk assessment.

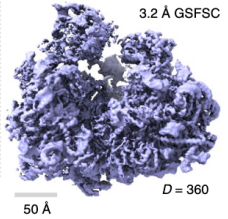


Scientific imaging & inverse problems

CryoDRGN: reconstruction of heterogeneous cryo-EM structures using neural networks

Ellen D. Zhong[1,2], Tristan Bepler[1,2], Bonnie Berger[2,3]✉ and Joseph H. Davis[1,4]✉

Cryo-electron microscopy (cryo-EM) single-particle analysis has proven powerful in dete
romolecules. However, many imaged protein complexes exhibit conformational and comp
major challenge to existing three-dimensional reconstruction methods. Here, we present
ages the representation power of deep neural networks to directly reconstruct continuous
map per-particle heterogeneity of single-particle cryo-EM datasets. Using cryoDRGN, w
in high-resolution datasets of the 80S ribosome and the RAG complex, revealed a new st
ribosome, and visualized large-scale continuous motions of a spliceosome complex. Cry
visualize a dataset's distribution of per-particle variability, generate density maps for
subsets for use with other tools and generate trajectories to visualize molecular motions
freely available at http://cryodrgn.csail.mit.edu.



NeRF for molecules!

Inference machinery for simulations

The frontier of simulation-based inference

Kyle Cranmer[a,b,1], Johann Brehmer[a,b], and Gilles Louppe[c]

[a]Center for Cosmology and Particle Physics, New York University, New York, NY 10003; [b]Center for Data Science, New Yor
and [c]Montefiore Institute, University of Liège, B-4000 Liège, Belgium

Edited by Jitendra Malik, University of California, Berkeley, CA, and approved April 10, 2020 (received for review Novemb

Many domains of science have developed complex simulations to describe phenomena of interest. While these simulations provide high-fidelity models, they are poorly suited for inference and lead to challenging inverse problems. We review the rapidly developing field of simulation-based inference and identify the forces giving additional momentum to the field. Finally, we describe how the frontier is expanding so that a broad audience can appreciate the profound influence these developments may have on science.

the simulator—is being recognized
sample efficiency of various inferen
of research has stopped treating the
focused on integrations that allow
into the internal details of the simu

Amidst this ongoing revolution,
based inference is changing rapid
provide the reader with a high-leve

LLMs for Science

FutureHouse

Building an AI Scientist.

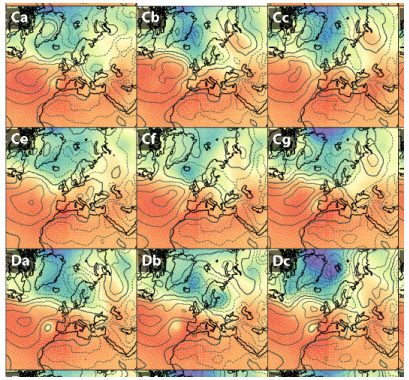
Not for today, but of great interest

Uncertainty propagation in complex physical systems

Generative emulation of weather forecast ensembles with diffusion models

Lizao Li, Robert Carver†, Ignacio Lopez-Gomez†, Fei Sha*, John Anderson

Uncertainty quantification is crucial to decision-making. A prominent example is probabilistic forecasting in numerical weather prediction. The dominant approach to representing uncertainty in weather forecasting is to generate an ensemble of forecasts by running physics-based simulations under different conditions, which is a computationally costly process. We propose to amortize the computational cost by emulating these forecasts with deep generative diffusion models learned from historical data. The learned models are highly scalable with respect to high-performance computing accelerators and can sample thousands of realistic weather forecasts at low cost. When designed to emulate operational ensemble forecasts, the generated ones are similar to physics-based ensembles in statistical properties and predictive skill. When designed to correct biases present in the operational forecasting system, the generated ensembles show improved probabilistic forecast metrics. They are more reliable and forecast probabilities of extreme weather events more accurately. While we focus on weather forecasting, this methodology may enable creating large climate projection ensembles for climate risk assessment.

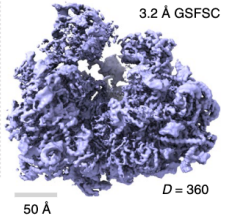


Scientific imaging & inverse problems

CryoDRGN: reconstruction of heterogeneous cryo-EM structures using neural networks

Ellen D. Zhong[1,2], Tristan Bepler[1,2], Bonnie Berger[2,3]✉ and Joseph H. Davis[1,4]✉

Cryo-electron microscopy (cryo-EM) single-particle analysis has proven powerful in dete
romolecules. However, many imaged protein complexes exhibit conformational and comp
major challenge to existing three-dimensional reconstruction methods. Here, we present
ages the representation power of deep neural networks to directly reconstruct continuous
map per-particle heterogeneity of single-particle cryo-EM datasets. Using cryoDRGN, w
in high-resolution datasets of the 80S ribosome and the RAG complex, revealed a new st
ribosome, and visualized large-scale continuous motions of a spliceosome complex. Cry
visualize a dataset's distribution of per-particle variability, generate density maps for
subsets for use with other tools and generate trajectories to visualize molecular motions
freely available at http://cryodrgn.csail.mit.edu.



NeRF for molecules!

Inference machinery for simulations

The frontier of simulation-based inference

Kyle Cranmer[a,b,1], Johann Brehmer[a,b], and Gilles Louppe[c]

[a]Center for Cosmology and Particle Physics, New York University, New York, NY 10003; [b]Center for Data Science, New Yor
and [c]Montefiore Institute, University of Liège, B-4000 Liège, Belgium

Edited by Jitendra Malik, University of California, Berkeley, CA, and approved April 10, 2020 (received for review Novemb

Many domains of science have developed complex simulations to describe phenomena of interest. While these simulations provide high-fidelity models, they are poorly suited for inference and lead to challenging inverse problems. We review the rapidly developing field of simulation-based inference and identify the forces giving additional momentum to the field. Finally, we describe how the frontier is expanding so that a broad audience can appreciate the profound influence these developments may have on science.

the simulator—is being recognized
sample efficiency of various inferen
of research has stopped treating the
focused on integrations that allow
into the internal details of the simu

Amidst this ongoing revolution,
based inference is changing rapid
provide the reader with a high-leve

LLMs for Science

FutureHouse

Building an AI Scientist.

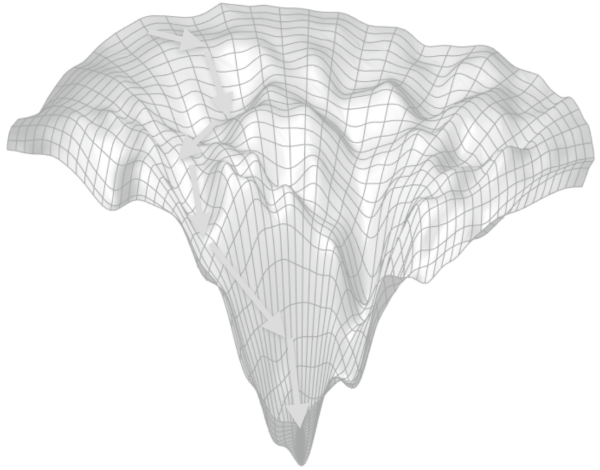
Three problems for generative modeling, three application areas

Genetics

AQKLYLTHIDAEVEGD
ADRLYMTKIHHQFDGD
ADTLFITEVKQVFEGD
ADRLYMTKIHHTFDGD
ADKLYCTLIHNSFEGD
ADRLYMTKIHHEFEGD

Likelihood estimation

Molecular Physics



Amortizing intractable inference

Molecular Engineering

function



Controllable generation

Three problems for generative modeling, three application areas

Genetics

AQKLYLTHIDAEVEGD
ADRLYMTKIHHQFDGD
ADTLFITEVKQVFEGD
ADRLYMTKIHHTFDGD
ADKLYCTLIHNSFEGD
ADRLYMTKIHHEFEGD

Likelihood estimation

Molecular Physics

Amortizing intractable inference

Molecular Engineering

Controllable generation

Mutations are constantly arising and now we can control them. What will they do?

We can read genomes at scale

The sequences of 150,119 genomes in the UK Biobank

https://doi.org/10.1038/s41586-022-04965-x

Received: 5 November 2021

Accepted: 10 June 2022

Published online: 20 July 2022

Open access

Check for updates

Bjarni V. Halldorsson[1,2]✉, Hannes P. Eggertsson[1], Kristjan H. S. Moore[1], Hannes Hauswedell[1], Ogmundur Eiriksson[1], Magnus O. Ulfarsson[1,3], Gunnar Palsson[1], Marteinn T. Hardarson[1,2], Asmundur Oddsson[1], Brynjar O. Jensson[1], Snaedis Kristmundsdottir[1,2], Brynja D. Sigurpalsdottir[1,2], Olafur A. Stefansson[1], Doruk Beyter[1], Guillaume Holley[1], Vinicius Tragante[1], Arnaldur Gylfason[1], Pall I. Olason[1], Florian Zink[1], Margret Asgeirsdottir[1], Sverrir T. Sverrisson[1], Brynjar Sigurdsson[1], Sigurjon A. Gudjonsson[1], Gunnar T. Sigurdsson[1], Gisli H. Halldorsson[1], Gardar Sveinbjornsson[1], Kristjan Norland[1], Unnur Styrkarsdottir[1], Droplaug N. Magnusdottir[1], Steinunn Snorradottir[1], Kari Kristinsson[1], Emilia Sobech[1], Helgi Jonsson[4,5], Arni J. Geirsson[4], Isleifur Olafsson[4], Palmi Jonsson[4,5], Ole Birger Pedersen[6], Christian Erikstrup[7,8], Søren Brunak[9], Sisse Rye Ostrowski[10,11], DBDS Genetic Consortium*, Gudmar Thorleifsson[1], Frosti Jonsson[1], Pall Melsted[1,3], Ingileif Jonsdottir[1,5], Thorunn Rafnar[1], Hilma Holm[1], Hreinn Stefansson[1], Jona Saemundsdottir[1], Daniel F. Gudbjartsson[1,3], Olafur T. Magnusson[1], Gisli Masson[1], Unnur Thorsteinsdottir[1,5], Agnar Helgason[1,12], Hakon Jonsson[1], Patrick Sulem[1] & Kari Stefansson[1]✉

sequencing of large cohorts with rich phenotypic data[1,2]. Here we describe the analysis of whole-genome sequencing of 150,119 individuals from the UK Biobank[3]. This constitutes a set of high-quality variants, including 585,040,410 single-nucleotide polymorphisms, representing 7.0% of all possible human single-nucleotide polymorphisms, and 58,707,036 indels. This large set of variants allows us to

We can write and edit genomes

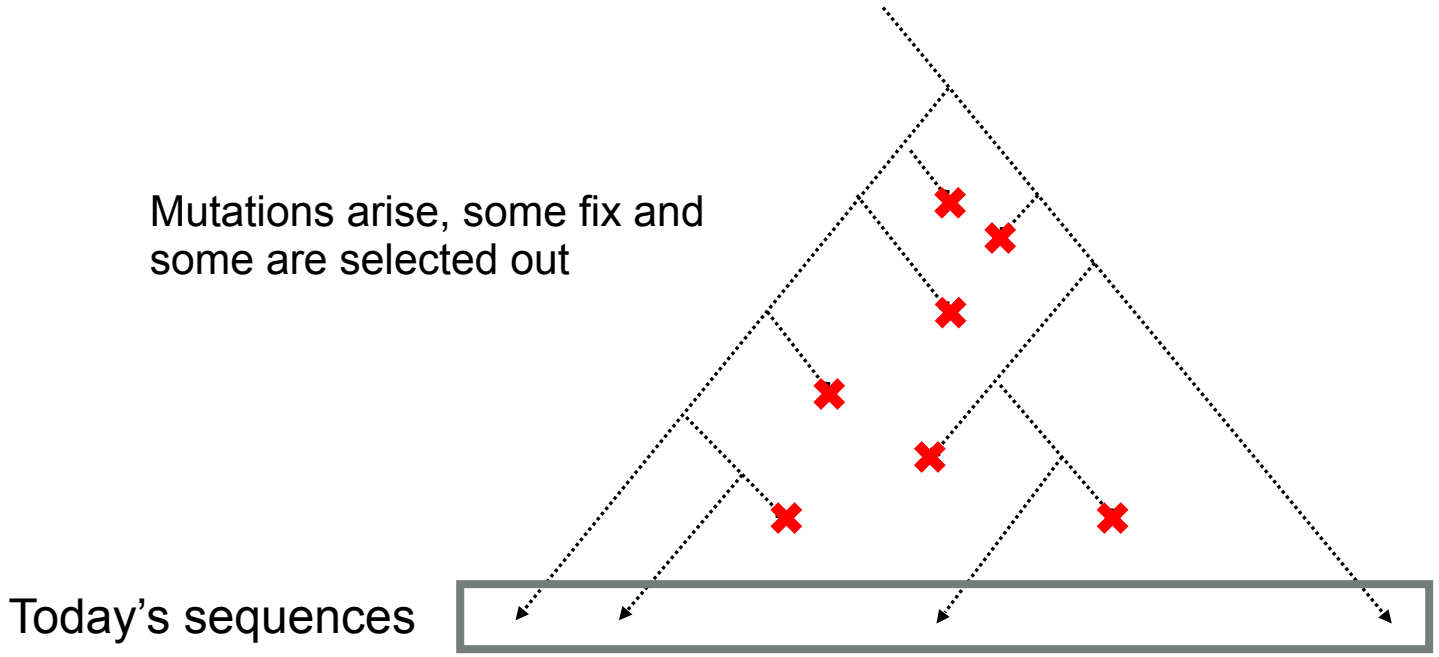
DNA WRITING TECHNOLOGIES MOVING TOWARD SYNTHETIC GENOMES

As new technologies hit the market, synthetic DNA is available faster and cheaper than ever before. Regulators are preparing to step in to limit opportunities for misuse. **By Caroline Seydel**

FDA NEWS RELEASE

FDA Approves First Gene Therapies to Treat Patients with Sickle Cell Disease

Idea: Sequences arose via evolution, so their **distribution** reflects **selection pressures of the past*.**

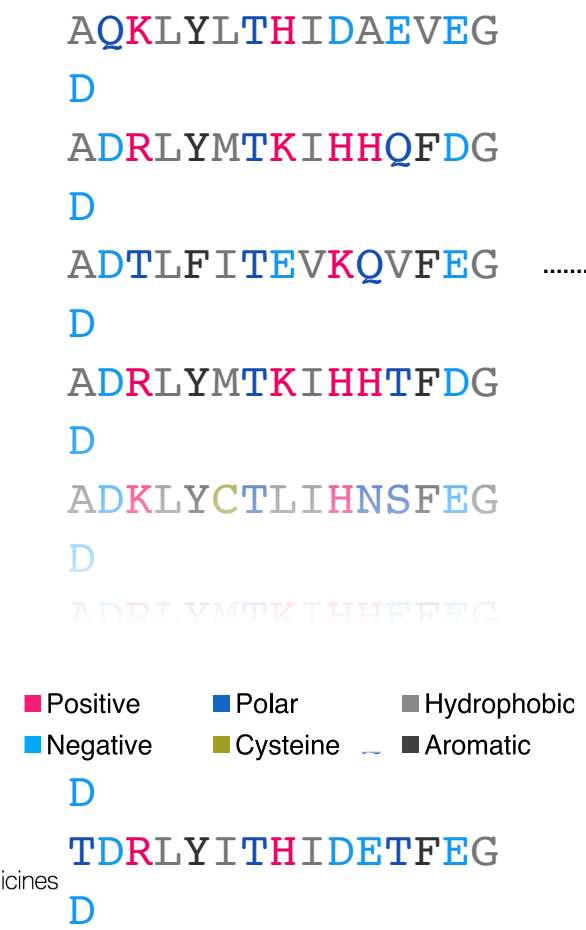


$$p(\text{sequence}|\text{evolution on earth})\,?$$

*ignoring population structure, shifting evolutionary pressures, and many other reasons why this shouldn't work

Idea: Estimate likelihoods for **unsupervised** estimation of **mutation effects**

1. Collect natural sequences



2. Learn a distribution

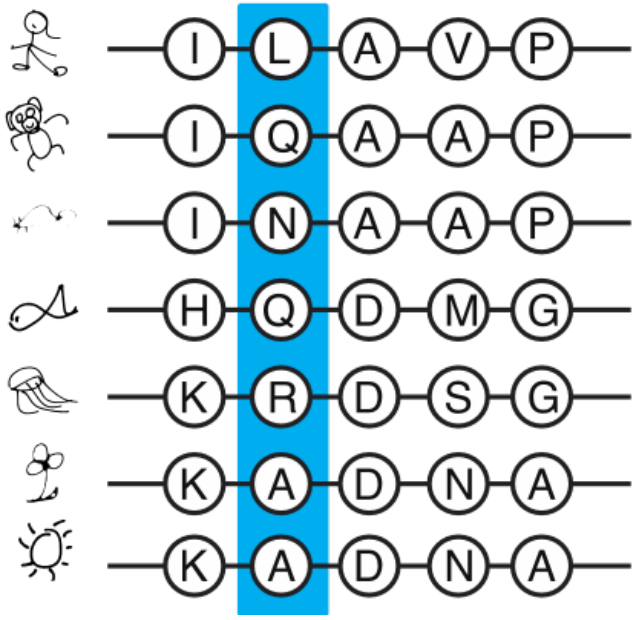
$$p_{\theta}(\mathbf{x})$$

3. Compute likelihood ratios for new mutations

$$\log \frac{p_{\theta}(\mathbf{x}^{\star})}{p_{\theta}(\mathbf{x})}$$

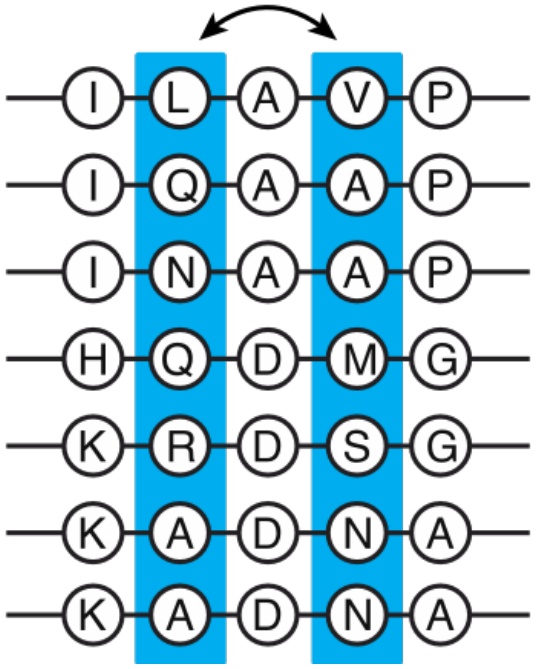
The data can be weak, so simple and/or regularized generative models have been favorable

Site-independent model



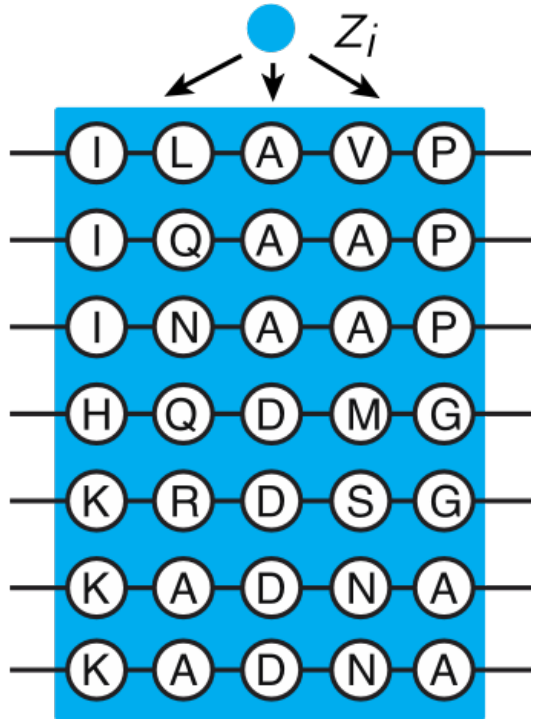
$$\prod_i p_i(x_i|\boldsymbol{\theta})$$

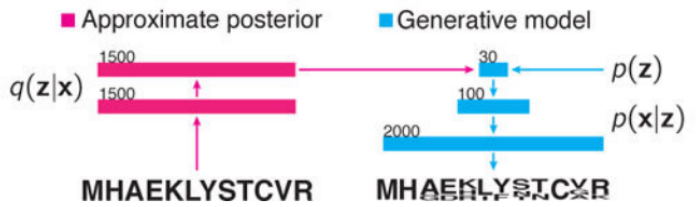
Potts model
(Markov Random Field)



$$\frac{1}{Z}\exp\left(\sum_{i<j} J_{ij}(x_i, x_j) + \sum_i h_i(x_i)\right)$$

(Bayesian) VAE

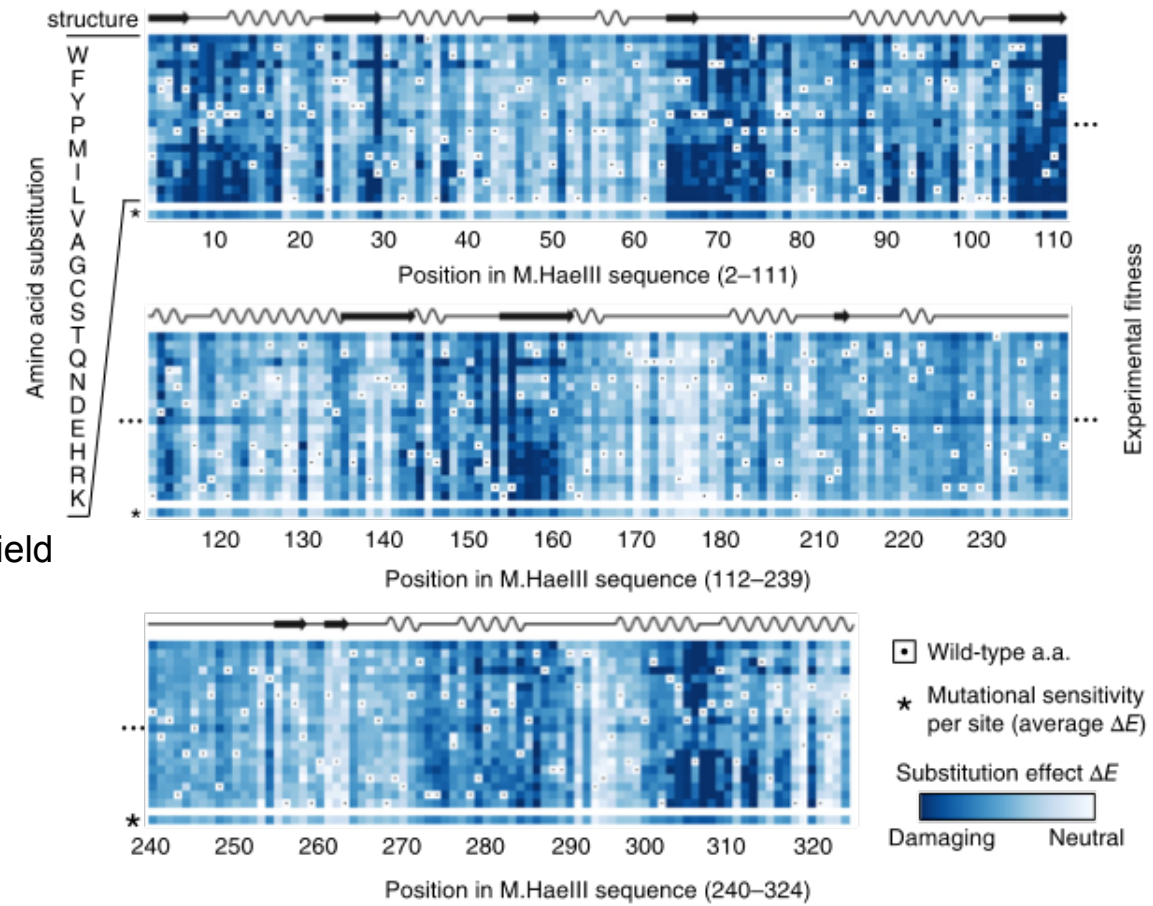




Sequence likelihoods predict experimental mutation effects, without training or fine tuning on experimental data

$$\log \frac{p_\theta(\mathbf{x}^\star)}{p_\theta(\mathbf{x})}$$

Markov Random Field
(Potts model)

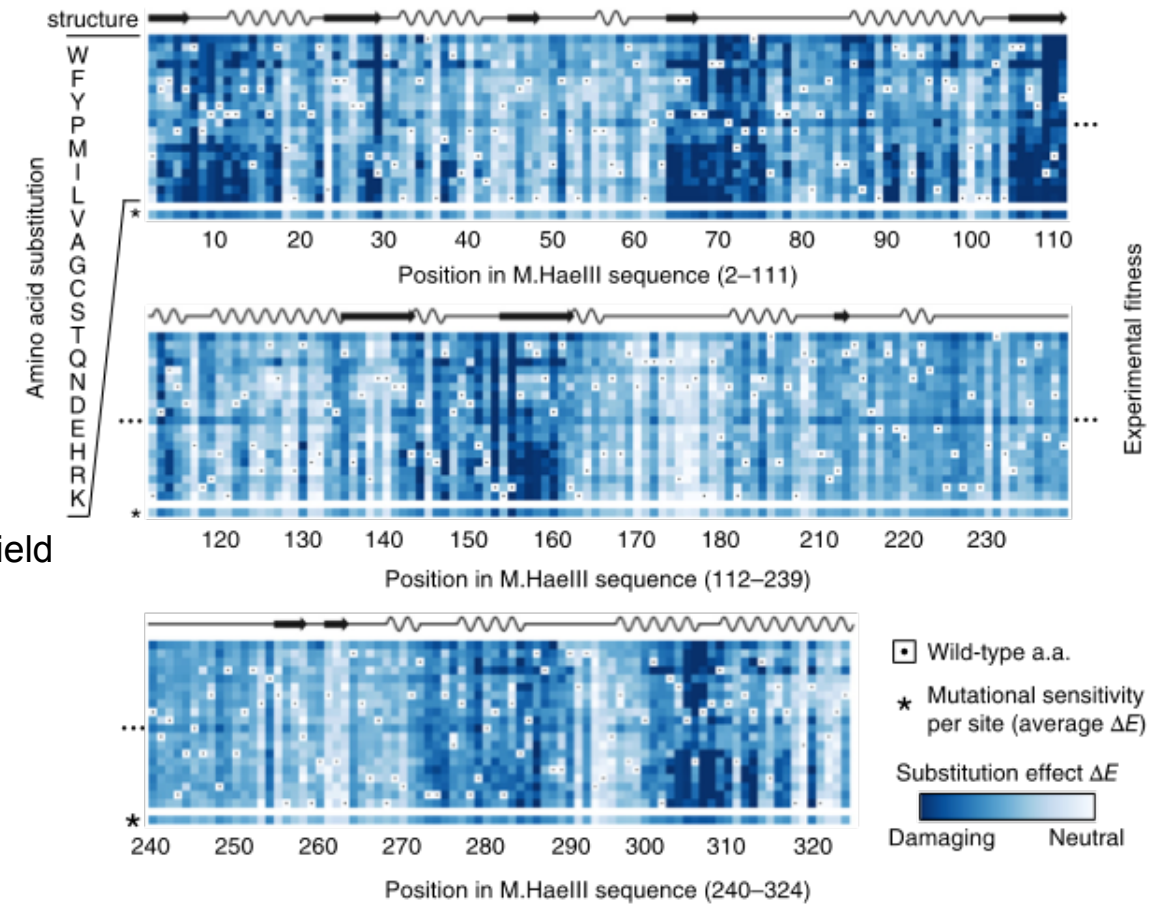


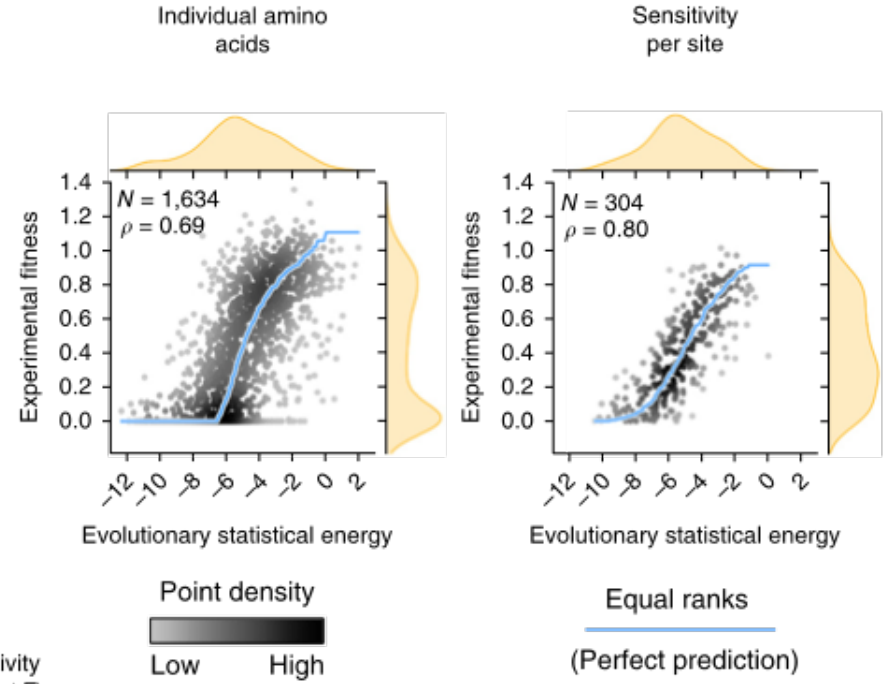
Hopf*, Ingraham,* et al. Nature Biotechnology (2017)

Sequence likelihoods predict experimental mutation effects, without training or fine tuning on experimental data

$$\log \frac{p_\theta(\mathbf{x}^\star)}{p_\theta(\mathbf{x})}$$

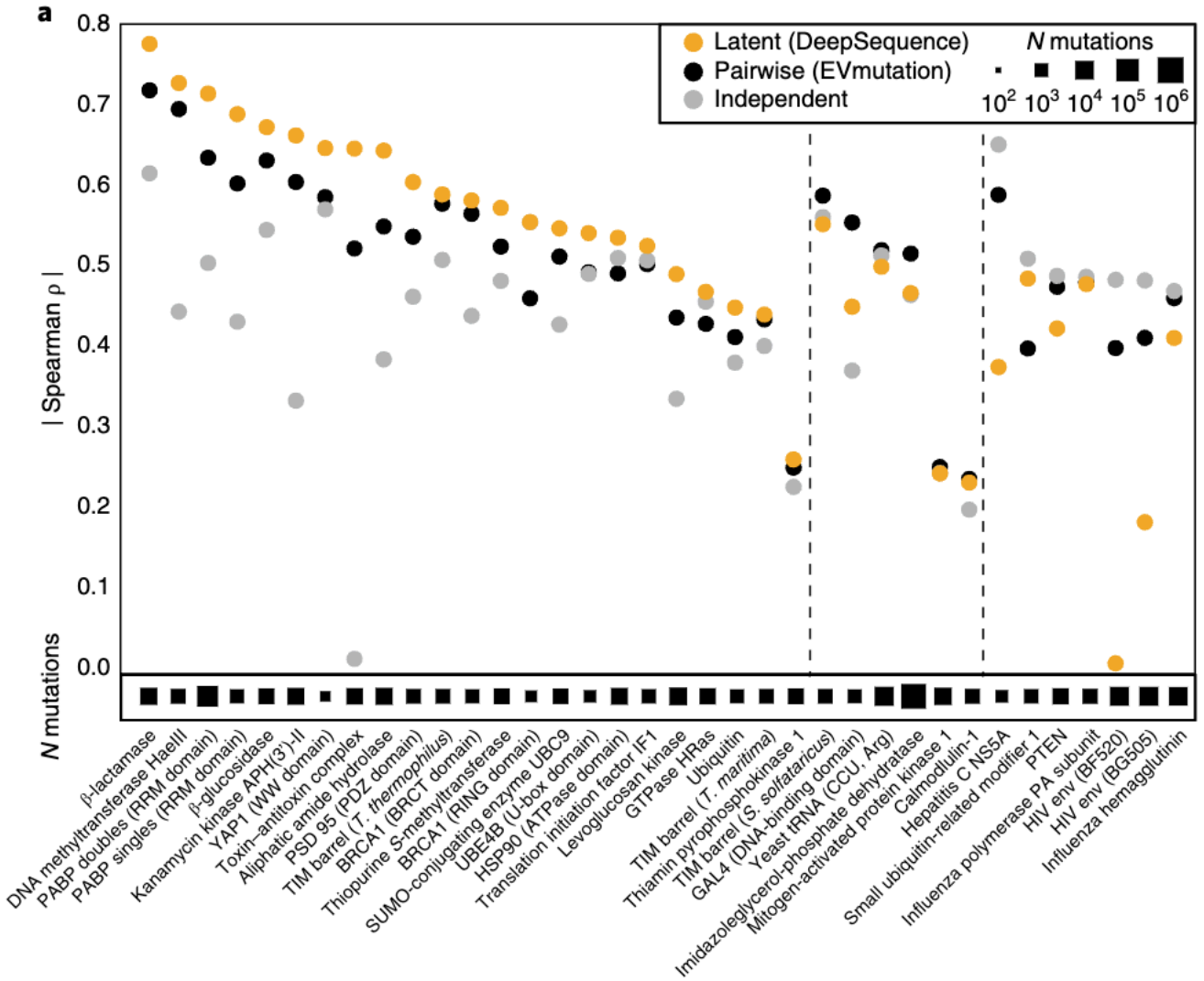
Markov Random Field (Potts model)





Hopf*, Ingraham,* et al. Nature Biotechnology (2017)

Sequence likelihoods predict experimental mutation effects across many different assay types and functions



What is function?

Whatever the protein was selected to do in evolution.

Interestingly, regularization was important, presumably because the data can be weak

	Variational Bayes for $\boldsymbol{\theta}$						*MAP* point estimate of $\boldsymbol{\theta}$								Pair	Site	N
Sparsity ★	✓	✓	✓														
Convolution ♦	✓	✓		✓													
Temperature ✠	✓		✓	✓	✓												
L2 Regularization							✓	✓	✓	✓	✓	✓					
Dropout							✓	✓			✓						
★ + ♦ + ✠							✓		✓								
Final ReLU								✓		✓			✓				
β-lactamase	0.73	0.73	**0.74**	0.73	0.73	**0.74**	0.56	0.61	0.04	0.40	0.56	0.37	0.34	0.42	0.70	0.60	4807
PDZ domain	**0.59**	**0.59**	**0.59**	0.57	0.57	0.55	0.56	0.48	0.32	0.47	0.50	0.41	0.37	0.47	0.54	0.47	1578
GAL4	0.48	0.46	0.52	**0.62**	0.60	0.58	0.60	0.53	0.30	0.47	0.52	0.43	0.42	0.47	0.59	0.41	1123
HSP90	**0.54**	0.53	**0.54**	0.51	0.52	0.52	0.48	0.45	0.03	0.34	0.44	0.26	0.22	0.33	0.49	0.43	4104
Kanamycin kinase	**0.63**	0.62	0.62	0.60	0.59	0.60	0.53	0.49	0.09	0.38	0.49	0.40	0.39	0.38	0.59	0.33	4385
HaeIII	**0.71**	**0.71**	0.69	0.70	0.68	0.68	0.66	0.64	0.08	0.54	0.64	0.50	0.49	0.54	0.69	0.44	1634
RRM domain	**0.66**	**0.66**	**0.66**	0.65	0.63	0.65	0.64	0.62	0.45	0.59	0.63	0.58	0.58	0.59	0.59	0.42	1188
Ubiquitin	0.44	**0.48**	0.47	0.44	**0.48**	0.43	0.37	0.39	0.09	0.38	0.37	0.29	0.31	0.38	0.43	0.46	1169
WW domain	**0.64**	0.63	**0.64**	0.63	0.63	**0.64**	**0.64**	0.58	0.31	0.50	0.61	0.49	0.44	0.50	0.57	0.58	319

(more) Bayesian VAE

Standard VAE

$$\iint \log \frac{p(\mathbf{X}|\mathbf{Z},\boldsymbol{\theta})p(\mathbf{Z})p(\boldsymbol{\theta})}{q(\mathbf{Z},\boldsymbol{\theta}|\mathbf{X},\boldsymbol{\phi})} q(\mathbf{Z},\boldsymbol{\theta}|\mathbf{X},\boldsymbol{\phi}) d\mathbf{Z} d\boldsymbol{\theta}$$

But there are *so many* protein sequences – Why settle for training on one family at a time?

Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences

Alexander Rives[a,b,1,2], Joshua Meier[a,1], Tom Sercu[a,1], Siddharth Goyal[a,1], Zeming Lin[b], Jason Liu[a], Demi Guo[c,3], Myle Ott[a], C. Lawrence Zitnick[a], Jerry Ma[d,e,3], and Rob Fergus[b]

[a]Facebook AI Research, New York, NY 10003; [b]Department of Computer Science, New York University, New York, NY 10012; [c]Harvard University, Cambridge, MA 02138; [d]Booth School of Business, University of Chicago, Chicago, IL 60637; and [e]Yale Law School, New Haven, CT 06511

Large language models generate functional protein sequences across diverse families

Received: 12 July 2022

Accepted: 17 November 2022

Ali Madani[1,2], Ben Krause[1,10], Eric R. Greene[3,10], Subu Subramanian[4,5], Benjamin P. Mohr[6], James M. Holton[7,8,9], Jose Luis Olmos Jr.[3], Caiming Xiong[1], Zachary Z. Sun[6], Richard Socher[1], James S. Fraser[3] & Nikhil Naik[1]

Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval

Pascal Notin [1] Mafalda Dias [2] Jonathan Frazer [2] Javier Marchena-Hurtado [2] Aidan Gomez [1 3]
Debora S. Marks [† 2] Yarin Gal [† 1]

Language models enable zero-shot prediction of the effects of mutations on protein function

Joshua Meier [1 2] Roshan Rao [3] Robert Verkuil [1] Jason Liu [1]
Tom Sercu [1] Alexander Rives [1 2]

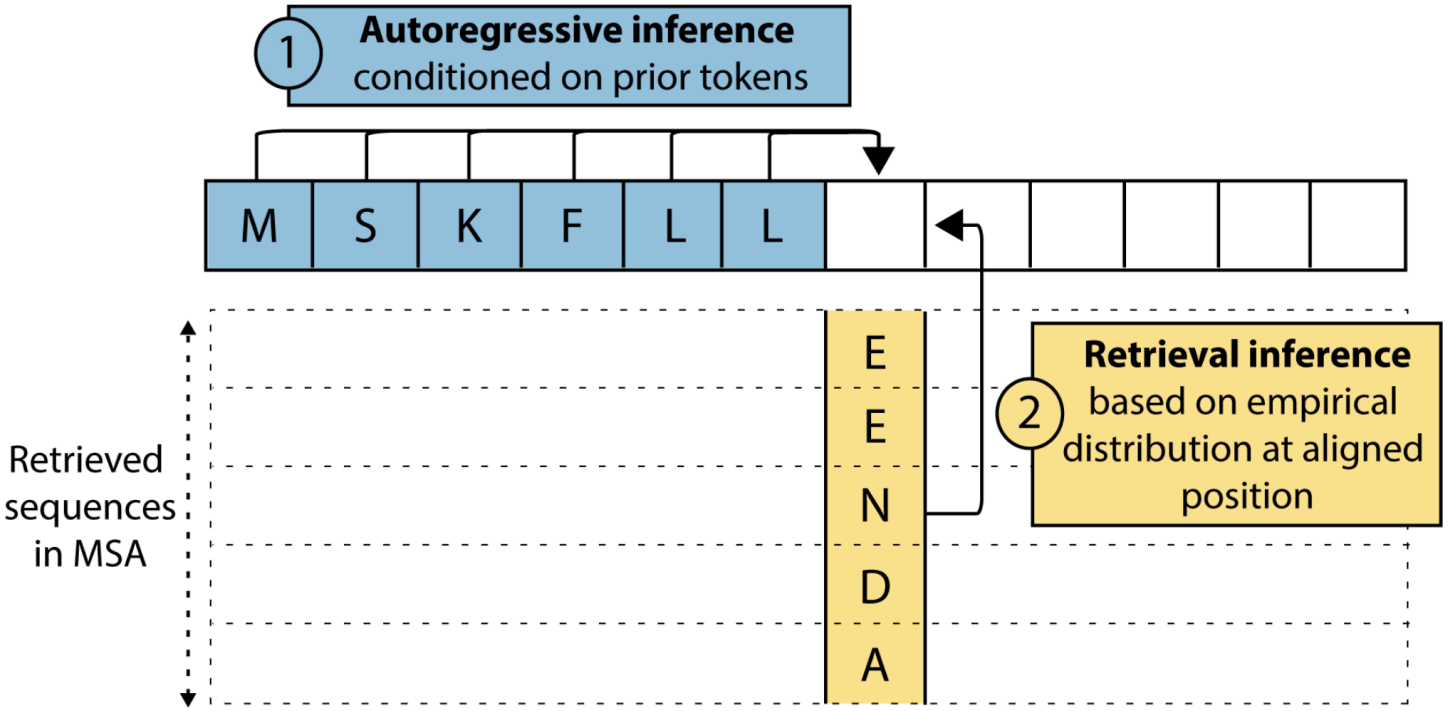
Interestingly, the one-family-at-a-time models can still be hard to beat

Model type	Model name	Spearman's rank correlation by MSA depth ↑ Low	Medium	High	All	AUC ↑ All
Alignment-based models	Site indep	0.428	0.403	0.350	0.397	0.725
	Wavenet	0.319	0.398	0.469	0.398	0.725
	DeepSequence	0.375	0.397	0.506	0.415	0.733
Potts	EVmutation	0.401	0.421	0.468	0.427	0.738
Bayesian VAEs	EVE	0.408	**0.440**	**0.507**	0.448	0.751
Protein language models	ESM-1v	0.321	0.348	0.484	0.371	0.713
	MSA Transformer	0.373	0.418	0.482	0.422	0.737
	Tranception (w/o retrieval)	0.394	0.398	0.439	0.406	0.728
LM	Tranception (w/ retrieval)	**0.453**	0.438	0.488	**0.451**	**0.754**

Tranception: Protein Fitness Prediction with Autoregressive Transformers and Inference-time Retrieval

Pascal Notin [1] Mafalda Dias [2] Jonathan Frazer [2] Javier Marchena-Hurtado [2] Aidan Gomez [1 3] Debora S. Marks [† 2] Yarin Gal [† 1]

Most successful pure language method based on inference with retrieval



LM likelihood

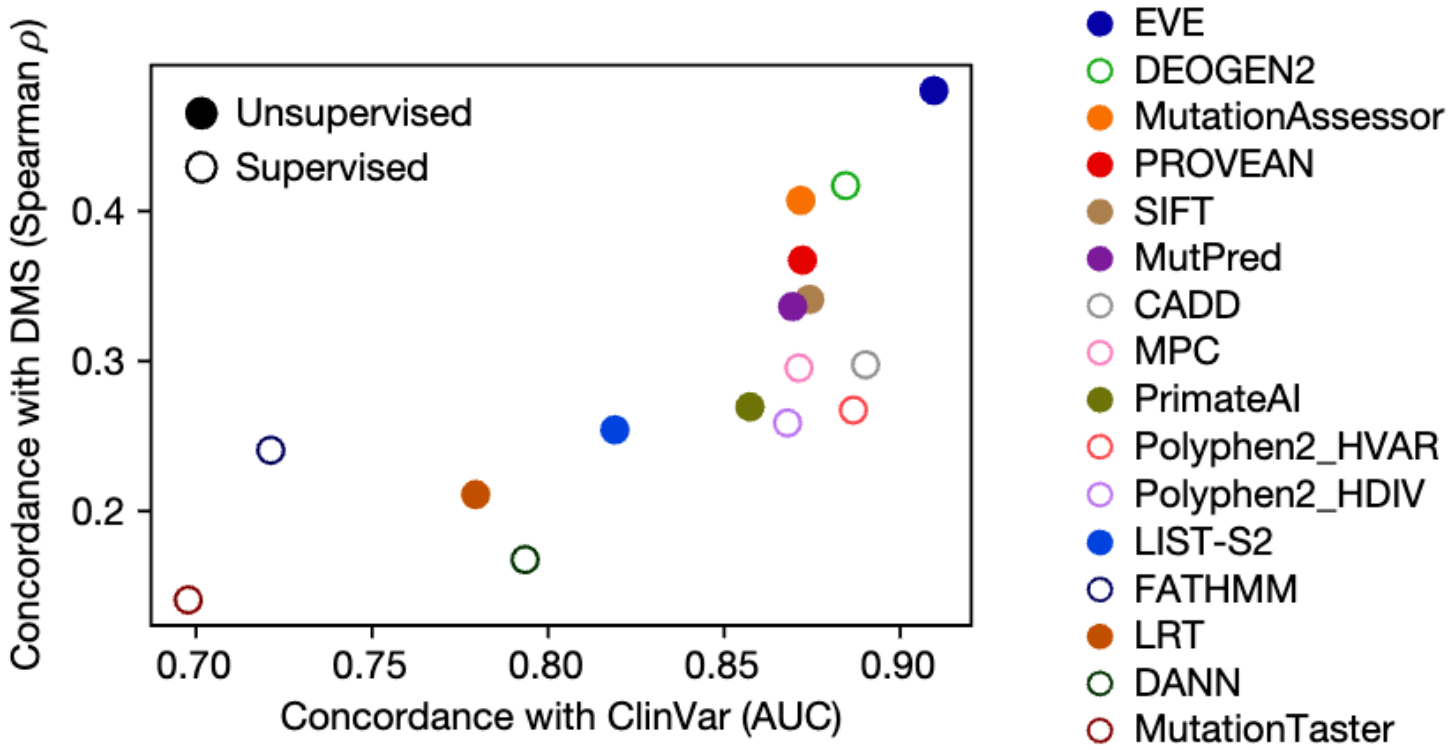
$$\log P(x) \propto \sum_{i=1}^{l} [(1-\alpha) \log P_A(x_i|x_{<i}) + \alpha \log P_R(x_i)]$$

Empirical first order frequencies

Tranception: Protein Fitness Prediction with Autoregressive Transformers and Inference-time Retrieval

Pascal Notin [1] Mafalda Dias [2] Jonathan Frazer [2] Javier Marchena-Hurtado [2] Aidan Gomez [1 3]
Debora S. Marks [† 2] Yarin Gal [† 1]

These approaches have proven useful for interpreting human genomic data



Disease variant prediction with deep generative models of evolutionary data

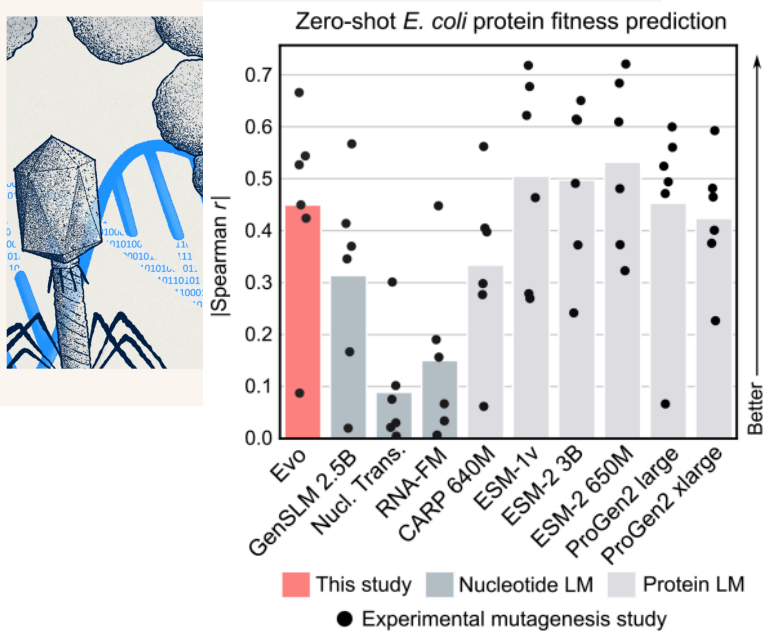
https://doi.org/10.1038/s41586-021-04043-8
Received: 18 December 2020

Jonathan Frazer[1,4], Pascal Notin[2,4], Mafalda Dias[1,4], Aidan Gomez[2], Joseph K. Min[1], Kelly Brock[1], Yarin Gal[2✉] & Debora S. Marks[1,3✉]

Future directions for generative models of genomic sequences

Genomic LMs?

Evo: DNA foundation modeling from molecular to genome scale



Unify with structure-oriented models?

Accurate proteome-wide missense variant effect prediction with AlphaMissense

Jun Cheng*, Guido Novati, Joshua Pan†, Clare Bycroft†, Akvilė Žemgulytė†, Taylor Applebaum†, Alexander Pritzel, Lai Hong Wong, Michal Zielinski, Tobias Sargeant, Rosalia G. Schneider, Andrew W. Senior, John Jumper, Demis Hassabis, Pushmeet Kohli*, Žiga Avsec*

Protein design using structure-based residue preferences

Received: 19 July 2023

Accepted: 29 January 2024

David Ding[1], Ada Y. Shaw[2], Sam Sinai[3], Nathan Rollins[4], Noam Prywes[1], David F. Savage[1,5,6], Michael T. Laub[7,8] & Debora S. Marks[2]

Algorithms still matter!

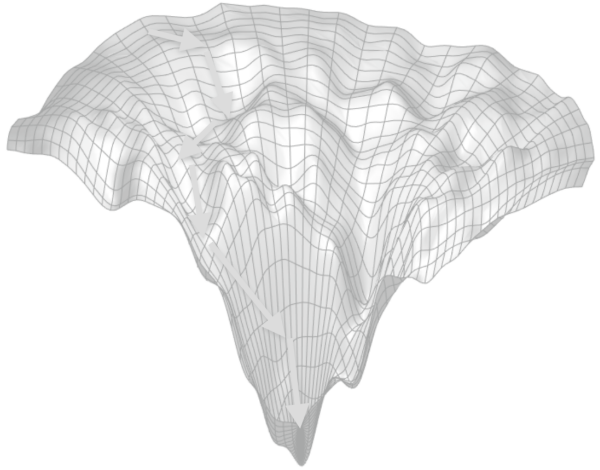
Three problems for generative modeling, three application areas

Genetics

AQKLYLTHIDAEVEGD
ADRLYMTKIHHQFDGD
ADTLFITEVKQVFEGD
ADRLYMTKIHHTFDGD
ADKLYCTLIHNSFEGD
ADRLYMTKIHHEFEGD

Likelihood estimation

Molecular Physics



Amortizing intractable inference

Molecular Engineering

function

Controllable generation

Molecular physics is … an inference problem?

“Well-defined*” energy function

$$U(\mathbf{x}) = \sum_{\text{bonds}} k_b(b - b_0)^2 + \sum_{\text{angles}} k_\theta(\theta - \theta_0)^2 + \sum_{\text{dihedrals}} k_\phi[1 + cos(n\phi - \delta)]$$
$$+ \sum_{\text{impropers}} k_\omega(\omega - \omega_0)^2 + \sum_{\text{Urey-Bradley}} k_u(u - u_0)^2$$
$$+ \sum_{\text{nonbonded}} \left(\epsilon_{ij} \left[\left(\frac{R_{min_{ij}}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{min_{ij}}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{\epsilon_r r_{ij}} \right)$$

*classical approximation of QM!
No bond breaking. This one is CHARM22

Well-defined distribution

$$p(\mathbf{x}) = \frac{1}{Z} \exp\left(-\beta U(\mathbf{x})\right)$$

Molecular physics is … an inference problem?

“Well-defined*” energy function

$$U(\mathbf{x}) = \sum_{\text{bonds}} k_b(b - b_0)^2 + \sum_{\text{angles}} k_\theta(\theta - \theta_0)^2 + \sum_{\text{dihedrals}} k_\phi[1 + cos(n\phi - \delta)]$$
$$+ \sum_{\text{impropers}} k_\omega(\omega - \omega_0)^2 + \sum_{\text{Urey-Bradley}} k_u(u - u_0)^2$$
$$+ \sum_{\text{nonbonded}} \left(\epsilon_{ij} \left[\left(\frac{R_{min_{ij}}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{min_{ij}}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{\epsilon_r r_{ij}} \right)$$

*classical approximation of QM!
No bond breaking. This one is CHARM22

Well-defined distribution

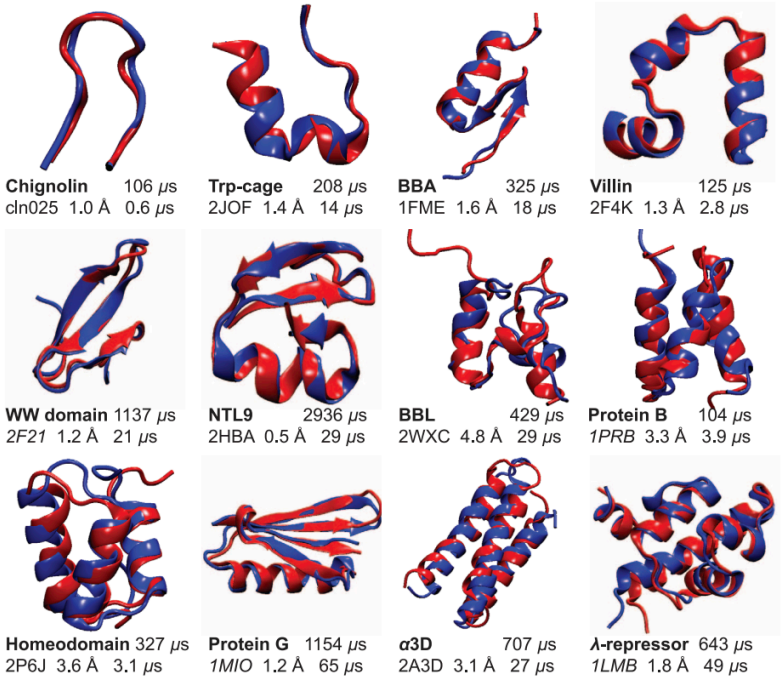
$$p(\mathbf{x}) = \frac{1}{Z} \exp\left(-\beta U(\mathbf{x})\right)$$

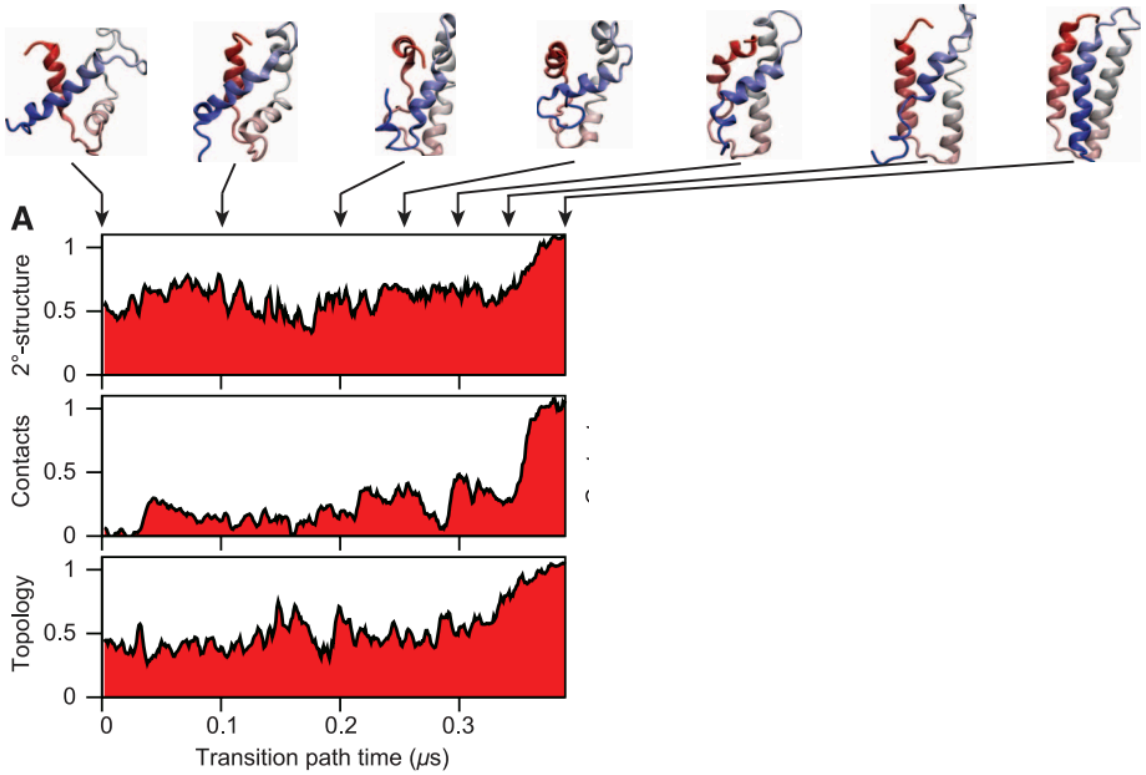


Brute force molecular dynamics can fold *small* proteins… with great effort

How Fast-Folding Proteins Fold

Kresten Lindorff-Larsen,[1]*† Stefano Piana,[1]*† Ron O. Dror,[1] David E. Shaw[1,2]†





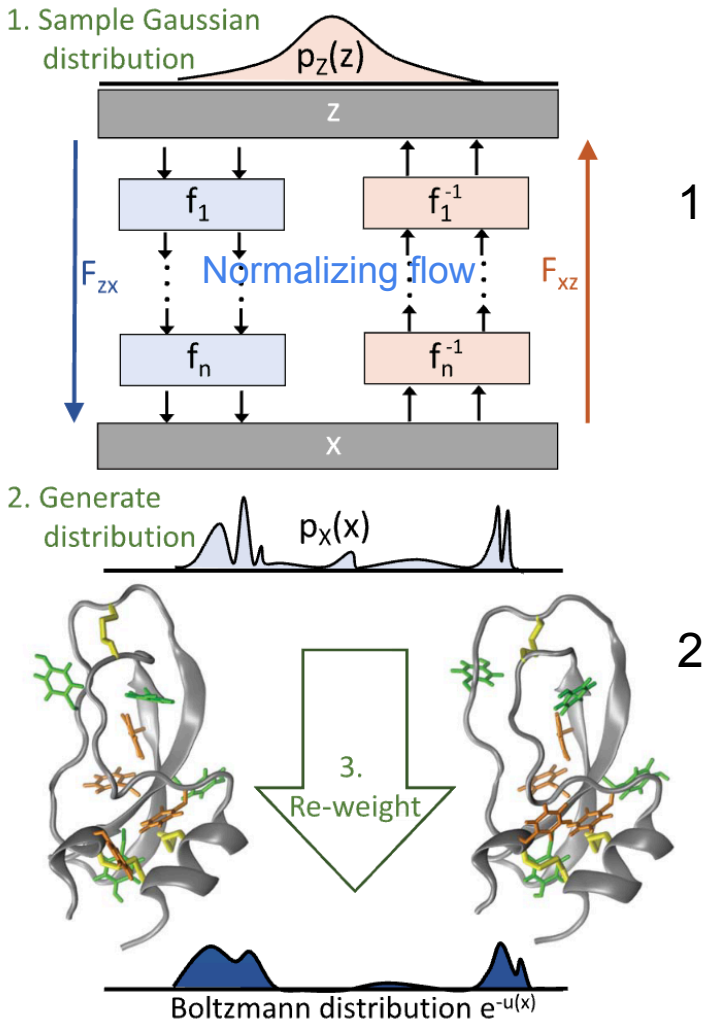
MD with a Thermostat (Langevin-like dynamics) on the energy

Typical time step ~2.5 femtoseconds (10^{-15} seconds)
→ **~400 million time steps** for a microsecond!

Can deep generative models help us attack this intractable sampling problem?
Idea 1: Boltzmann Generators

Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning

Frank Noé[1,2,3]*†, Simon Olsson[1]*, Jonas Köhler[1]*, Hao Wu[4,1]



1. Learn a normalizing flow by minimizing KL Divergence* (similar to distillation)

$$J_{KL} = \mathbb{E}_{\mathbf{z}}[u(F_{zx}(\mathbf{z})) - \log R_{zx}(\mathbf{z})]$$

Energy LogDetJac

2. Use importance weighting to clean up distributional inconsistencies

$$w_X(\mathbf{x}) = \frac{\mu_X(\mathbf{x})}{q_X(\mathbf{x})} = \frac{q_Z(\mathbf{z})}{\mu_Z(\mathbf{z})} \qquad \mathbb{E}[O] \approx \frac{\sum_{i=1}^{N} w_X(\mathbf{x}) O(\mathbf{x})}{\sum_{i=1}^{N} w_X(\mathbf{x})}$$

Importance weights

*Finnicky, bidirectional KL + MLE side estimation used in practice

Can deep generative models help us attack this intractable sampling problem?
Idea 1: Boltzmann Generators

Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning

Frank Noé[1,2,3]*†, Simon Olsson[1]*, Jonas Köhler[1]*, Hao Wu[4,1]

$$J = w_{ML}J_{ML} + w_{KL}J_{KL} + w_{RC}J_{RC}$$

In practice, need to pre-run some molecular dynamics for MLE co-training

SCALABLE NORMALIZING FLOWS ENABLE BOLTZMANN GENERATORS FOR MACROMOLECULES

Joseph C. Kim[1,2], **David Bloore**[1], **Karan Kapoor**[1], **Jun Feng**[1]
Ming-Hong Hao[1], **Mengdi Wang**[2]
[1]Ensem Therapeutics, Waltham, MA
[2]Department of Electrical and Computer Engineering, Princeton University

In their seminal paper, Noé et al. (2019) used a convex combination of the two loss terms, in the context of BGs, in order to both avoid mode-collapse and generate low-energy samples. However, for larger molecules, target evaluation is computationally expensive and dramatically slows iterative training with the reverse KL divergence objective. In addition, during the early stages of training, the KL divergence explodes and leads to unstable training. One way to circumvent these issues is to train with the NLL loss, followed by a combination of both loss terms. Unfortunately, for larger systems, the KL term tends to dominate and training often get stuck at non-optimal local minima. In order to remedy these issues, we consider a sequential training scheme, whereby we smooth the transition from maximum likelihood training to the combination of maximum likelihood and reverse KL divergence minimization.

(1) As mentioned previously, we first train with maximum likelihood to convergence.

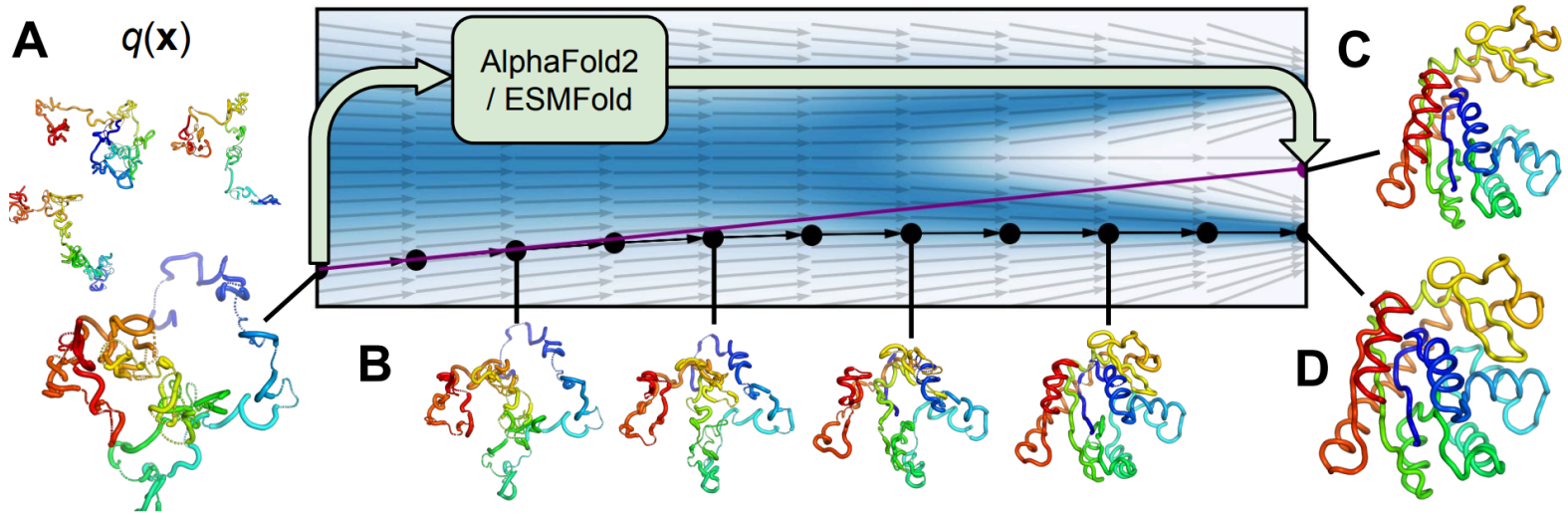
(2) Afterward, we train with a combination of the NLL and the 2-Wasserstein loss with respect to distance matrices of the backbone atoms:

Boltzmann generators remain tricky to train and currently focus on only modeling equilibrium fluctuations around a *known* fold.
(i.e. They are **not** modeling folding or structure from scratch)
Lots of room for further development

Can deep generative models help us attack this intractable sampling problem?
Idea 2: Fine-tune a point estimate (e.g. AlphaFold) with Diffusion/Flow-matching

AlphaFold Meets Flow Matching for Generating Protein Ensembles

Bowen Jing [1] **Bonnie Berger** [1 2] **Tommi Jaakkola** [1]



Algorithm 1 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 $\mathbf{x}_1 \leftarrow \text{BetaCarbons}(S_i)$
  Sample $\mathbf{x}_0 \sim \text{HarmonicPrior}(\text{length}(A_i))$
  Align $\mathbf{x}_0 \leftarrow \text{RMSDAlign}(\mathbf{x}_0, \mathbf{x}_1)$
  Sample $t \sim \text{Uniform}[0, 1]$
  Interpolate $\mathbf{x}_t \leftarrow t \cdot \mathbf{x}_1 + (1 - t) \cdot \mathbf{x}_0$
  Predict $\hat{S}_i \leftarrow \text{AlphaFold}(A_i, M_i, \mathbf{x}_t, t)$
  Optimize loss $\mathcal{L} = \text{FAPE}^2(\hat{S}_i, S_i)$

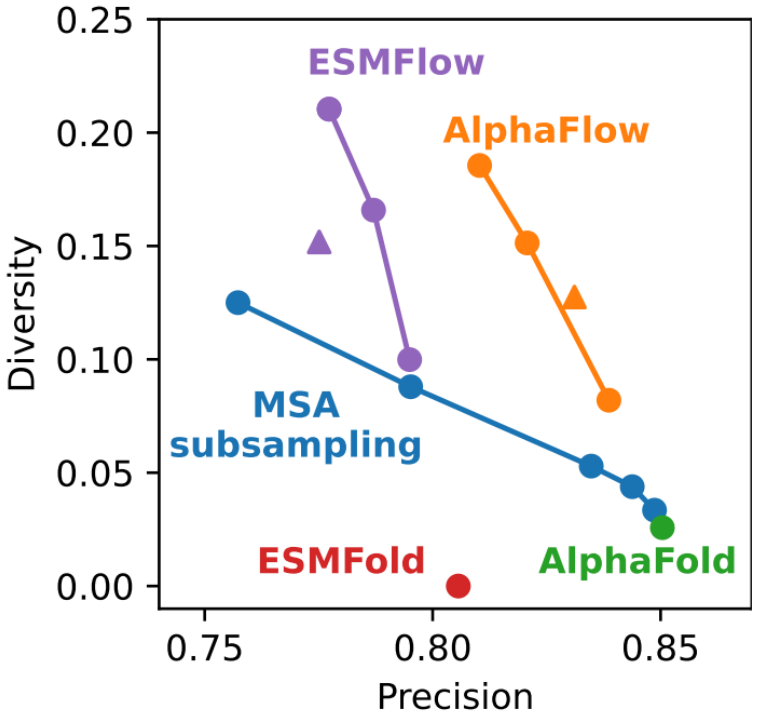
Algorithm 2 INFERENCE

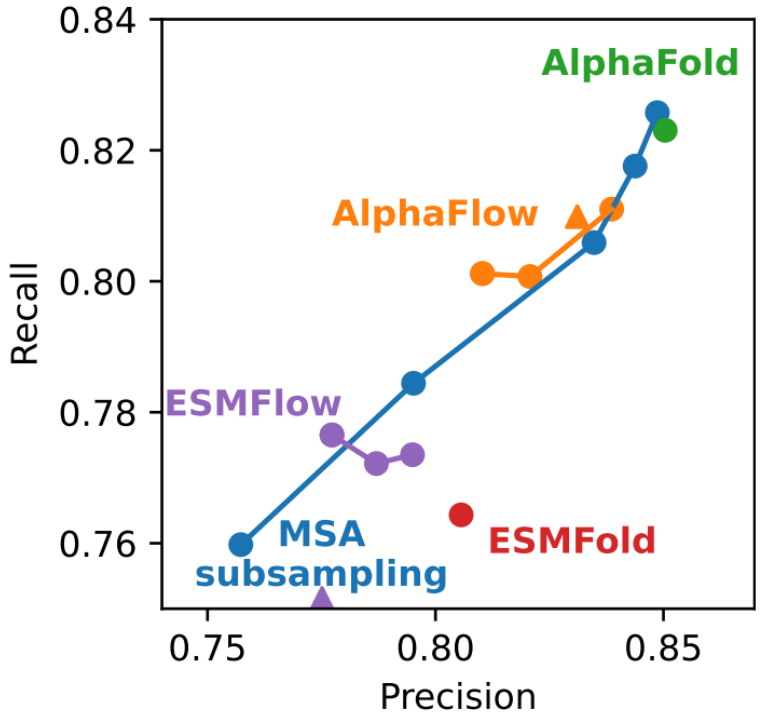
Input: Sequence and MSA (A, M)
Output: Sampled all-atom structure $\hat{S}$
Sample $\mathbf{x}_0 \sim \text{HarmonicPrior}(\text{length}(A))$
for $n \leftarrow 0$ to $N - 1$ **do**
  Let $t \leftarrow n/N$ and $s \leftarrow t + 1/N$
  Predict $\hat{S} \leftarrow \text{AlphaFold}(A, M, \mathbf{x}_t, t)$
  if $n = N - 1$ **then**
    return $\hat{S}$
  Extract $\hat{\mathbf{x}}_1 \leftarrow \text{BetaCarbons}(\hat{S})$
  Align $\mathbf{x}_t \leftarrow \text{RMSDAlign}(\mathbf{x}_t, \hat{\mathbf{x}}_1)$
  Interpolate $\mathbf{x}_s \leftarrow \frac{s-t}{1-t} \cdot \hat{\mathbf{x}}_1 + \frac{1-s}{1-t} \cdot \mathbf{x}_t$

Can deep generative models help us attack this intractable sampling problem?
Idea 2: Fine-tune a point estimate (e.g. AlphaFold) with Diffusion/Flow-matching

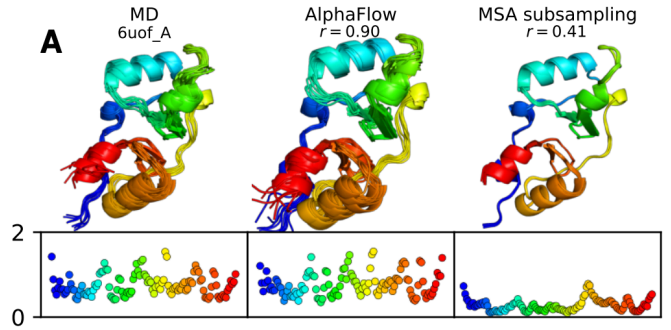
AlphaFold Meets Flow Matching for Generating Protein Ensembles

Bowen Jing [1] **Bonnie Berger** [1 2] **Tommi Jaakkola** [1]





Add significant diversity with little cost to precision



Three problems for generative modeling, three application areas

Genetics

AQKLYLTHIDAEVEGD
ADRLYMTKIHHQFDGD
ADTLFITEVKQVFEGD
ADRLYMTKIHHTFDGD
ADKLYCTLIHNSFEGD
ADRLYMTKIHHEFEGD

Likelihood estimation

Molecular Physics

Amortizing intractable inference

Molecular Engineering

function ⋯⋯▸



Controllable generation

Proteins are routinely best in class solutions to complex problems

Cancer and complex disease

Infectious disease

Nanotechnology and materials science

Proteins are routinely best in class solutions to complex problems

Cancer and complex disease

Infectious disease

Nanotechnology and materials science

Making these nanomachines requires?

$200M EUV Machine?



Proteins are routinely best in class solutions to complex problems

Cancer and complex disease

Infectious disease

Nanotechnology and materials science

Making these nanomachines requires?

$200M EUV Machine?



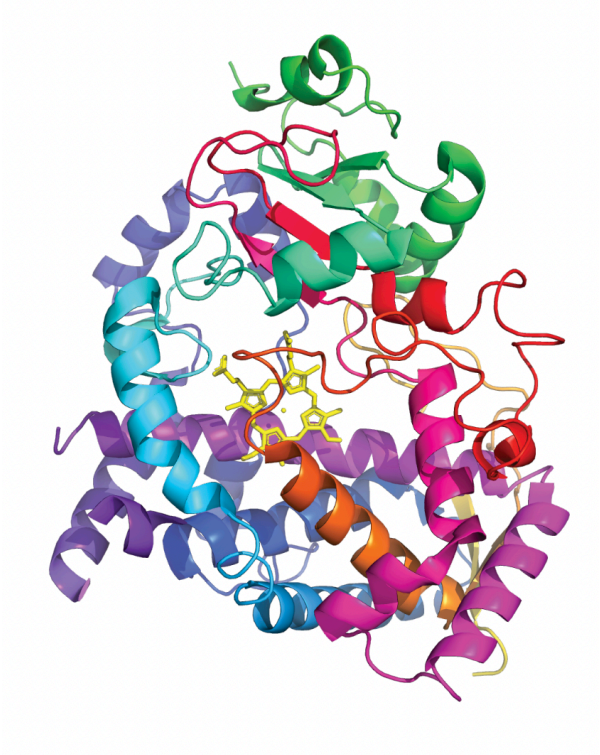
Beer brewing reactor



Most industrially relevant proteins at least today are edits of natural proteins, not designed from scratch

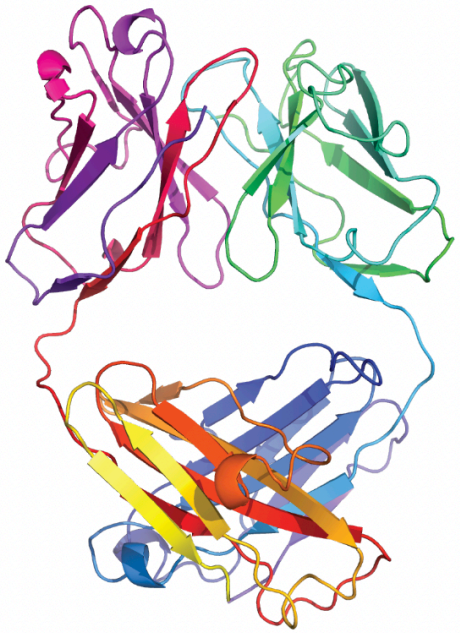
Directed Evolution of enzymes

Cytochrome P450

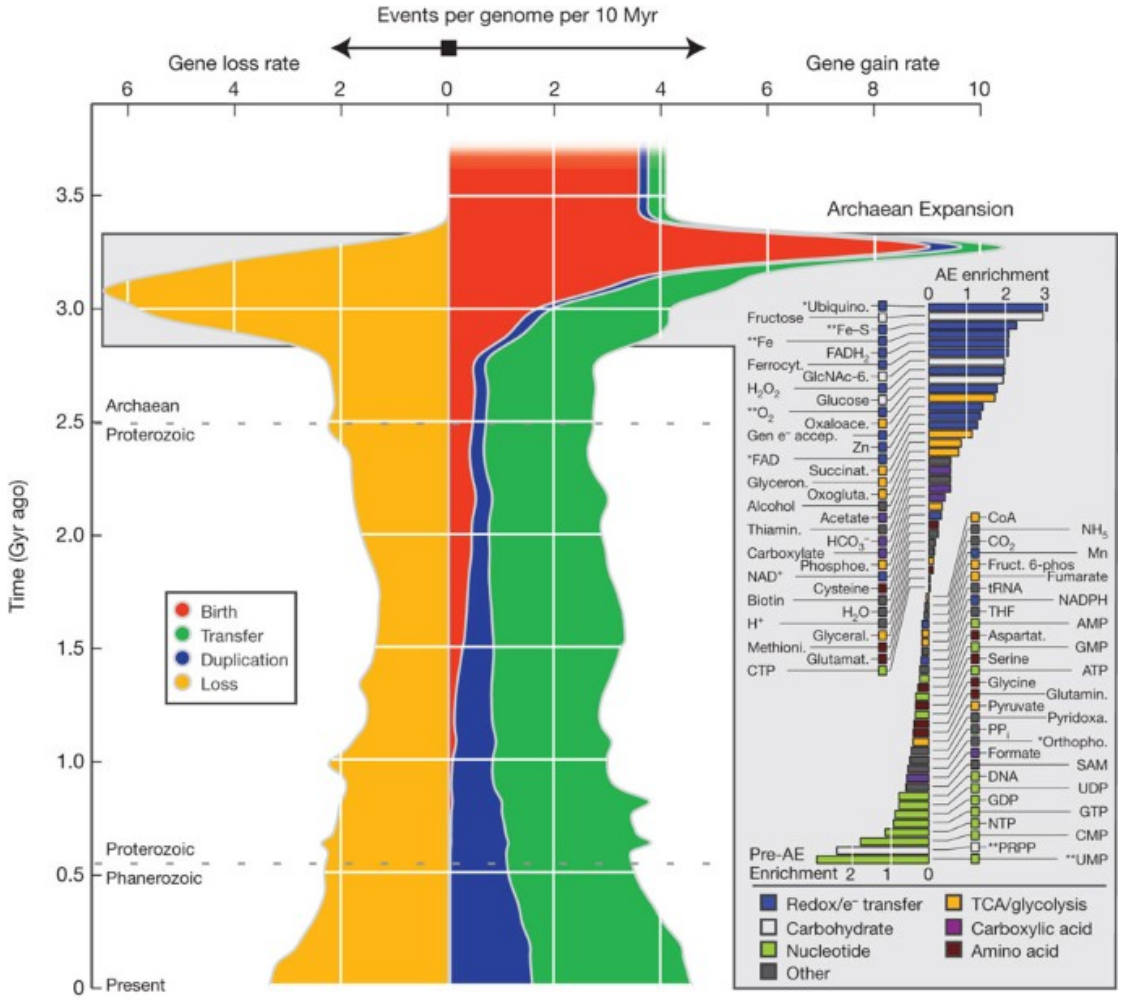


Discovery of antibodies

Human Antibody



In fact, Evolution discovered most interesting proteins ~3 billion years ago and then focused on mutating

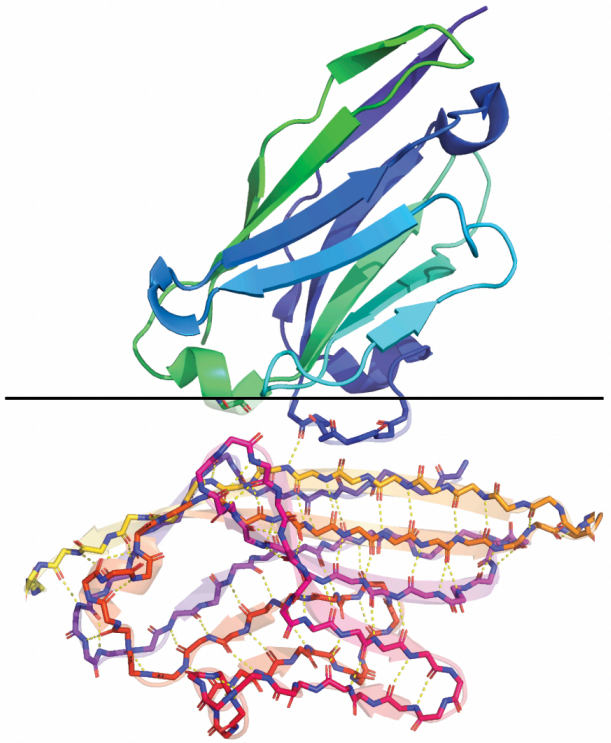


Rapid evolutionary innovation during an Archaean genetic expansion

Lawrence A. David[1] & Eric J. Alm[1,2,3]

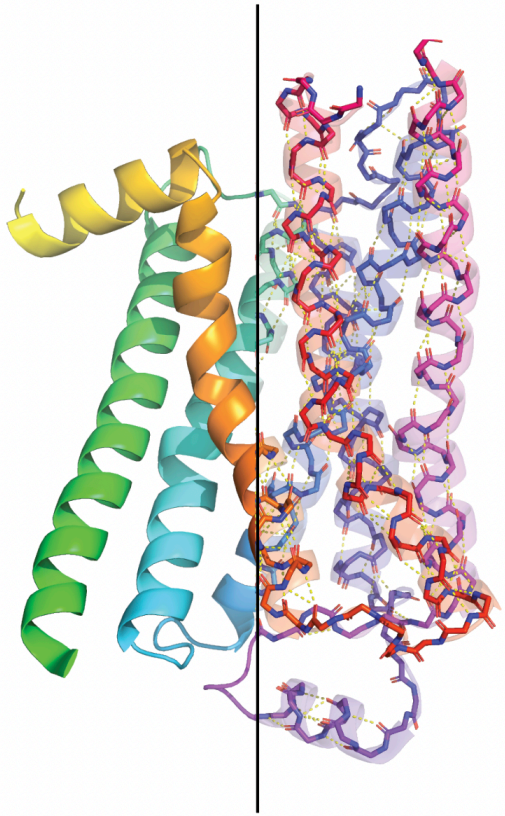
Why have there been so few new proteins?
Proteins function by **complex constraints** that are unforgiving if not satisfied all at once

Nanobody complex



Long range hydrogen bonds in β sheets

GPCR



Local hydrogen bonds in α helices

- Self-avoiding curve (polymer)
- Need to “lock” into special conformations full of hydrogen bonds
- To force a specific fold, must stabilize with 20 letter side chain alphabet (sequence)

It’s not like images are without complex constraints?

It’s not like images are without complex constraints?



OpenAI DALL-E 3

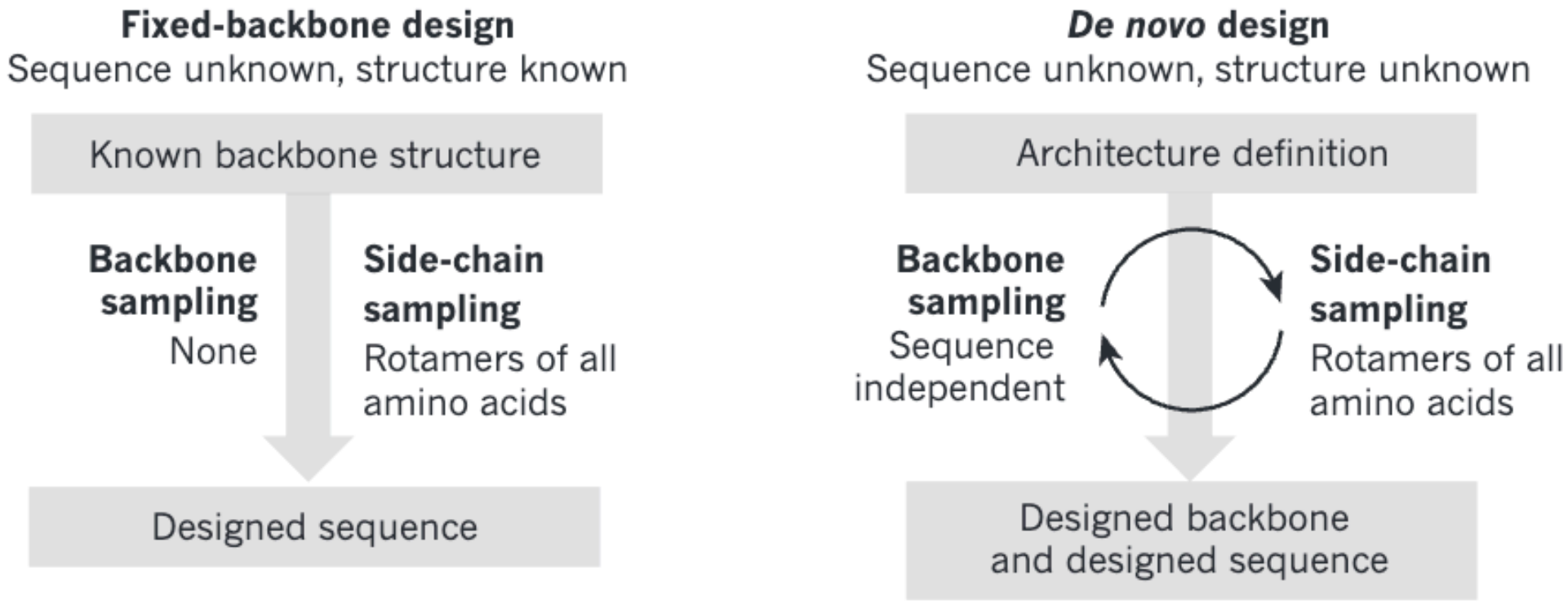
It’s not like images are without complex constraints?



OpenAI DALL-E 3

Spoiler: In the past few years, deep generative models “ate” computational protein design

The classic approaches

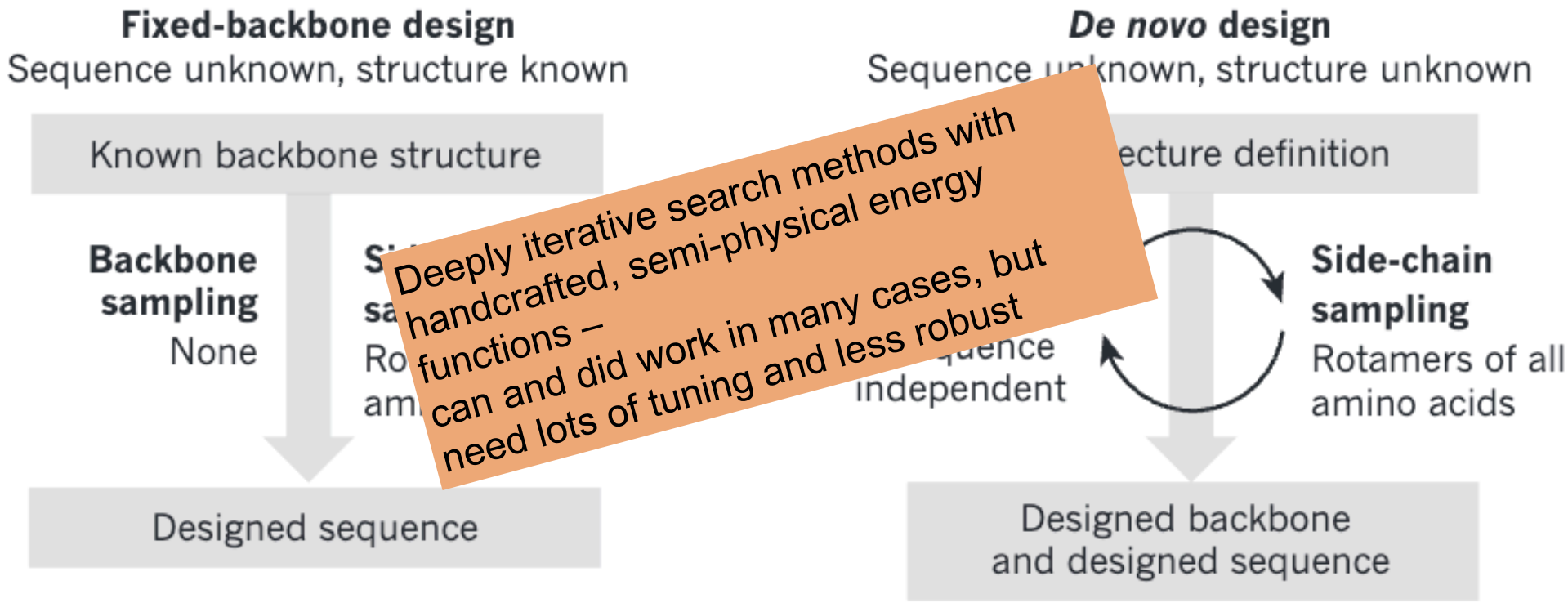


The coming of age of *de novo* protein design

Po-Ssu Huang[1,2]*†, Scott E. Boyken[1,2,3]* & David Baker[1,2,3]

Spoiler: In the past few years, deep generative models “ate” computational protein design

The classic approaches



Deeply iterative search methods with handcrafted, semi-physical energy functions –
can and did work in many cases, but need lots of tuning and less robust

The coming of age of *de novo* protein design

Po-Ssu Huang[1,2]*†, Scott E. Boyken[1,2,3]* & David Baker[1,2,3]

Spoiler: In the past few years, deep generative models “ate” de novo protein design

Structure-conditioned sequence design

Generative models for graph-based protein design

John Ingraham, Vikas K. Garg, Regina Barzilay, Tommi Jaakkola
Computer Science and Artificial Intelligence Lab, MIT
{ingraham, vgarg, regina, tommi}@csail.mit.edu

LEARNING FROM PROTEIN STRUCTURE WITH GEOMETRIC VECTOR PERCEPTRONS

Bowen Jing*, Stephan Eismann*, Patricia Suriana, Raphael J.L. Townshend, Ron O. Dror
Stanford University
{bjing, seismann, psuriana, raphael, rondror}@cs.stanford.edu

Learning inverse folding from millions of predicted structures

Chloe Hsu [1] Robert Verkuil [2] Jason Liu [2] Zeming Lin [2 3] Brian Hie [2]
Tom Sercu [2] Adam Lerer [* 2] Alexander Rives [* 2]

Robust deep learning–based protein sequence design using ProteinMPNN

J. Dauparas[1,2], I. Anishchenko[1,2], N. Bennett[1,2,3], H. Bai[1,2,4], R. J. Ragotte[1,2], L. F. Milles[1,2], B. I. M. Wicky[1,2], A. Courbet[1,2,4], R. J. de Haas[5], N. Bethel[1,2,4], P. J. Y. Leung[1,2,3], T. F. Huddy[1,2], S. Pellock[1,2], D. Tischer[1,2], F. Chan[1,2], B. Koepnick[1,2], H. Nguyen[1,2], A. Kang[1,2], B. Sankaran[6], A. K. Bera[1,2], N. P. King[1,2], D. Baker[1,2,4]*

+Backbone / Full *de novo* design

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

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]✉

Illuminating protein space with a programmable generative model

https://doi.org/10.1038/s41586-023-06728-8
Received: 20 December 2022
Accepted: 6 October 2023
Published online: 15 November 2023

John B. Ingraham[1], Max Baranov[1], Zak Costello[1], Karl W. Barber[1], Wujie Wang[1], Ahmed Ismail[1], Vincent Frappier[1], Dana M. Lord[1], Christopher Ng-Thow-Hing[1], Erik R. Van Vlack[1], Shan Tie[1], Vincent Xue[1], Sarah C. Cowles[1], Alan Leung[1], João V. Rodrigues[1], Claudio L. Morales-Perez[1], Alex M. Ayoub[1], Robin Green[1], Katherine Puentes[1], Frank Oplinger[1], Nishant V. Panwar[1], Fritz Obermeyer[1], Adam R. Root[1], Andrew L. Beam[1], Frank J. Poelwijk[1] & Gevorg Grigoryan[1]✉

Protein Structure and Sequence Generation with Equivariant Denoising Diffusion Probabilistic Models

Namrata Anand, Tudor Achim

Fast protein backbone generation with SE(3) flow matching

Jason Yim, Andrew Campbell, Andrew Y. K. Foong, Michael Gastegger, José Jiménez-Luna, Sarah Lewis, Victor Garcia Satorras, Bastiaan S. Veeling, Regina Barzilay, Tommi Jaakkola, Frank Noé

+ many more! Fast moving area.

Illuminating protein space with a programmable generative model

https://doi.org/10.1038/s41586-023-06728-8

Received: 20 December 2022

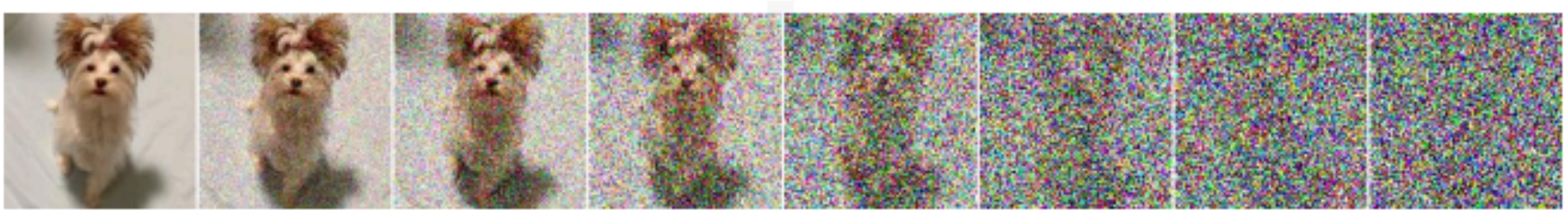
Accepted: 6 October 2023

Published online: 15 November 2023

Open access

John B. Ingraham[1], Max Baranov[1], Zak Costello[1], Karl W. Barber[1], Wujie Wang[1], Ahmed Ismail[1], Vincent Frappier[1], Dana M. Lord[1], Christopher Ng-Thow-Hing[1], Erik R. Van Vlack[1], Shan Tie[1], Vincent Xue[1], Sarah C. Cowles[1], Alan Leung[1], João V. Rodrigues[1], Claudio L. Morales-Perez[1], Alex M. Ayoub[1], Robin Green[1], Katherine Puentes[1], Frank Oplinger[1], Nishant V. Panwar[1], Fritz Obermeyer[1], Adam R. Root[1], Andrew L. Beam[1], Frank J. Poelwijk[1] & Gevorg Grigoryan[1]✉

(as we know) **Diffusion models** handle complex constraints by breaking them into a sequences of smaller changes



Deep Unsupervised Learning using Nonequilibrium Thermodynamics

Jascha Sohl-Dickstein JASCHA@STANFORD.EDU
Stanford University
Eric A. Weiss EAWEISS@BERKELEY.EDU
University of California, Berkeley
Niru Maheswaranathan NIRUM@STANFORD.EDU
Stanford University
Surya Ganguli SGANGULI@STANFORD.EDU
Stanford University

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

SCORE-BASED GENERATIVE MODELING THROUGH STOCHASTIC DIFFERENTIAL EQUATIONS

Yang Song*
Stanford University
yangsong@cs.stanford.edu

Jascha Sohl-Dickstein
Google Brain
jaschasd@google.com

Diederik P. Kingma
Google Brain
durk@google.com

Abhishek Kumar
Google Brain
abhishk@google.com

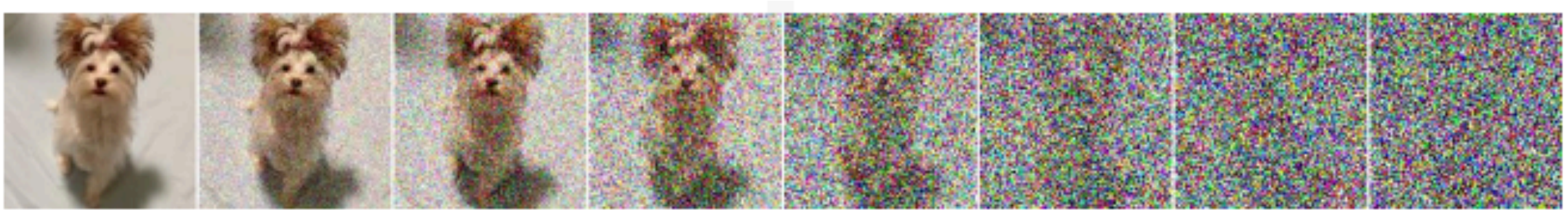
Stefano Ermon
Stanford University
ermon@cs.stanford.edu

Ben Poole
Google Brain
pooleb@google.com

* equivalently, score function, flow predictor, etc…

(as we know) **Diffusion models** handle complex constraints by breaking them into a sequences of smaller changes

1. Define a forwards diffusion process



Deep Unsupervised Learning using Nonequilibrium Thermodynamics

Jascha Sohl-Dickstein JASCHA@STANFORD.EDU
Stanford University
Eric A. Weiss EAWEISS@BERKELEY.EDU
University of California, Berkeley
Niru Maheswaranathan NIRUM@STANFORD.EDU
Stanford University
Surya Ganguli SGANGULI@STANFORD.EDU
Stanford University

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

SCORE-BASED GENERATIVE MODELING THROUGH STOCHASTIC DIFFERENTIAL EQUATIONS

Yang Song*
Stanford University
yangsong@cs.stanford.edu

Jascha Sohl-Dickstein
Google Brain
jaschasd@google.com

Diederik P. Kingma
Google Brain
durk@google.com

Abhishek Kumar
Google Brain
abhishk@google.com

Stefano Ermon
Stanford University
ermon@cs.stanford.edu

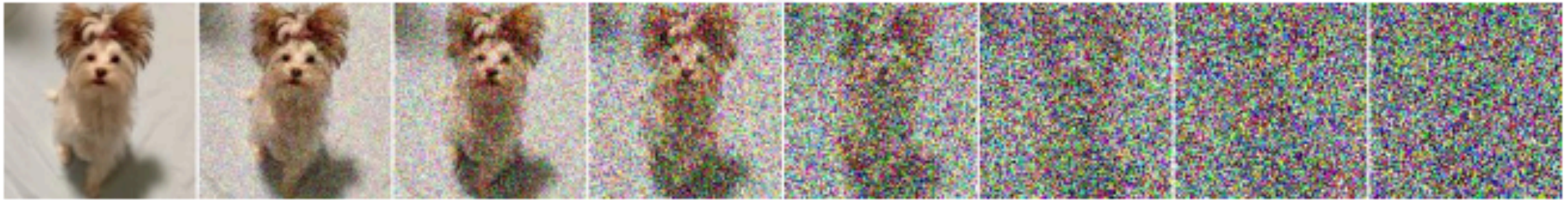
Ben Poole
Google Brain
pooleb@google.com

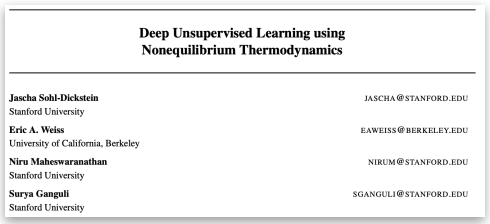
* equivalently, score function, flow predictor, etc…

(as we know) **Diffusion models** handle complex constraints by breaking them into a sequences of smaller changes

1. Define a forwards diffusion process

2. Train a neural network denoiser*





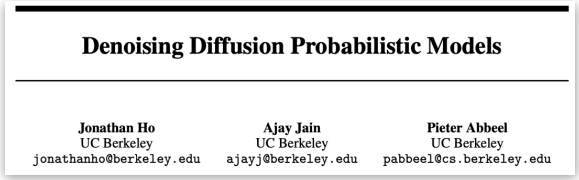
Deep Unsupervised Learning using Nonequilibrium Thermodynamics

Jascha Sohl-Dickstein JASCHA@STANFORD.EDU
Stanford University

Eric A. Weiss EAWEISS@BERKELEY.EDU
University of California, Berkeley

Niru Maheswaranathan NIRUM@STANFORD.EDU
Stanford University

Surya Ganguli SGANGULI@STANFORD.EDU
Stanford University

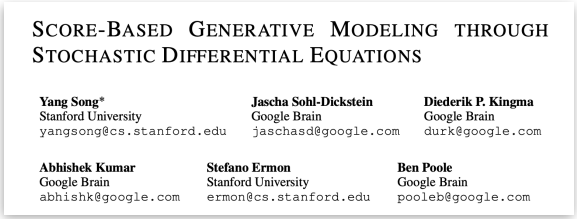


Denoising Diffusion Probabilistic Models

Jonathan Ho
UC Berkeley
jonathanho@berkeley.edu

Ajay Jain
UC Berkeley
ajayj@berkeley.edu

Pieter Abbeel
UC Berkeley
pabbeel@cs.berkeley.edu



SCORE-BASED GENERATIVE MODELING THROUGH STOCHASTIC DIFFERENTIAL EQUATIONS

Yang Song*
Stanford University
yangsong@cs.stanford.edu

Jascha Sohl-Dickstein
Google Brain
jaschasd@google.com

Diederik P. Kingma
Google Brain
durk@google.com

Abhishek Kumar
Google Brain
abhishk@google.com

Stefano Ermon
Stanford University
ermon@cs.stanford.edu

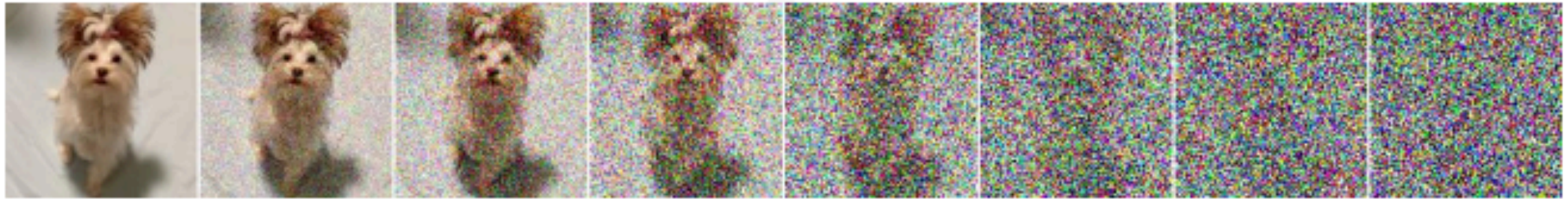
Ben Poole
Google Brain
pooleb@google.com

* equivalently, score function, flow predictor, etc…

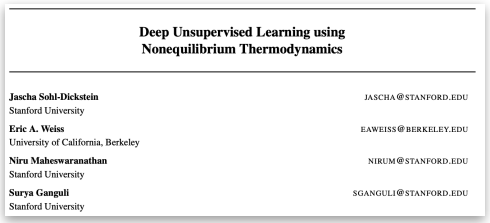
(as we know) **Diffusion models** handle complex constraints by breaking them into a sequences of smaller changes

1. Define a forwards diffusion process

2. Train a neural network denoiser*

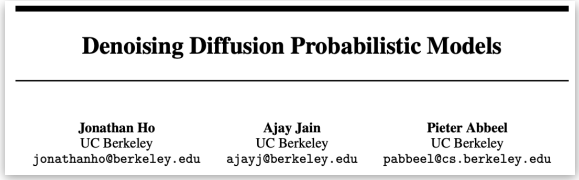


3. Reverse the diffusion process with our learned model



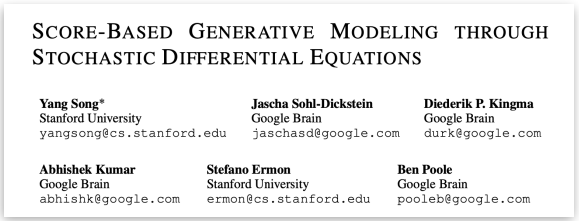
Deep Unsupervised Learning using Nonequilibrium Thermodynamics

Jascha Sohl-Dickstein — Stanford University — JASCHA@STANFORD.EDU
Eric A. Weiss — University of California, Berkeley — EAWEISS@BERKELEY.EDU
Niru Maheswaranathan — Stanford University — NIRUM@STANFORD.EDU
Surya Ganguli — Stanford University — SGANGULI@STANFORD.EDU



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



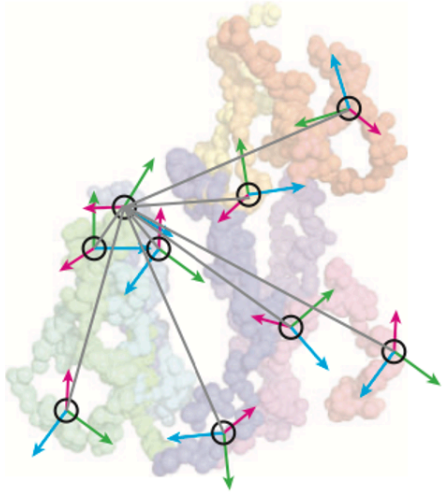
SCORE-BASED GENERATIVE MODELING THROUGH STOCHASTIC DIFFERENTIAL EQUATIONS

Yang Song* — Stanford University — yangsong@cs.stanford.edu
Jascha Sohl-Dickstein — Google Brain — jaschasd@google.com
Diederik P. Kingma — Google Brain — durk@google.com
Abhishek Kumar — Google Brain — abhishk@google.com
Stefano Ermon — Stanford University — ermon@cs.stanford.edu
Ben Poole — Google Brain — pooleb@google.com

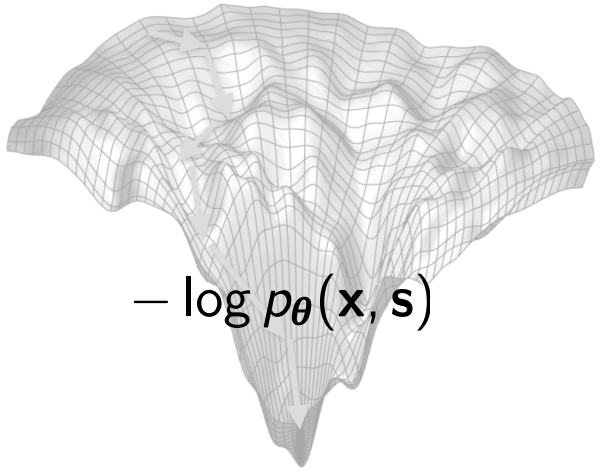
* equivalently, score function, flow predictor, etc…

Three properties we find important for a protein generative modeling framework

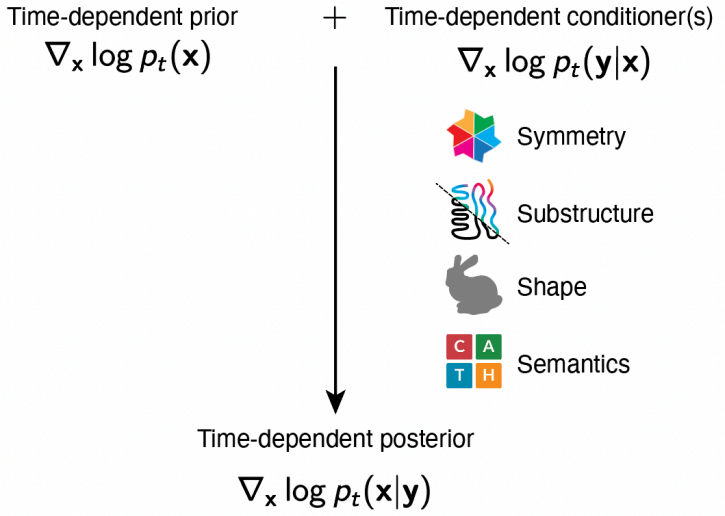
Subquadratic architecture



Likelihood that is useful and *controllable*

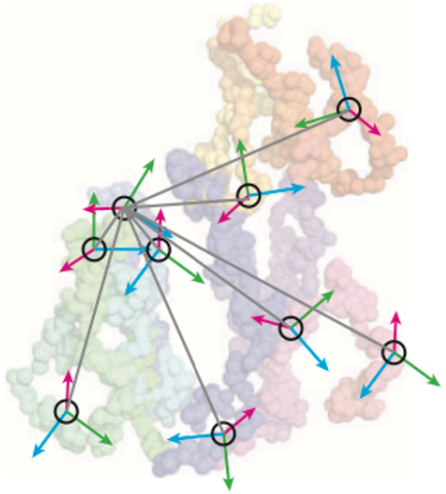


Programmable generation



Three properties we find important for a protein generative model

Subquadratic architecture



Likelihood that is useful and *controllable*

Programmable generation

Should structural ML really need quadratic cost?

Structural ML
(typical)

$O(N^2)$
$O(N^3)$

Scales with number of interactions

Image diffusion models

$O(N)$

Scales with number of pixels

Should structural ML really need quadratic cost?

Structural ML
(typical)

$O(N^2)$
$O(N^3)$

Scales with number of interactions

Image diffusion models

$O(N)$

Scales with number of pixels

But wait, proteins need interactions!! What about long-range context?

Should structural ML really need quadratic cost?

Structural ML
(typical)

$O(N^2)$
$O(N^3)$

Scales with number of interactions

Image diffusion models

$O(N)$

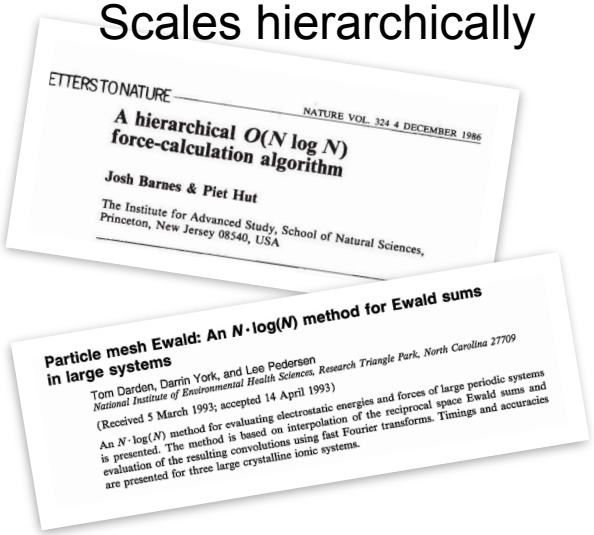
Scales with number of pixels

But wait, proteins need interactions!! What about long-range context?

Computational physics

$O(N \log N)$

Scales hierarchically



ETTERS TO NATURE NATURE VOL. 324 4 DECEMBER 1986

A hierarchical $O(N \log N)$ force-calculation algorithm

Josh Barnes & Piet Hut

The Institute for Advanced Study, School of Natural Sciences, Princeton, New Jersey 08540, USA

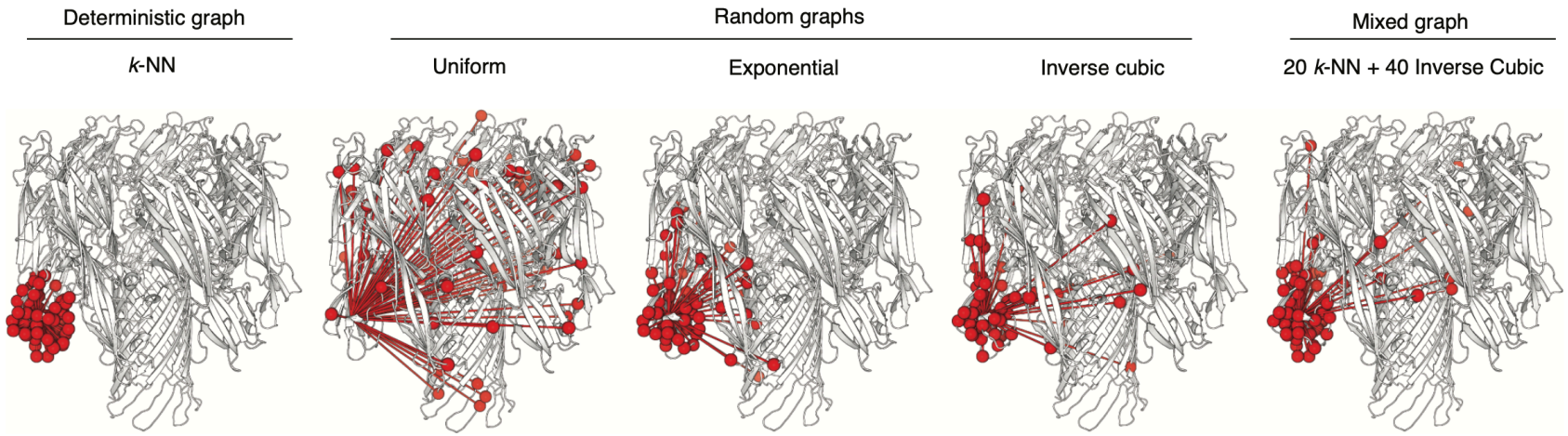
Particle mesh Ewald: An $N \cdot \log(N)$ method for Ewald sums in large systems

Tom Darden, Darrin York, and Lee Pedersen
National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina 27709

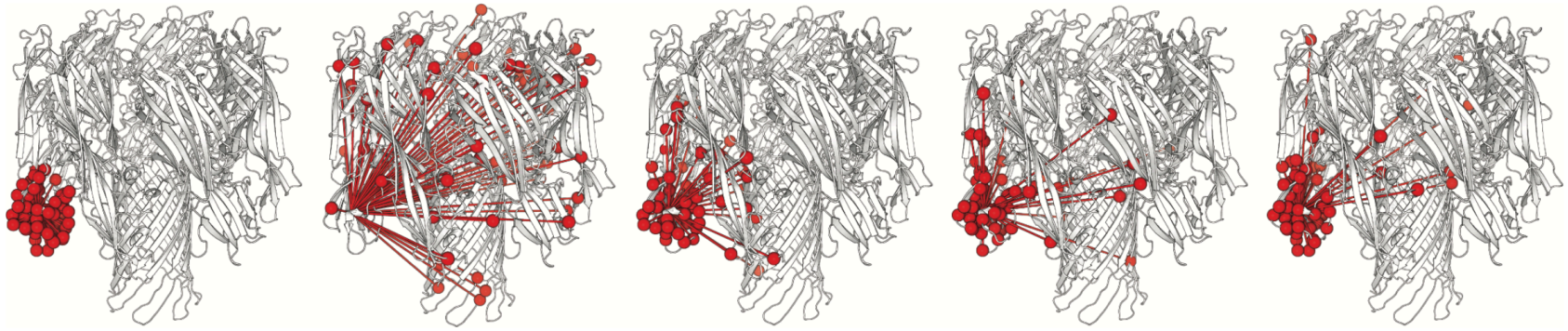
(Received 5 March 1993; accepted 14 April 1993)

An $N \cdot \log(N)$ method for evaluating electrostatic energies and forces of large periodic systems is presented. The method is based on interpolation of the reciprocal space Ewald sums and evaluation of the resulting convolutions using fast Fourier transforms. Timings and accuracies are presented for three large crystalline ionic systems.

Architecture trick 1: **Sparse long range connections** are subsquadratic while preserving long range context

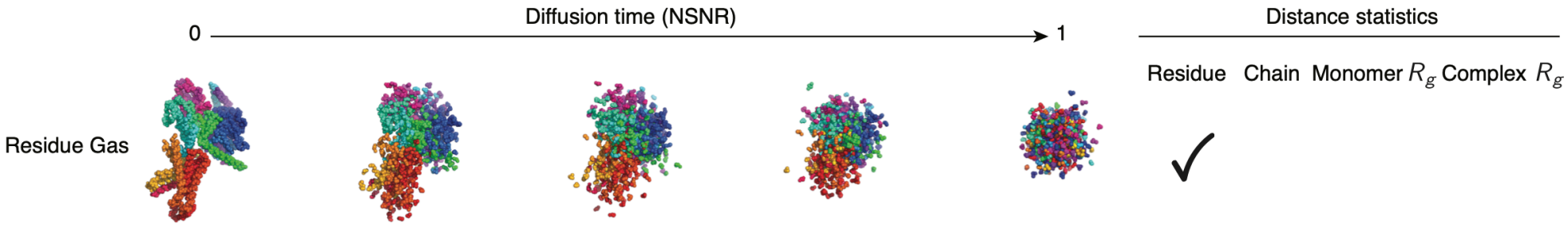


Architecture trick 1: **Sparse long range connections** are subsquadratic while preserving long range context

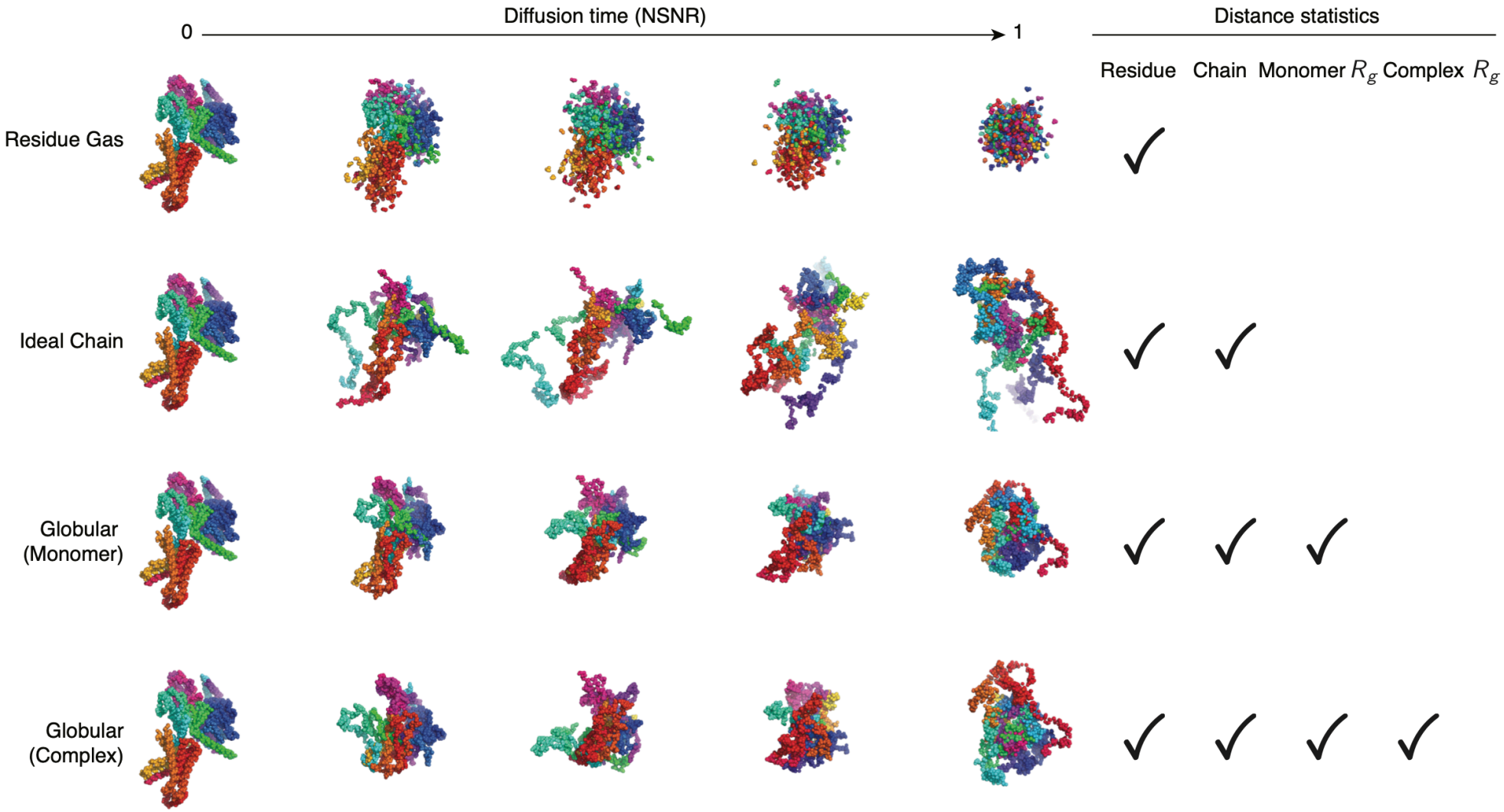


Edge propensity $p(\mathcal{E}_{ij} \in \mathcal{G}(\mathbf{x}) \mid D_{ij}(\mathbf{x})) \propto$	constant	$\exp\left(-\frac{D_{ij}(\mathbf{x})}{l}\right)$	$\frac{1}{D_{ij}(\mathbf{x})^3}$	
Marginal distance propensity (uniform grid) $p(D_{ij}(\mathbf{x}) \mid \mathcal{E}_{ij} \in \mathcal{G}(\mathbf{x})) \propto$	$D_{ij}(\mathbf{x})^2$	$D_{ij}(\mathbf{x})^2 \exp\left(-\frac{D_{ij}(\mathbf{x})}{l}\right)$	$\frac{1}{D_{ij}(\mathbf{x})}$	
Long-range attachment	✓	✓	✓	✓
Non-vanishing local attachment		✓	✓	✓
Monotonic decreasing distance propensity			✓	✓

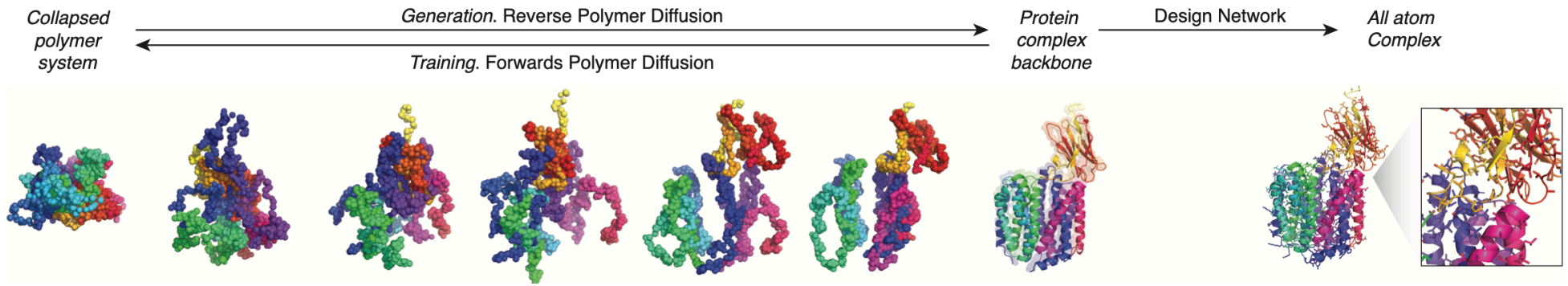
Architecture trick 2: **distance-preserving diffusions** avoid breaking polymer structure



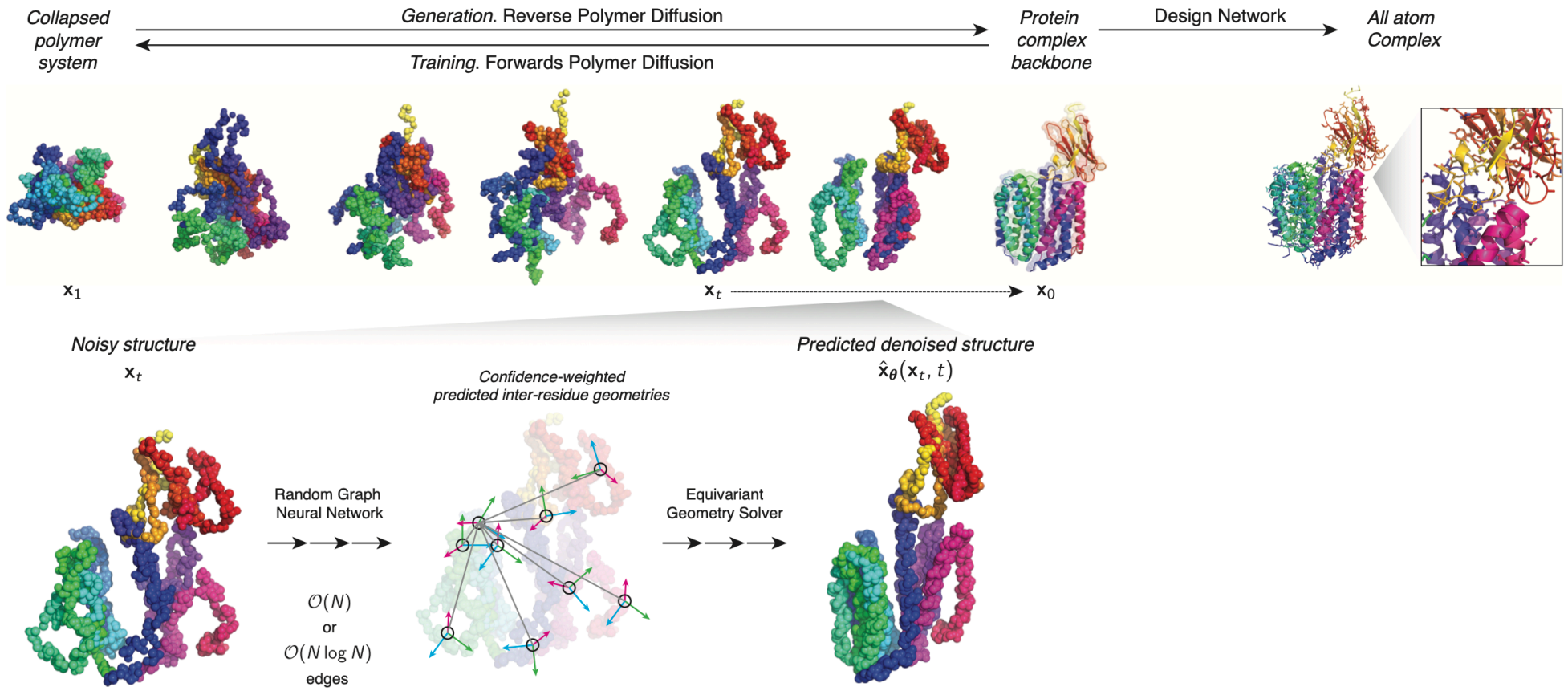
Architecture trick 2: **distance-preserving diffusions** avoid breaking polymer structure



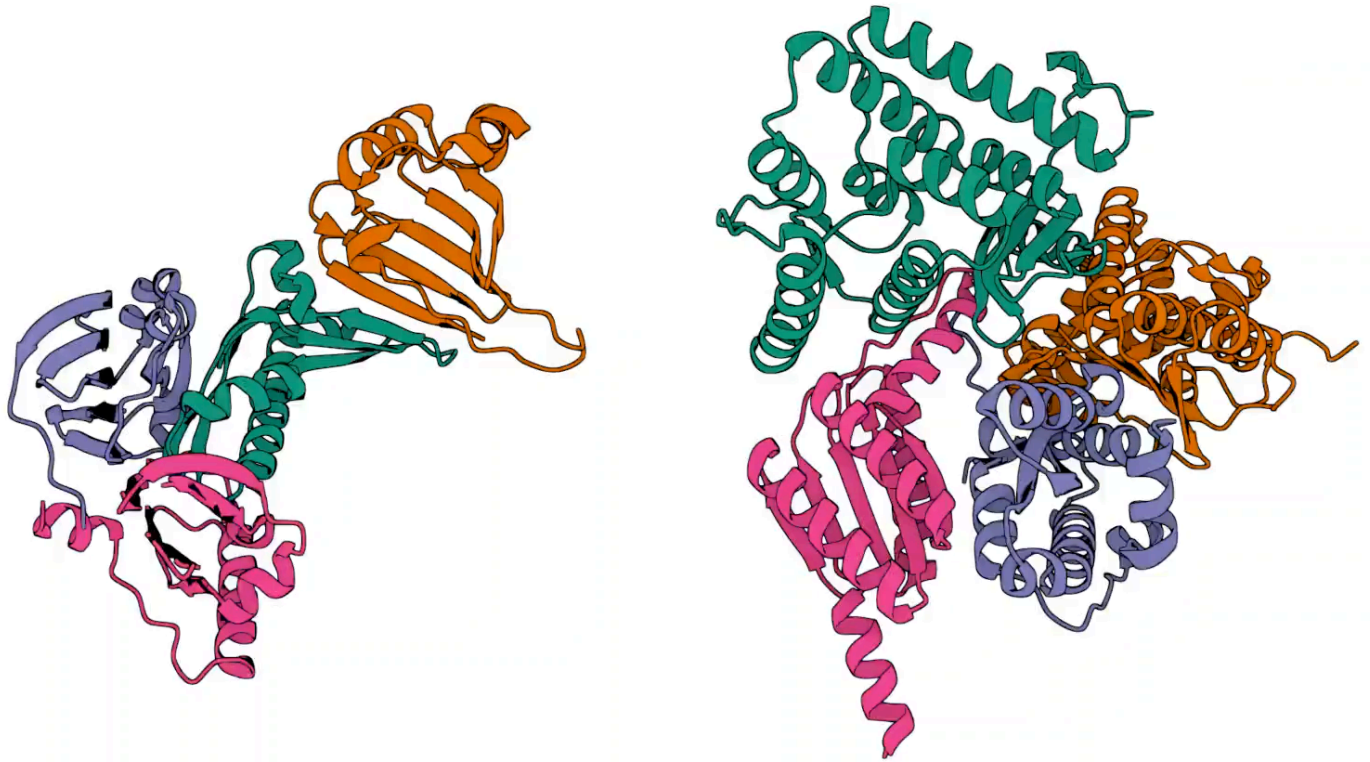
Putting together with some other things, a fast polymer diffusion model



Putting together with some other things, a fast polymer diffusion model

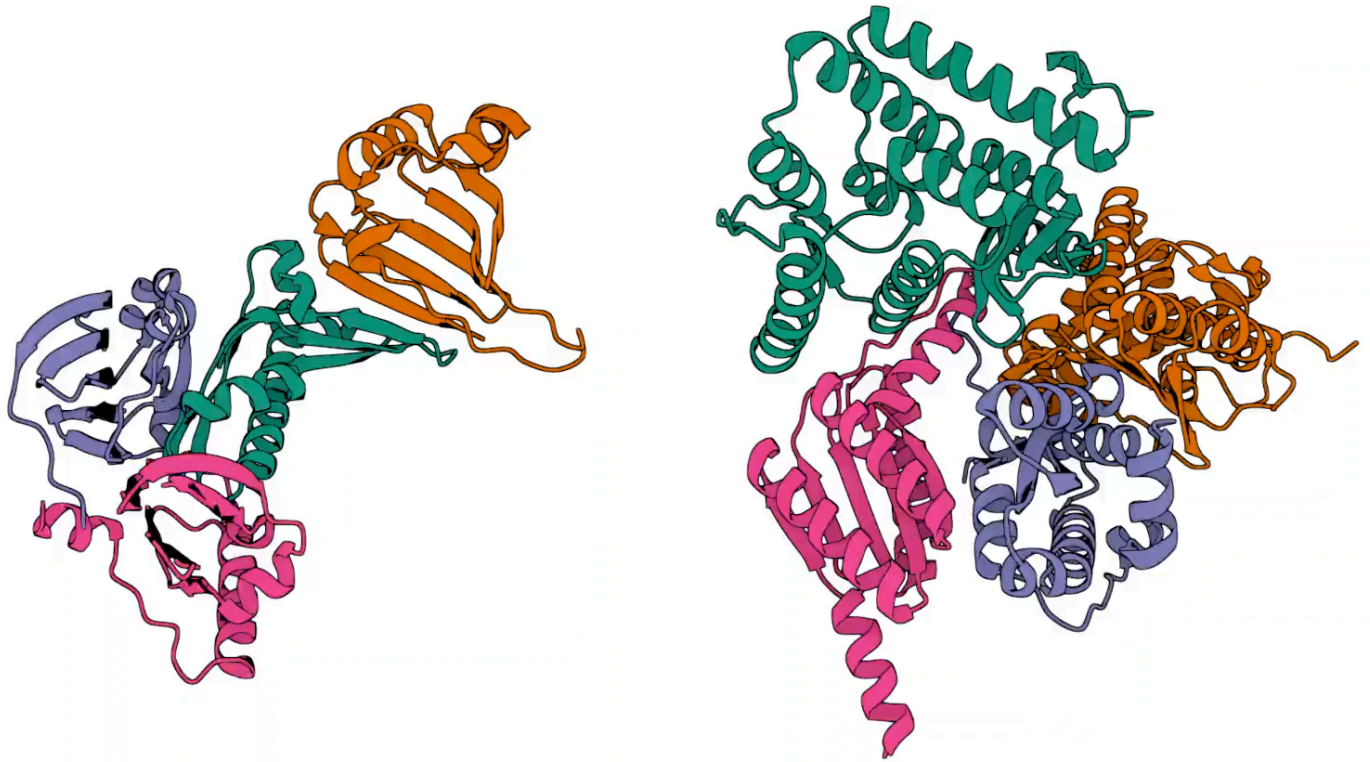


Punchline: Our system can generate **large complexes on 1 GPU in about a minute**



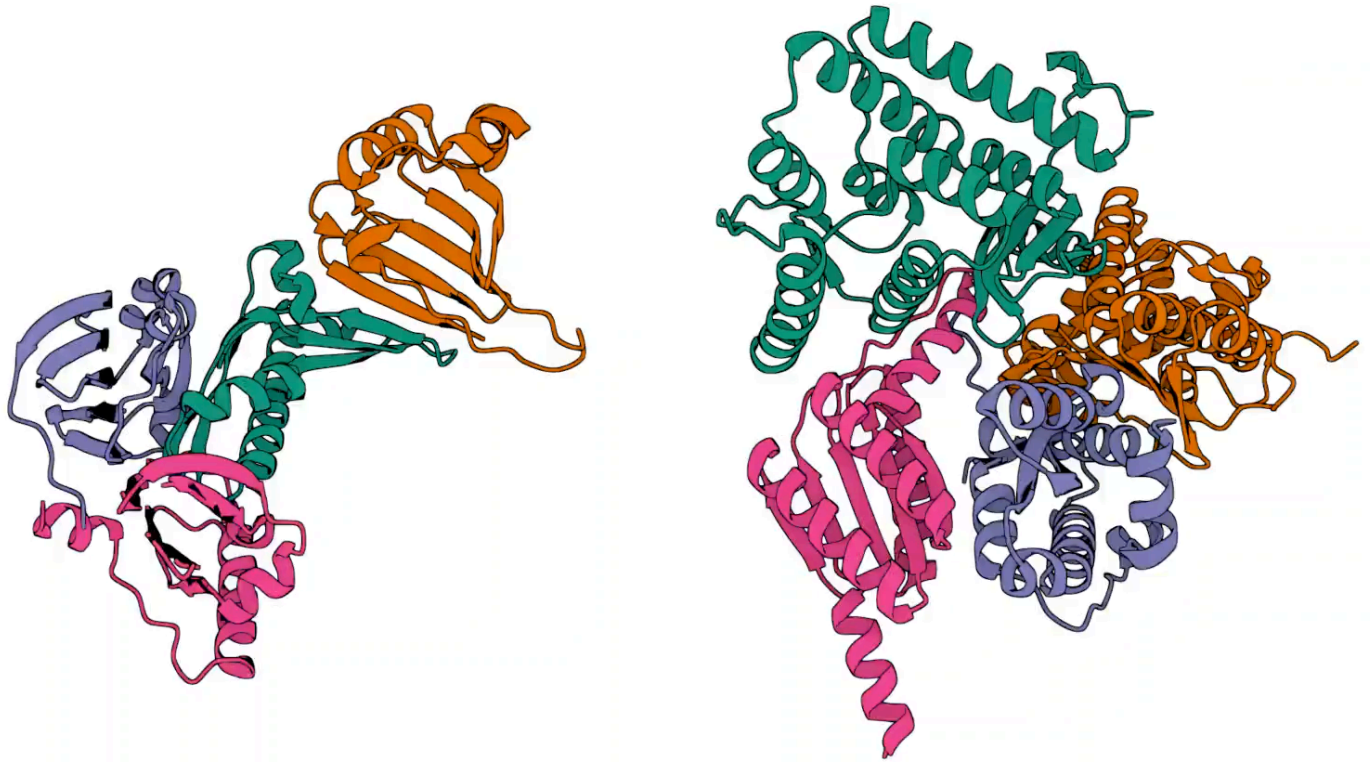
In principle this could keep scaling

Punchline: Our system can generate **large complexes on 1 GPU in about a minute**



In principle this could keep scaling

Punchline: Our system can generate **large complexes on 1 GPU in about a minute**

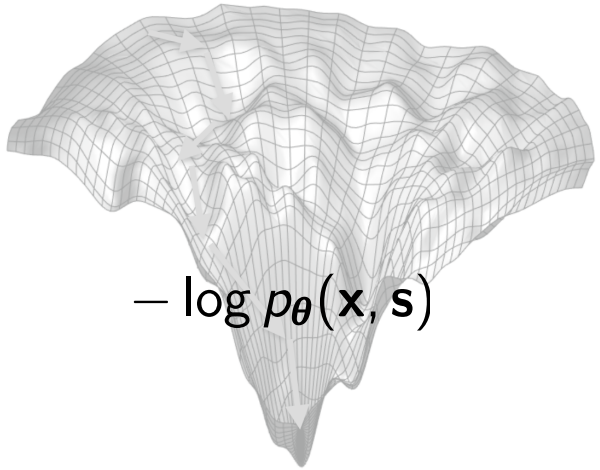


In principle this could keep scaling

Three properties we find important for a protein generative model

Subquadratic architecture

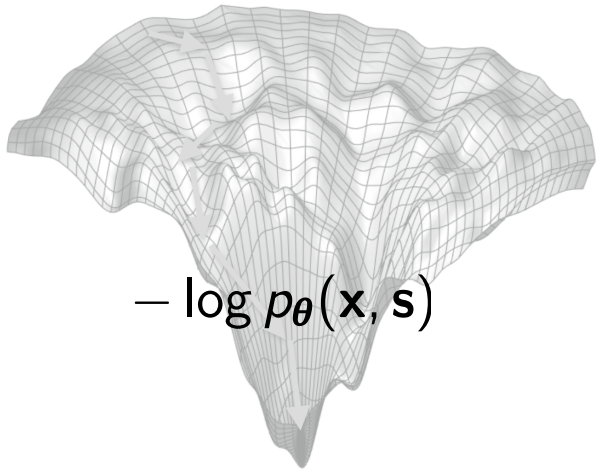
Likelihood that is useful and *controllable*



Programmable generation

Likelihood (aka free energy) has been a north star in protein biophysics and design

Energy landscape

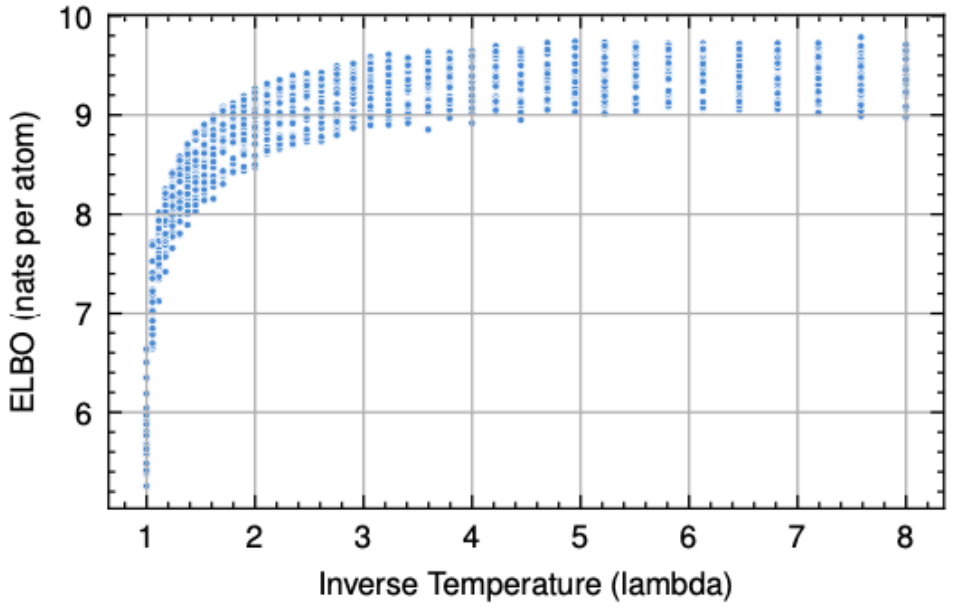


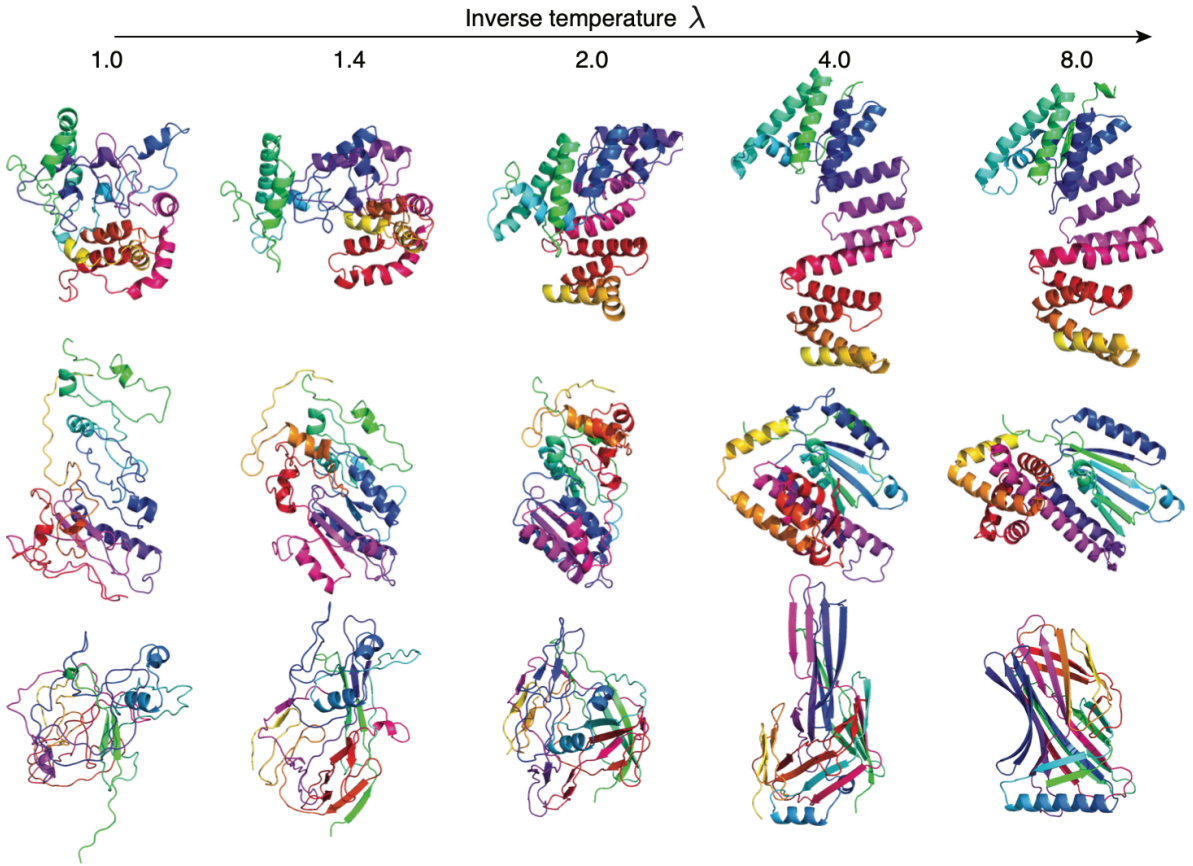
Energy answers questions like:

- What state(s) will proteins fold up into and what competes?
- How stable is your molecule?
- How specific is your binder?
- What mutations are possible?

We can evaluate likelihood and also control it with a new low temperature sampling algorithm

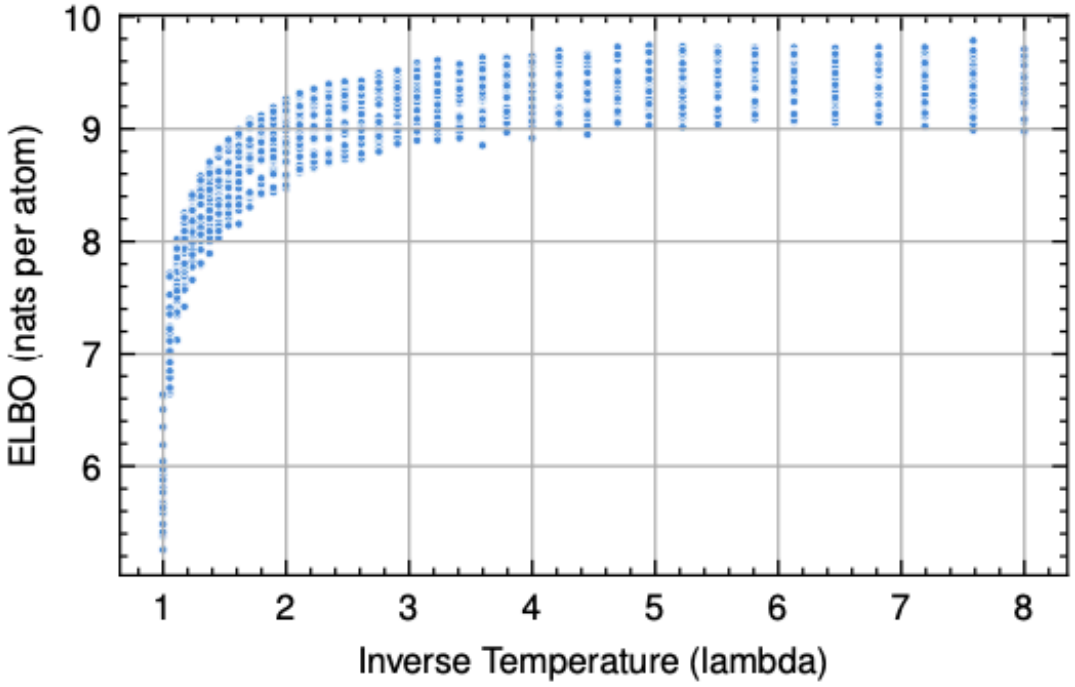
Increases likelihood



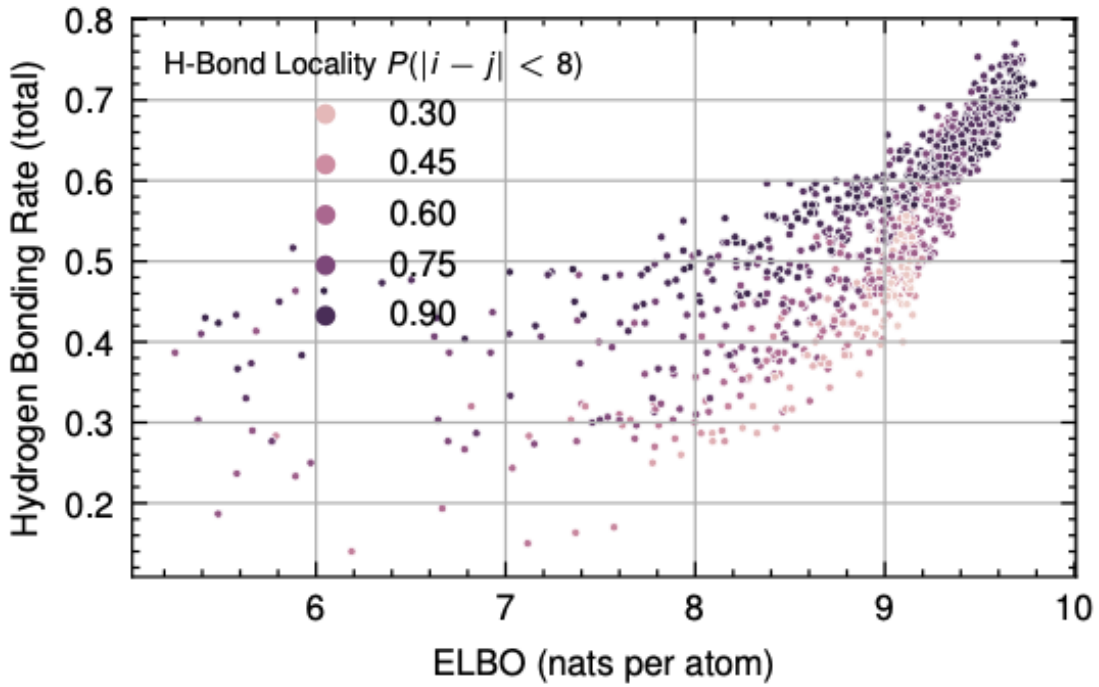


Low temperature drives towards high likelihood and favorable structural properties

Higher likelihood



More hydrogen bonds

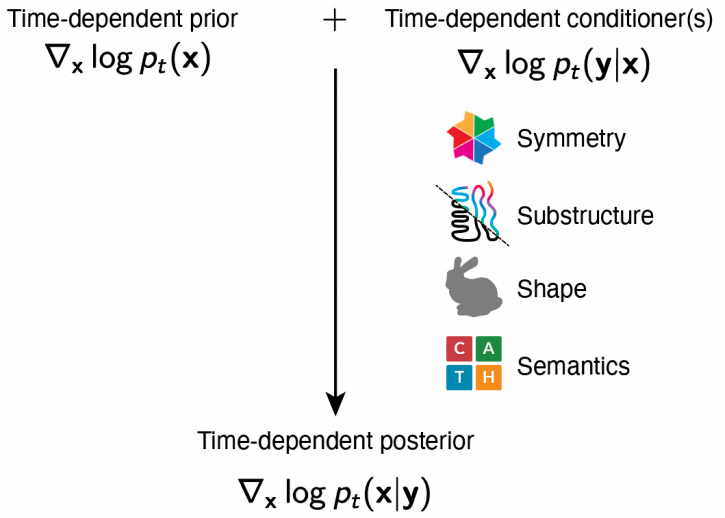


Three properties we find important for a protein generative model

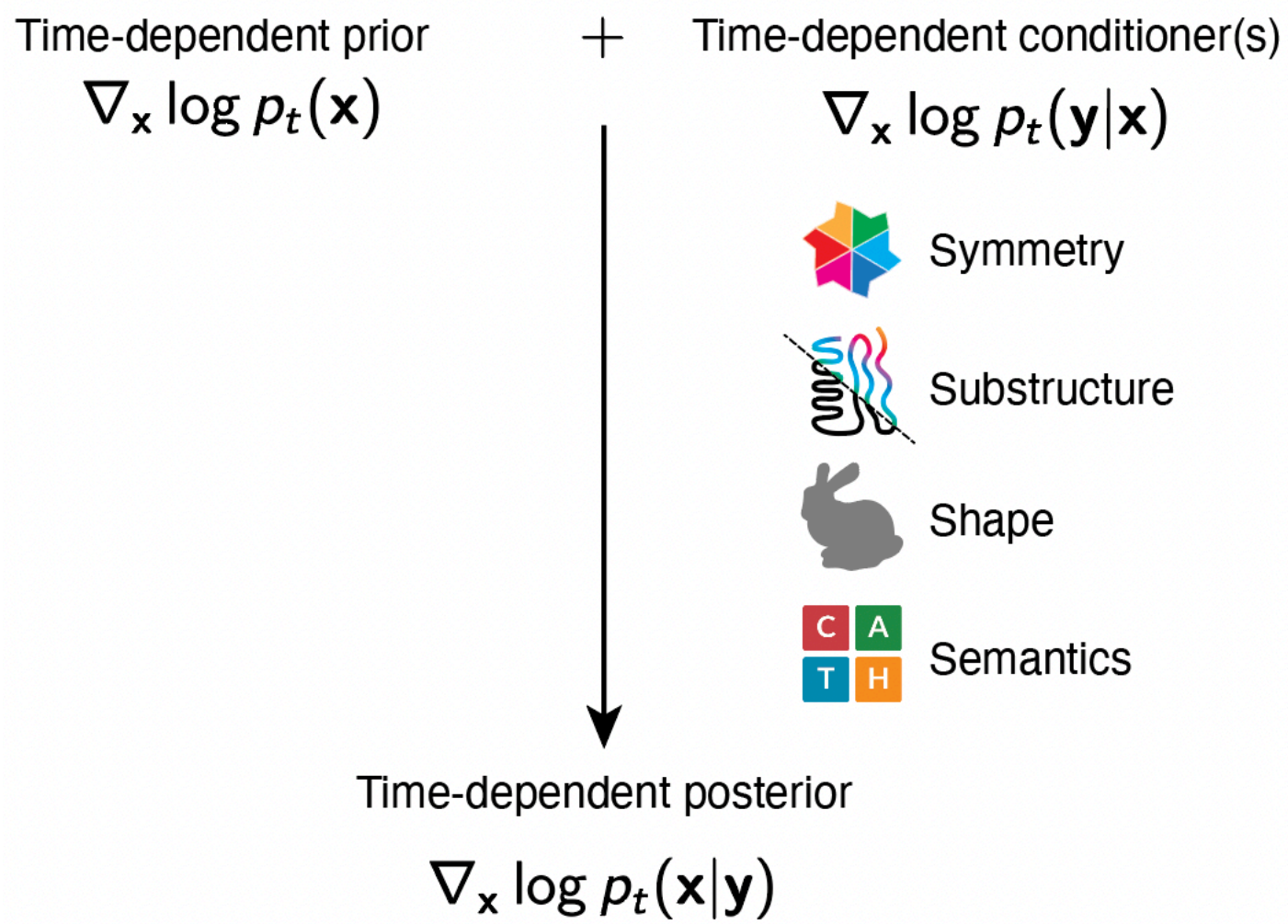
Subquadratic architecture

Likelihood that is useful and *controllable*

Programmable generation



We are most excited about **programmability**



We have built a “zoo” of conditioner models, which can be mixed and match to solve for functional proteins



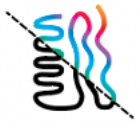
Symmetry

Cyclic, Dihedral, Icosahedral…
(coming soon)
Crystal, helices

Nanomaterials

Vaccines

Vectors



Substructure

Motif grafting,
Protein infilling
Outfilling

de novo binders

Enzyme miniaturization

Copy-paste protein engineering



Shape

Arbitrary 3D shapes

Pores,
Interaction geometry control

Cryo EM



Semantics

ProClass:
Secondary structure
CATH

ProCap: Text

Computer, give me a protein now

Diffusion conditioning: How can we make it modular, composable, extensible?

Requirements We propose four desiderata for a *programmable protein design* framework:

- **Compositionality.** Problems are expressed as *design programs* which are composed from "building blocks" encoding different required attributes.
- **Restraints.** Building blocks should be able to express soft restraints (e.g. classifier guidance) as a special case.
- **Constraints.** Building blocks should be able to express hard constraints, such as manifold constraints, as a special case.
- **Automatic sampling.** It should be feasible to automatically synthesize a valid sampling algorithm for any *design program* without requiring additional logic to be implemented by the user.

Key idea: diffusion conditioning as **state and energy transformations**

State transform
(e.g. change of variables)

Energy adjustment
(e.g. classifier guidance)

$$U(\tilde{\mathbf{x}}_t; U_f, f, t) = \underbrace{\frac{1}{2}\left\|\sigma_t^{-1}\mathbf{R}^{-1}\left(f(\tilde{\mathbf{x}}_t, U_0; t) - \alpha_t \hat{\mathbf{x}}_t\left(f(\tilde{\mathbf{x}}_t, U_0; t), t\right)\right)\right\|_2^2}_{\text{Diffusion Energy}} + \underbrace{U_f(\tilde{\mathbf{x}}_t, U_0; t)}_{\text{Conditioner Energy}}$$

MCMC-enriched reverse diffusion

When state transform is identity, energy adjustment is zero, and pink terms are 1,

This is standard reverse diffusion

$$\mathrm{d}\mathbf{x} = h_t\,\mathbf{x} - g_t^2\left(\lambda_t + \frac{\lambda_0\psi}{2}\right)\mathbf{R}\mathbf{R}^\intercal\nabla_{\mathbf{x}_t}\log p_t(\mathbf{x}_t)\,\mathrm{d}t + g_t\sqrt{(1+\psi)}\,\mathbf{R}\,\mathrm{d}\bar{\mathbf{w}}.$$

Some related ideas in

Reduce, Reuse, Recycle: Compositional Generation with Energy-Based Diffusion Models and MCMC

Yilun Du [1] Conor Durkan [2] Robin Strudel [2] Joshua B. Tenenbaum [1] Sander Dieleman [2] Rob Fergus [2] Jascha Sohl-Dickstein [2] Arnaud Doucet [2] Will Grathwohl [2]

DREAMFUSION: TEXT-TO-3D USING 2D DIFFUSION

Ben Poole[1], Ajay Jain[2], Jonathan T. Barron[1], Ben Mildenhall[1]
[1]Google Research, [2]UC Berkeley
{pooleb, barron, bmild}@google.com, ajayj@berkeley.edu

All of our conditioners as state and energy transformations

$$U(\tilde{\mathbf{x}}_t; U_f, f, t) = \underbrace{\frac{1}{2}\left\|\sigma_t^{-1}\mathbf{R}^{-1}\left(f(\tilde{\mathbf{x}}_t, U_0; t) - \alpha_t \hat{\mathbf{x}}_t\left(f(\tilde{\mathbf{x}}_t, U_0; t), t\right)\right)\right\|_2^2}_{\text{Diffusion Energy}} + \underbrace{U_f(\tilde{\mathbf{x}}_t, U_0; t)}_{\text{Conditioner Energy}}$$

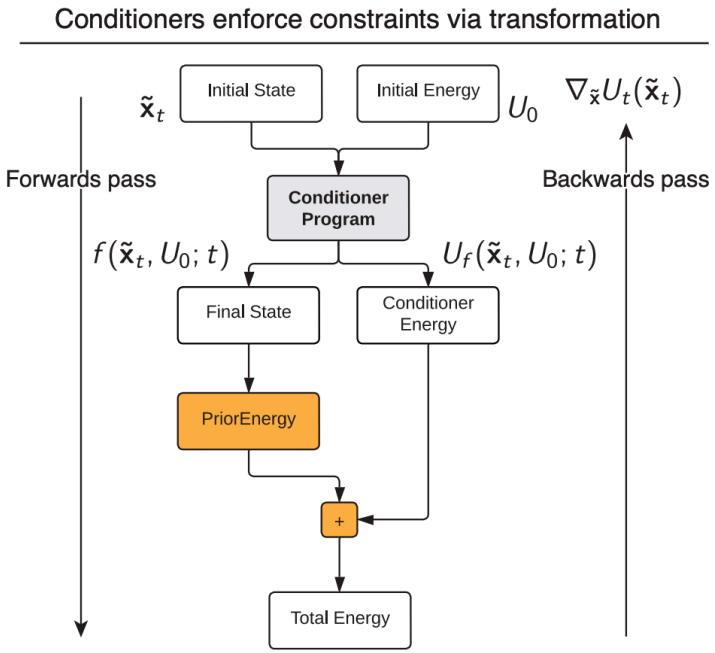
Conditioner	$f(\tilde{\mathbf{x}}, U, t)$	$U_f(\tilde{\mathbf{x}}, U, t)$	**Examples and applications**
Likelihood restraint	$\tilde{\mathbf{x}}$	$U - \log p_t(\cdot\|\tilde{\mathbf{x}})$	Biasing towards specifications
Linear constraint	$\mathbf{A}\tilde{\mathbf{x}} + \mathbf{b}$	U	Exactly enforcing specifications
Nonlinear constraint	$f(\tilde{\mathbf{x}})$	$U + \log\det \frac{\mathrm{d}f}{\mathrm{d}\tilde{\mathbf{x}}}$	Exactly enforcing specifications

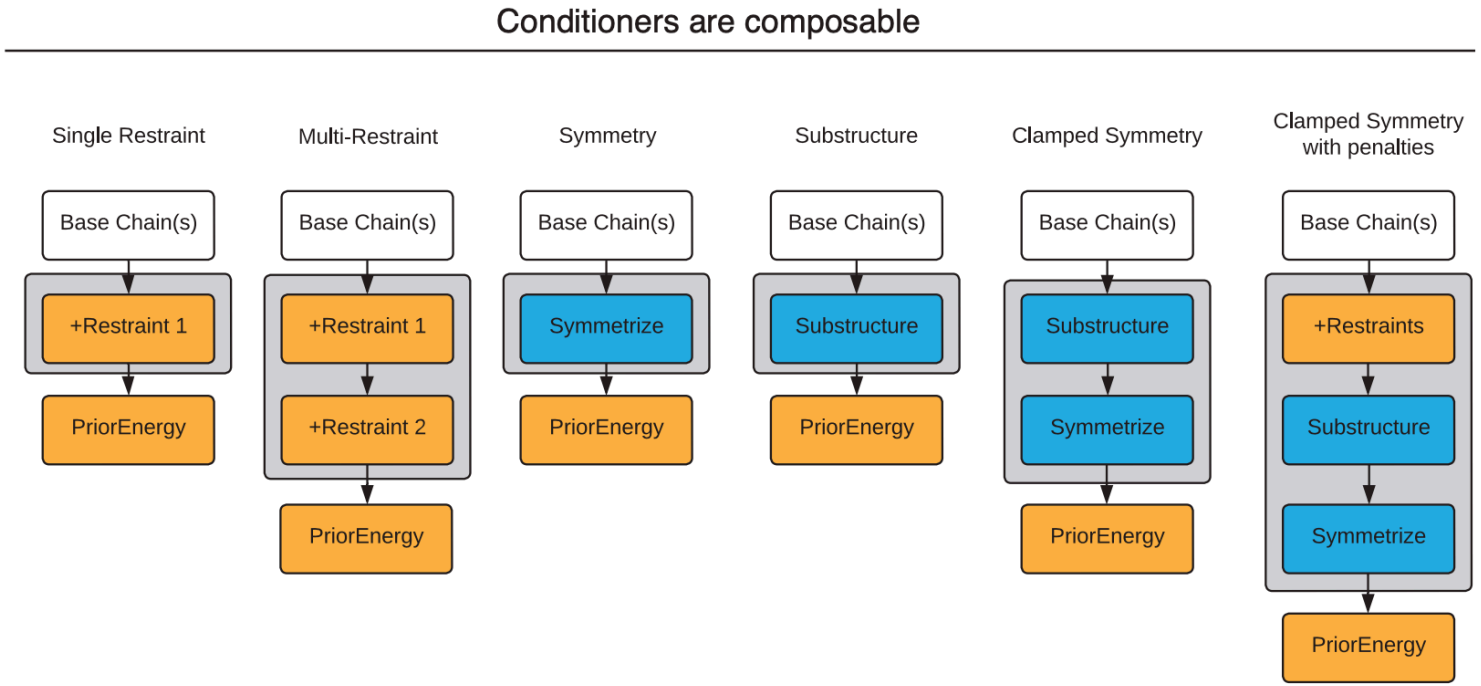
All of our conditioners as state and energy transformations

$$U(\tilde{\mathbf{x}}_t; U_f, f, t) = \underbrace{\frac{1}{2}\left\|\sigma_t^{-1}\mathbf{R}^{-1}\left(f(\tilde{\mathbf{x}}_t, U_0; t) - \alpha_t \hat{\mathbf{x}}_t\left(f(\tilde{\mathbf{x}}_t, U_0; t), t\right)\right)\right\|_2^2}_{\text{Diffusion Energy}} + \underbrace{U_f(\tilde{\mathbf{x}}_t, U_0; t)}_{\text{Conditioner Energy}}$$

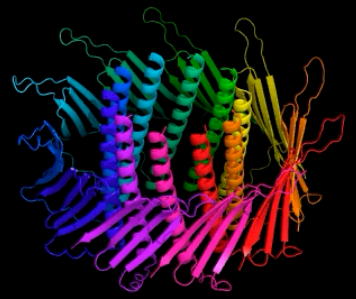
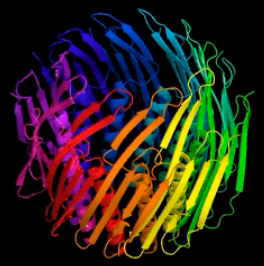
Conditioner	$f(\tilde{\mathbf{x}}, U, t)$	$U_f(\tilde{\mathbf{x}}, U, t)$	**Examples and applications**
Symmetry constraint	$\mathbf{G}\tilde{\mathbf{x}}$	U	Large assemblies
Substructure constraint	$\bar{\mathbf{R}}\mathbf{R}^{-1}\tilde{\mathbf{x}} + \bar{\mu}$	$U + \|\hat{\mathbf{x}}_\theta(f(\tilde{\mathbf{x}}, U, t), t)^{\mathcal{M}} - \mathbf{x}_t^{\mathcal{M}}\|_2^2$	Substructure grafting
Substructure distances	$\tilde{\mathbf{x}}$	$U - \log p_t(d_{ij}\|\tilde{\mathbf{x}})$	Interface and contact constraints
Substructure motif	$\tilde{\mathbf{x}}$	$U + \eta \log\left(1 + e^{\zeta[\rho(\mathbf{x}_t) - \rho_{max}]}\right)$	Motif-conditioned scaffolds
Shape constraint	$\tilde{\mathbf{x}}$	$U + \text{ShapeLoss}_t(\mathbf{x}, \mathbf{r})$	Molecular shape control
Sequence	$\tilde{\mathbf{x}}$	$U - \log p_t(\text{sequence}\|\tilde{\mathbf{x}})$	Sequence constraints
Secondary structure	$\tilde{\mathbf{x}}$	$U - \log p_t(\text{ss}\|\tilde{\mathbf{x}})$	Topological constraints
Domain classification	$\tilde{\mathbf{x}}$	$U - \log p_t(\text{domain}\|\tilde{\mathbf{x}})$	Pfam, CATH, Taxonomy
Text caption	$\tilde{\mathbf{x}}$	$U - \log p_t(\text{caption}\|\tilde{\mathbf{x}})$	Natural language prompting
Likelihood restraint	$\tilde{\mathbf{x}}$	$U - \log p_t(\cdot\|\tilde{\mathbf{x}})$	Biasing towards specifications
Linear constraint	$\mathbf{A}\tilde{\mathbf{x}} + \mathbf{b}$	U	Exactly enforcing specifications
Nonlinear constraint	$f(\tilde{\mathbf{x}})$	$U + \log\det\frac{\mathrm{d}f}{\mathrm{d}\tilde{\mathbf{x}}}$	Exactly enforcing specifications

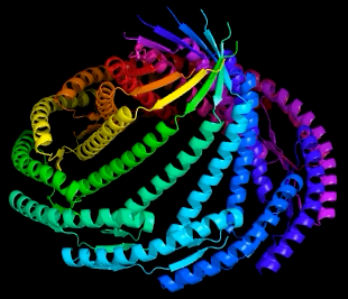
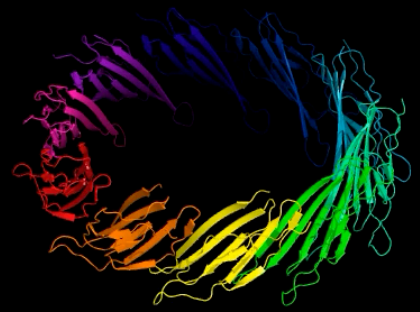
Because input-output types are matched, Conditioners are **composable**

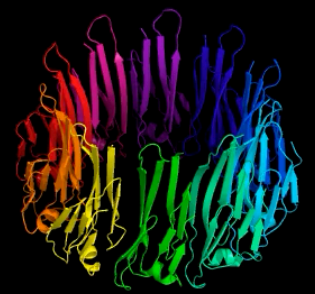




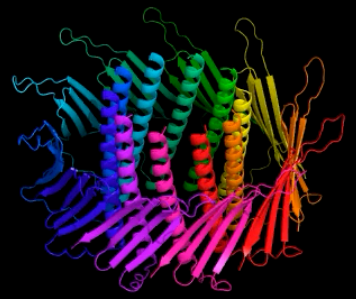
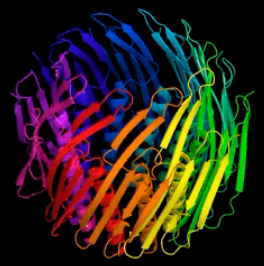
Programmable Symmetry in Complexes

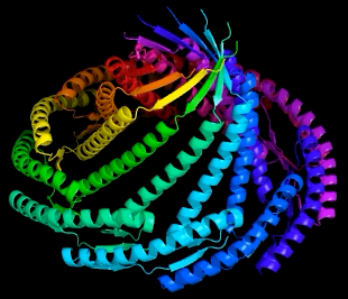
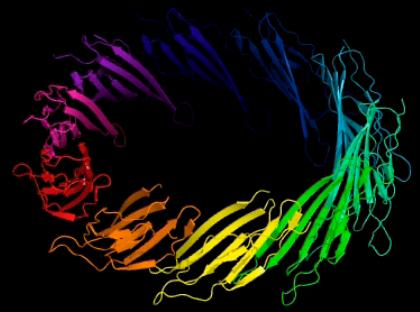




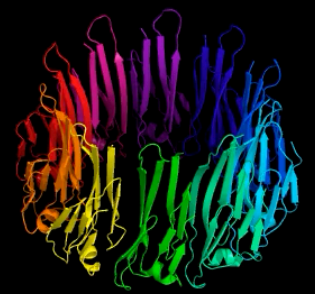





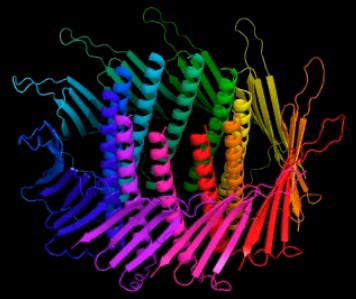
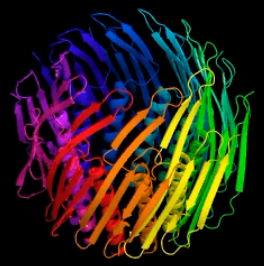
Programmable Symmetry in Complexes

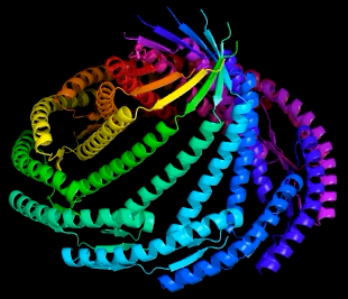
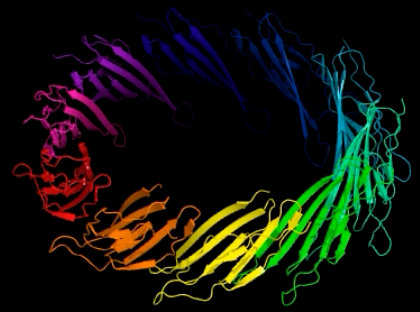




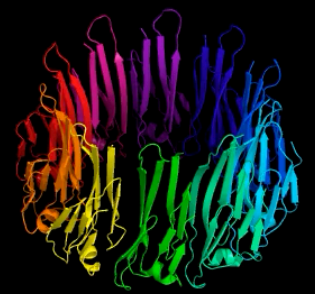





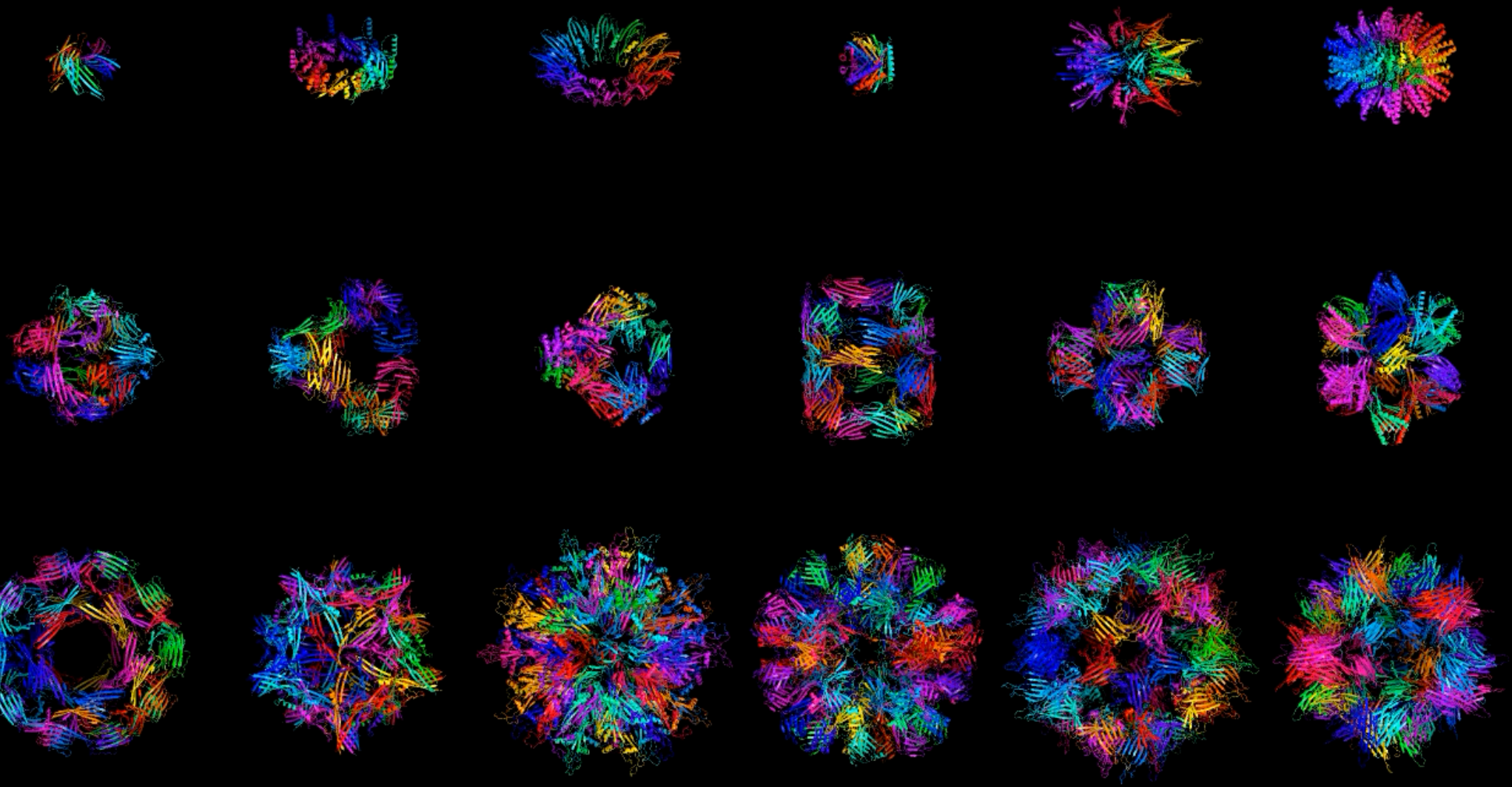
Programmable Symmetry in Complexes

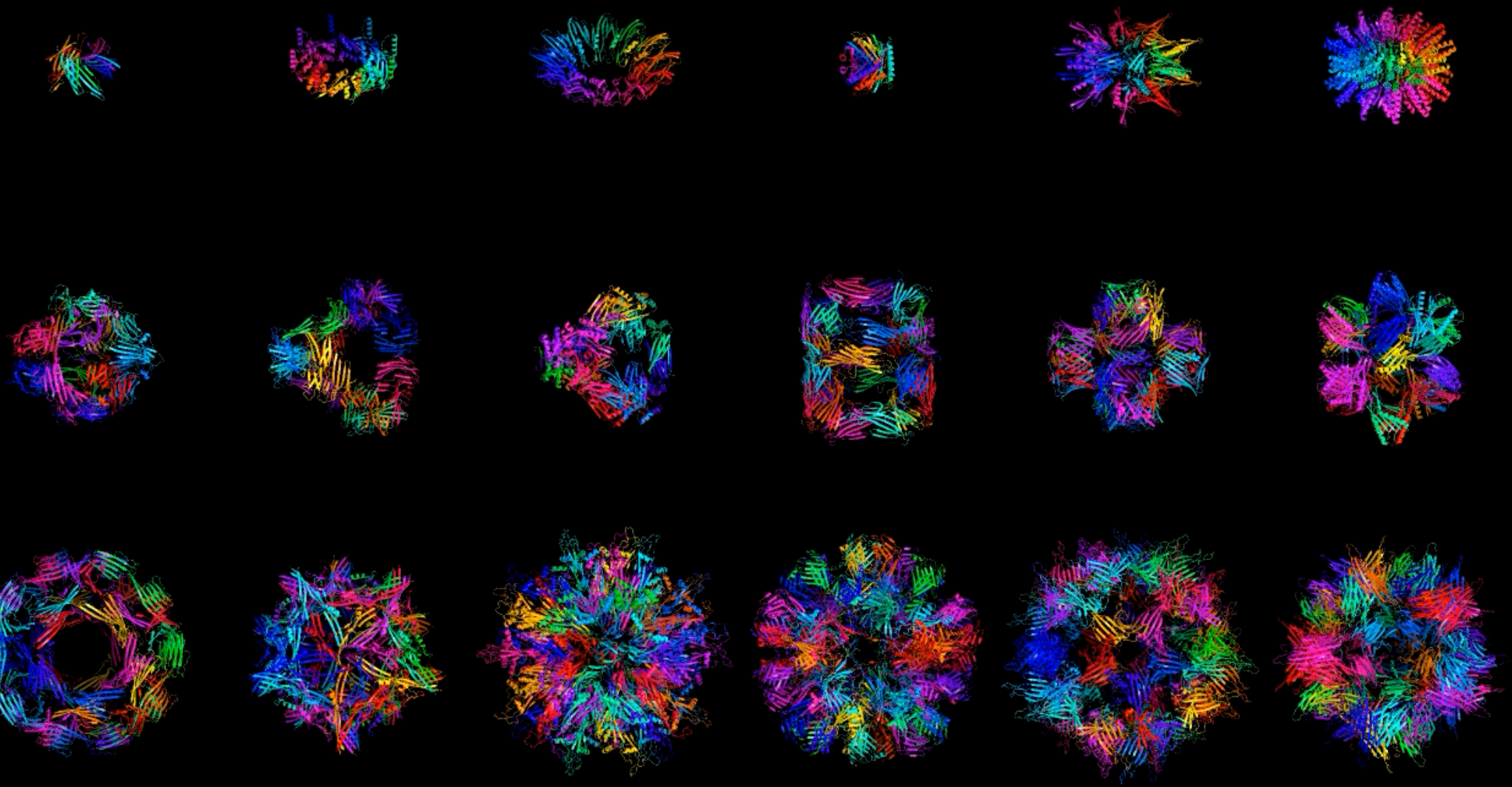





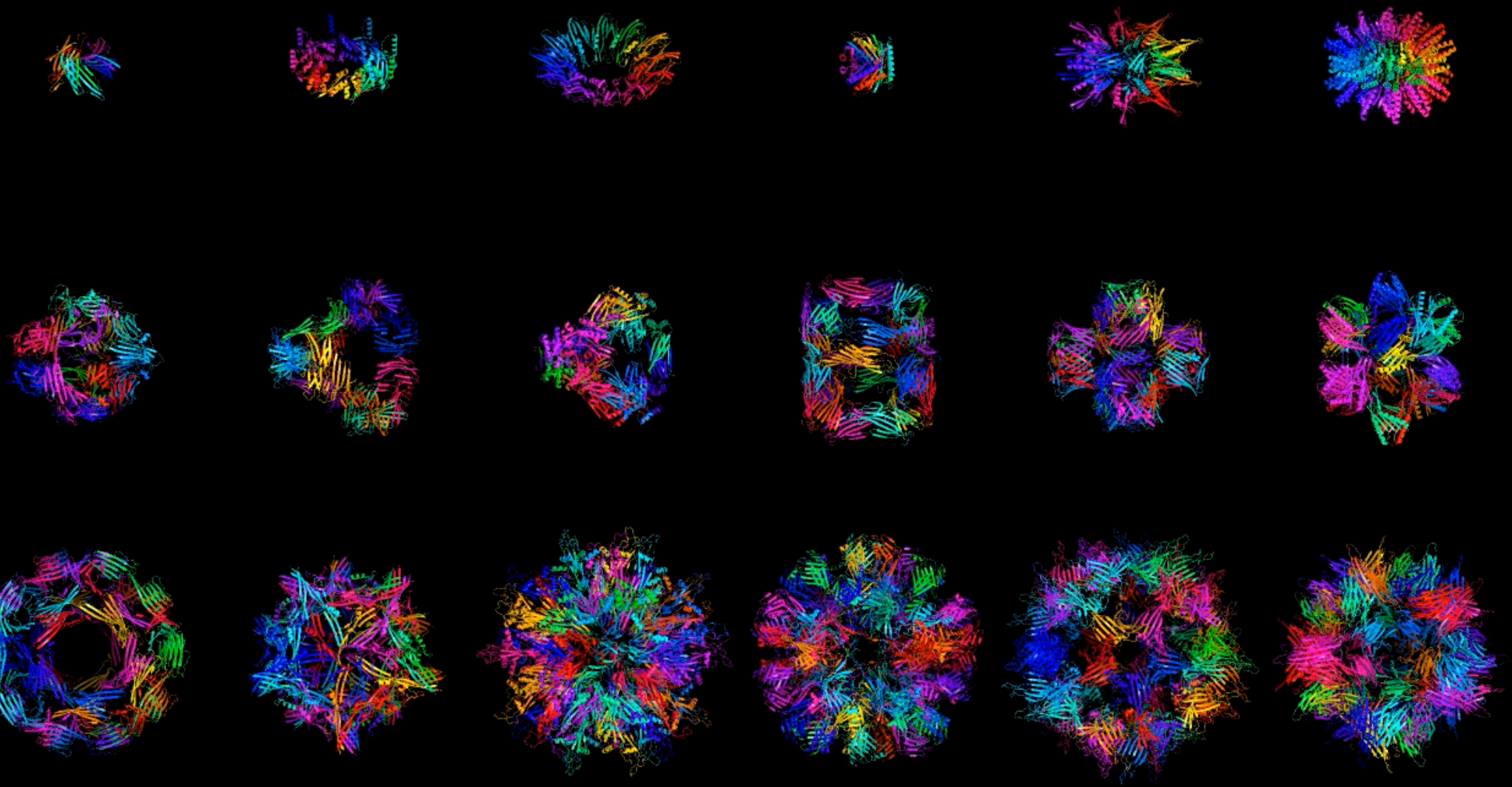
Programmable Symmetry in Complexes (scales)



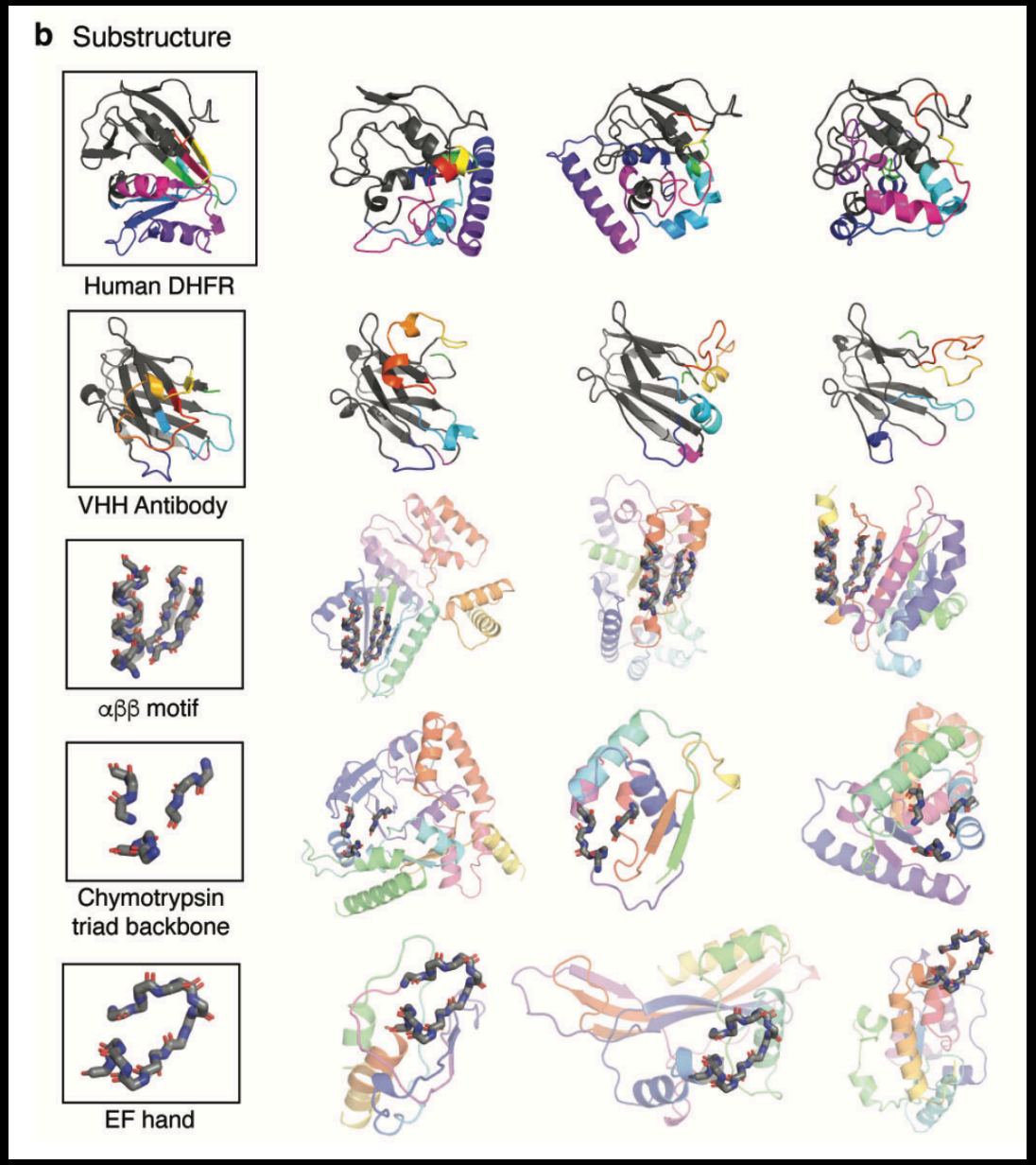
Programmable Symmetry in Complexes (scales)

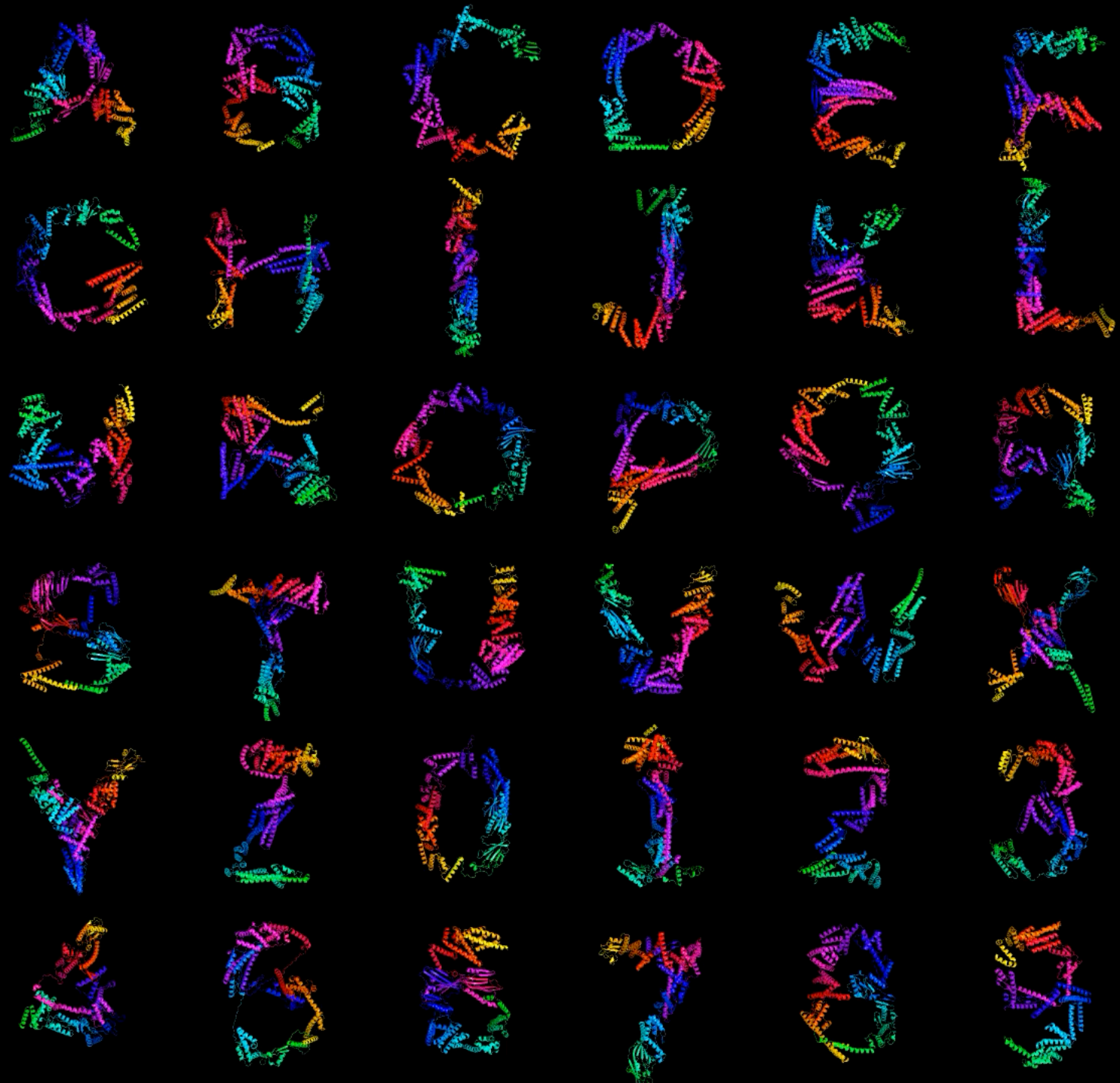


Programmable Symmetry in Complexes (scales)

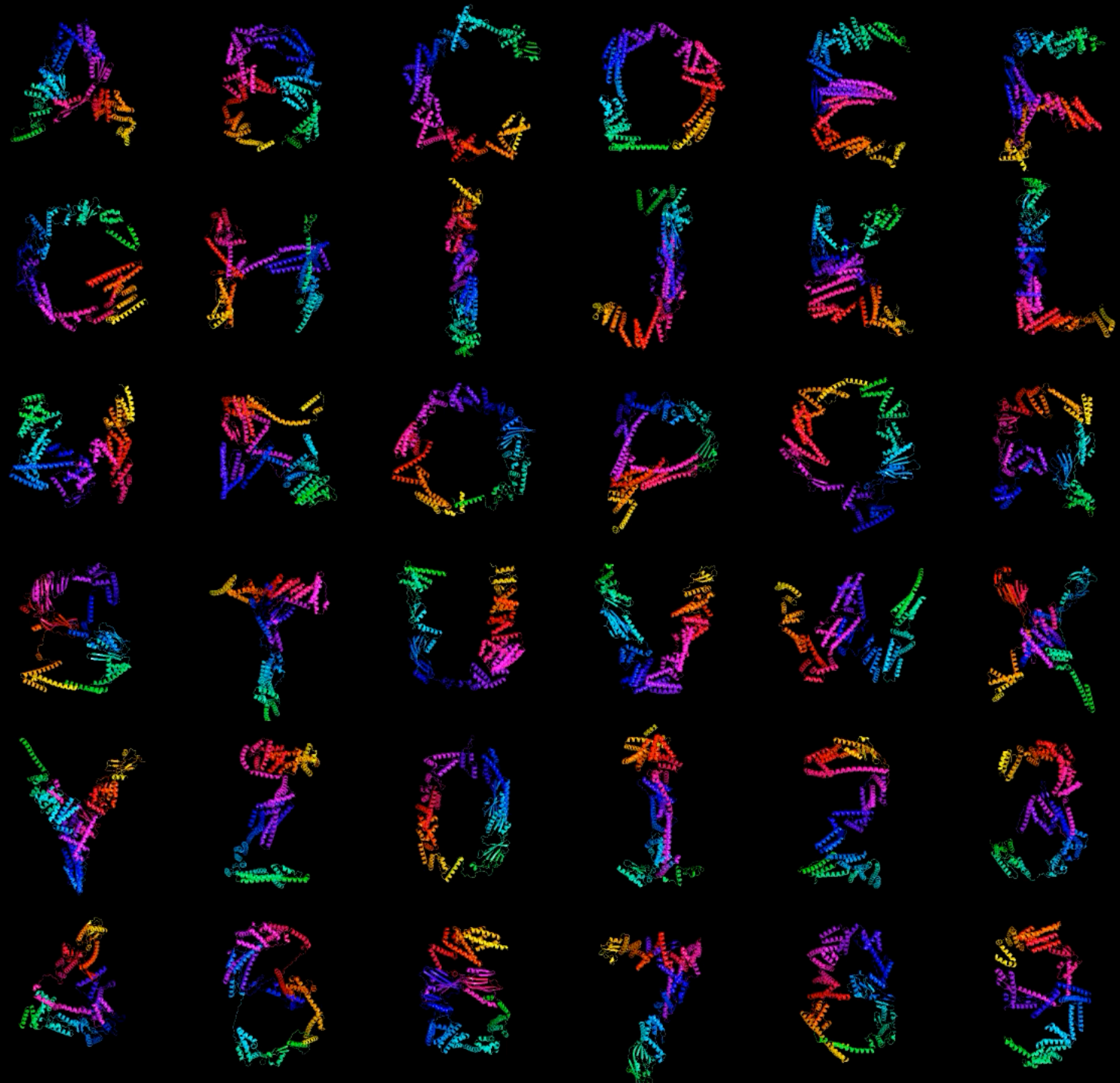


Programmable Substructure

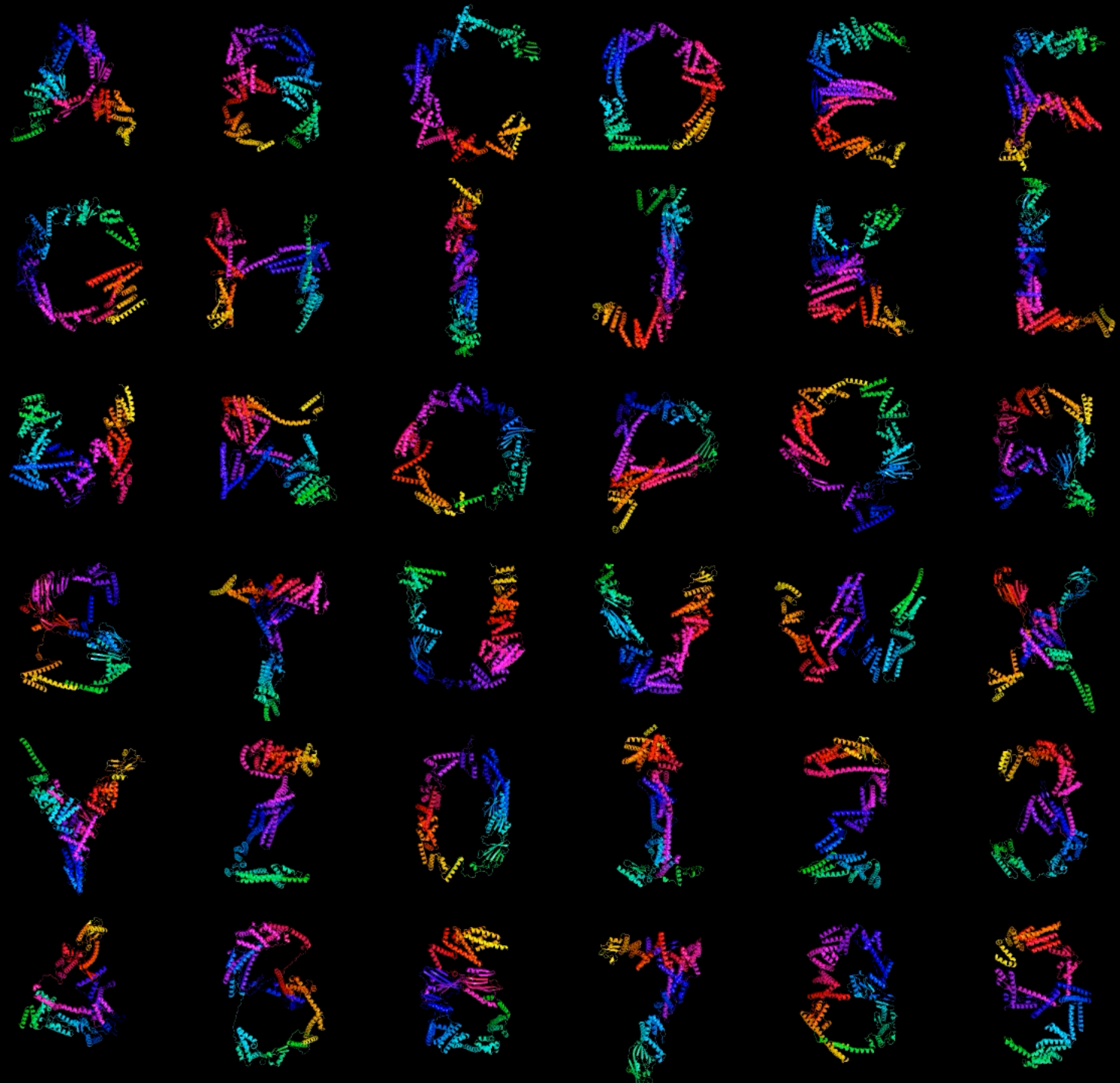




Generate : Biomedicines

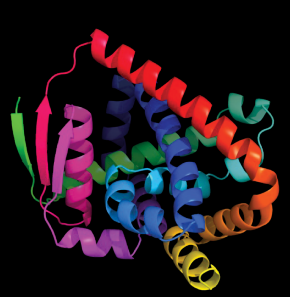
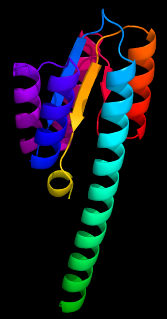
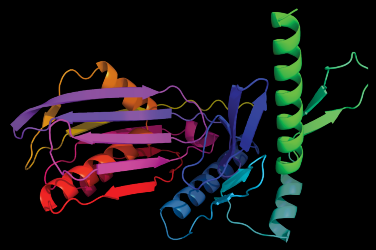
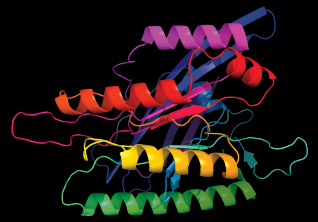
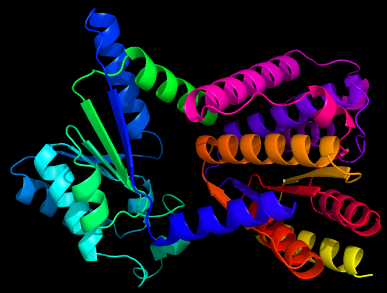
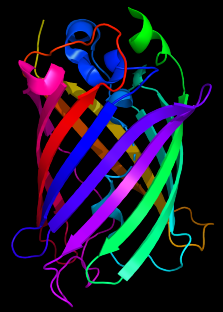
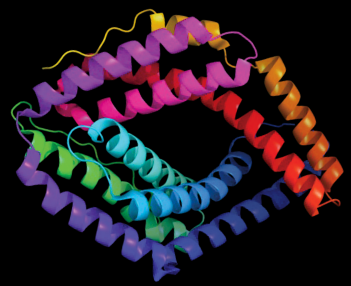


Generate : Biomedicines

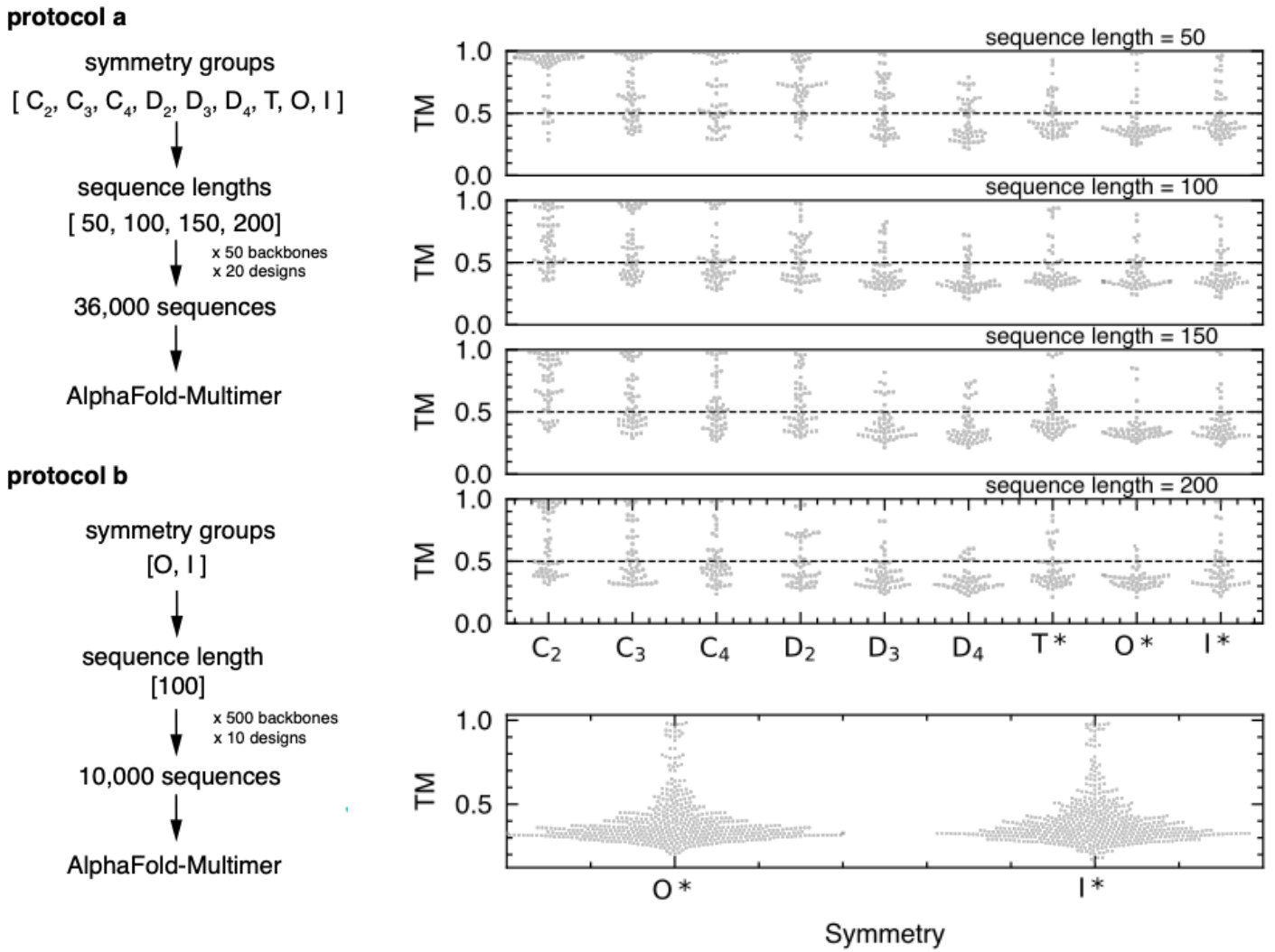


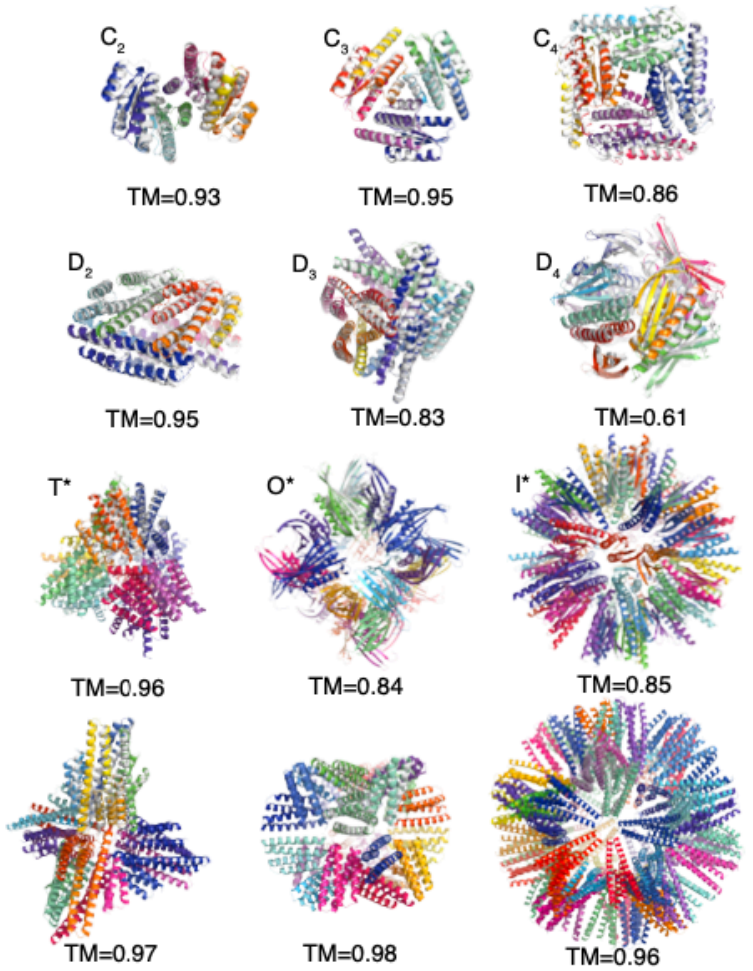
Generate : Biomedicines

We are beginning to see benefits of *semantic* programmability

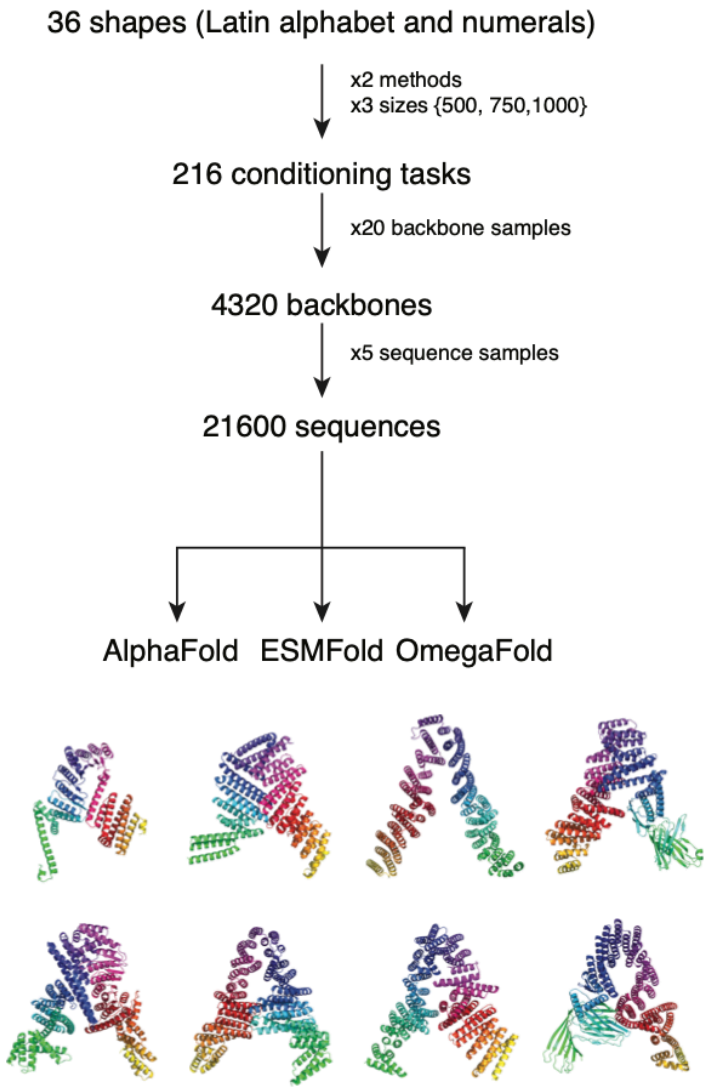
	Beta Barrel (2.40.155)	IG Fold (2.60.40)	"Protein with CHAD domain"	"Crystal Structure of Aminotransferase"
Not Conditioned				
Conditioned				
Real Example				

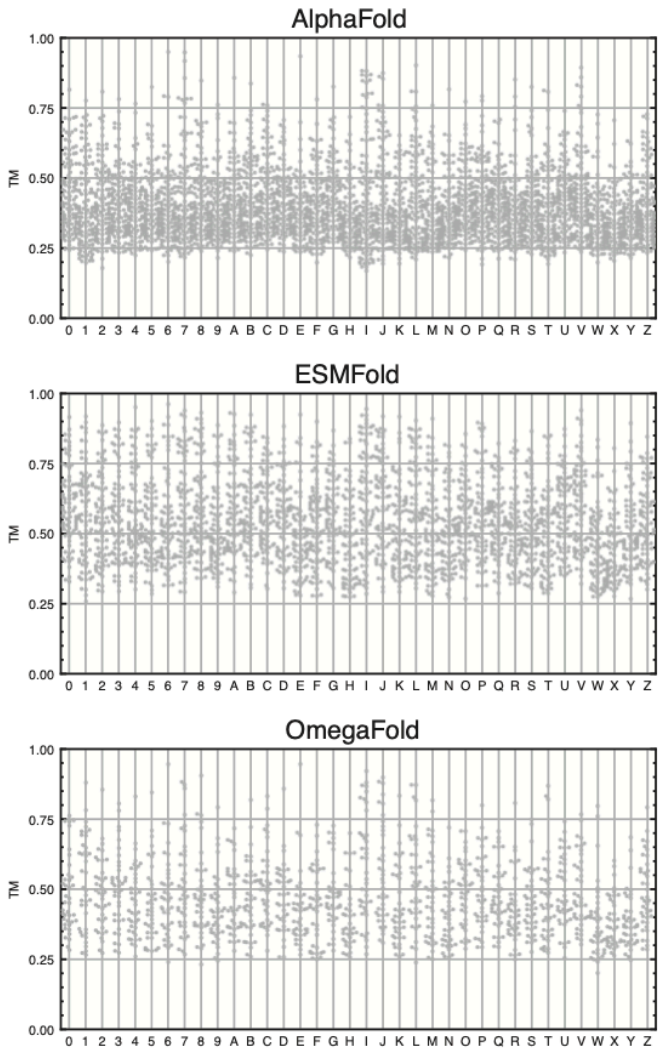
Symmetry-conditioned samples can refold *in silico*

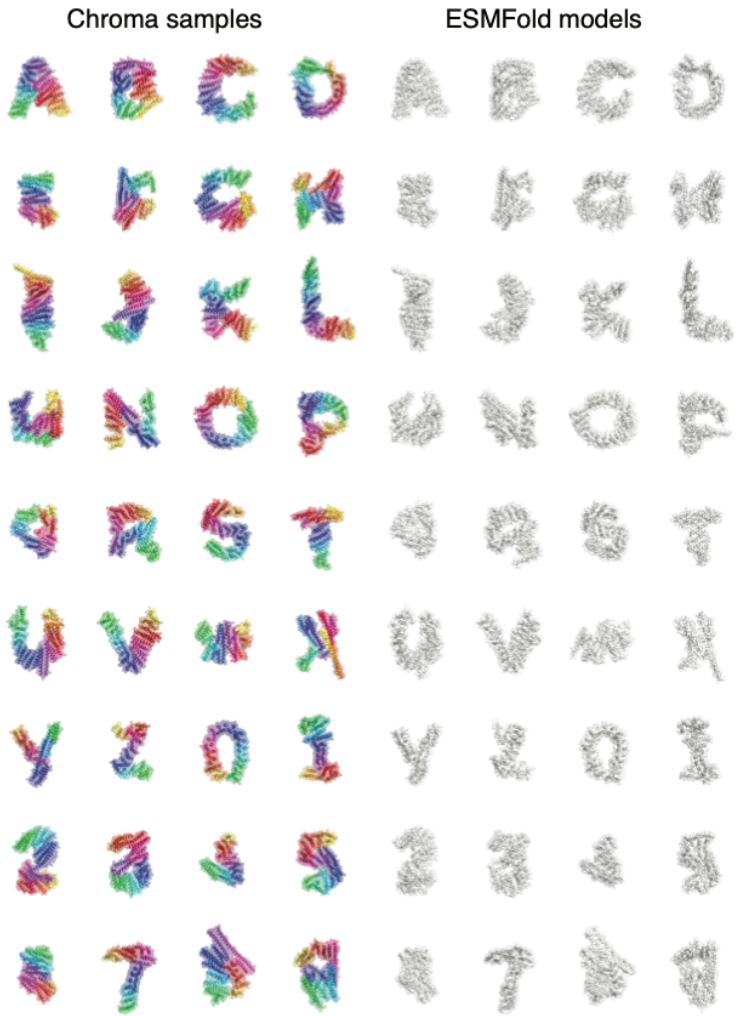




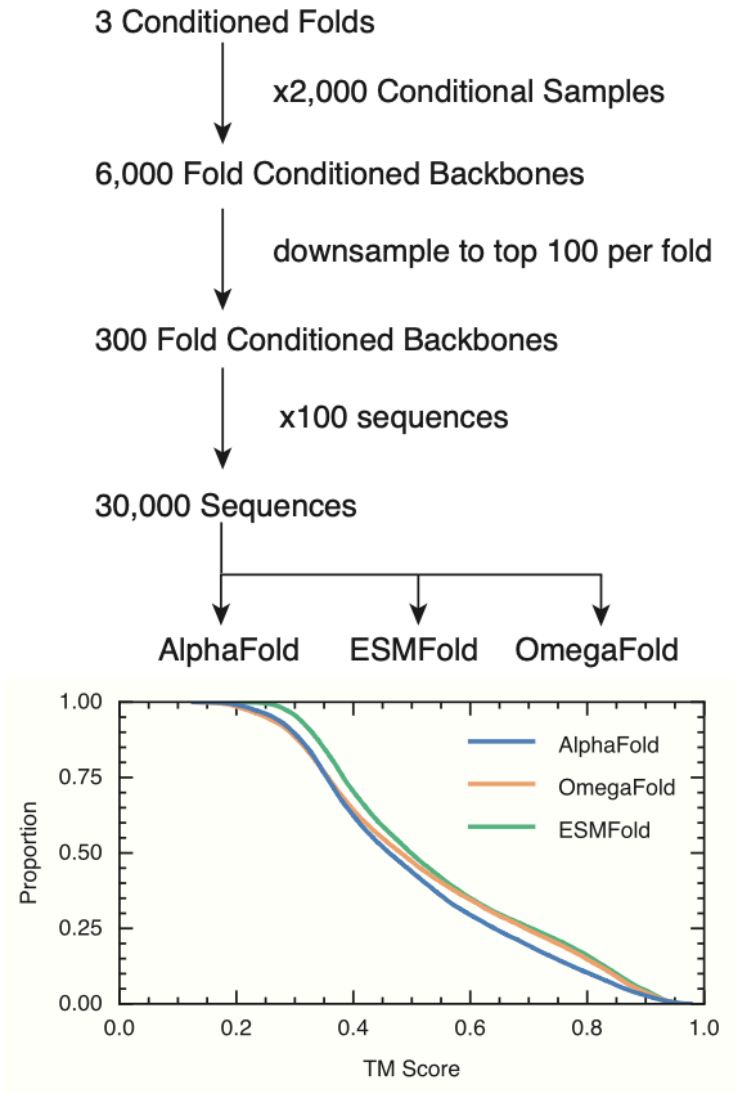
Shape-conditioned samples can refold *in silico*

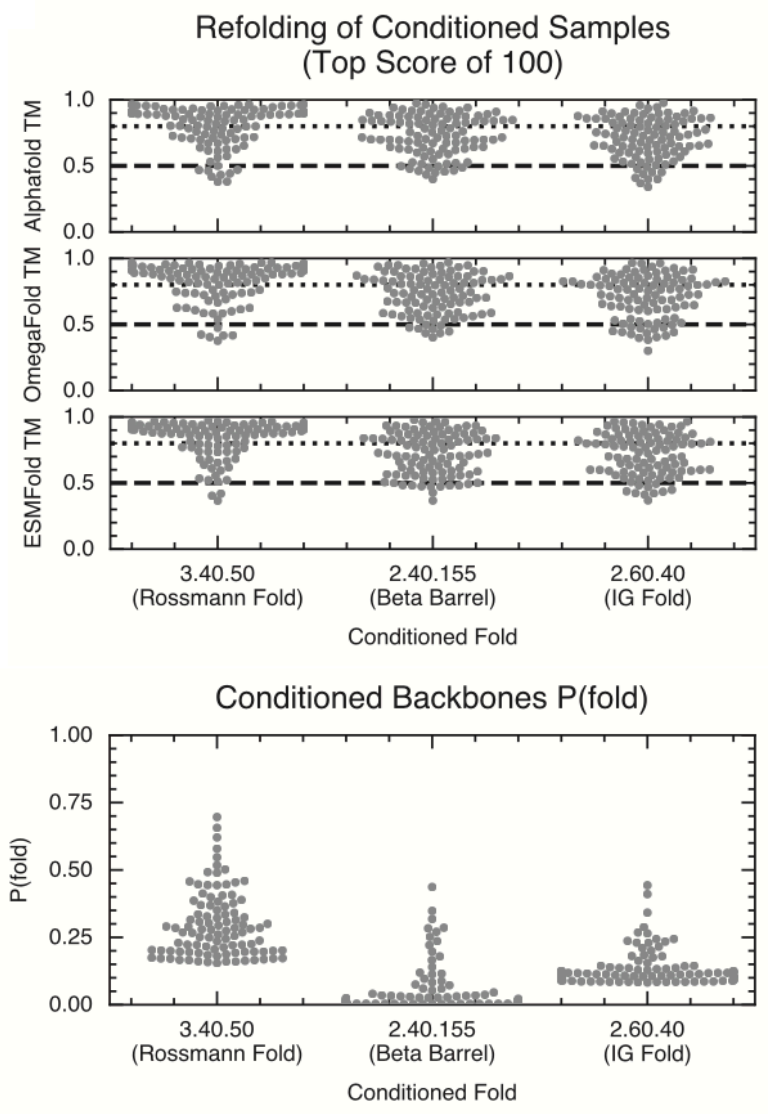


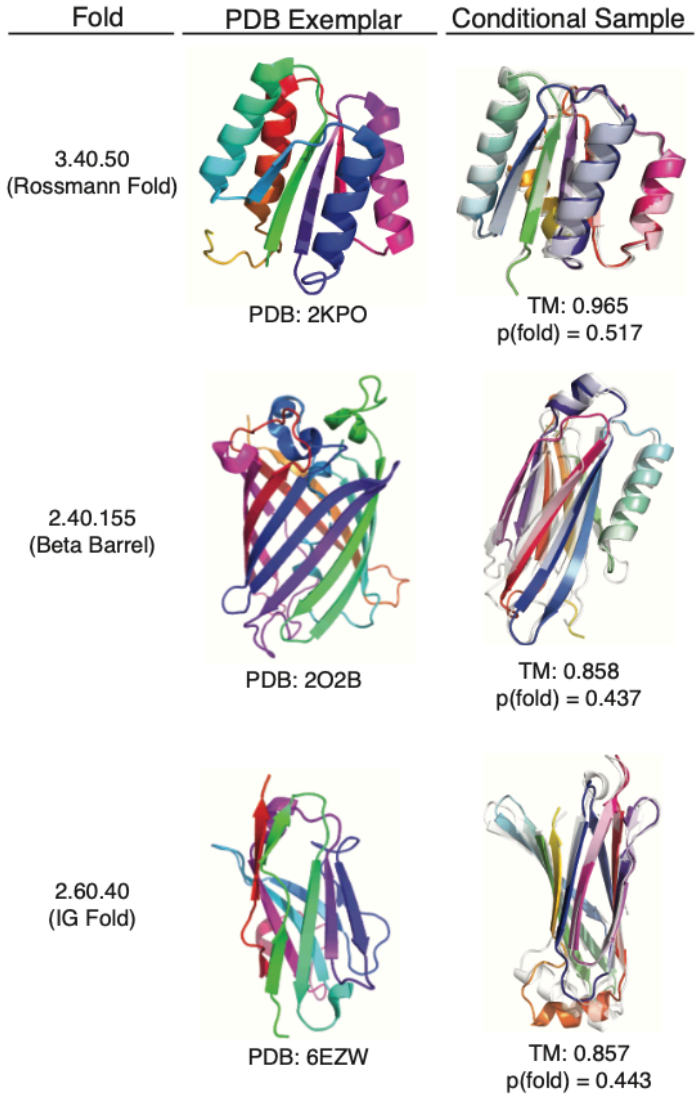




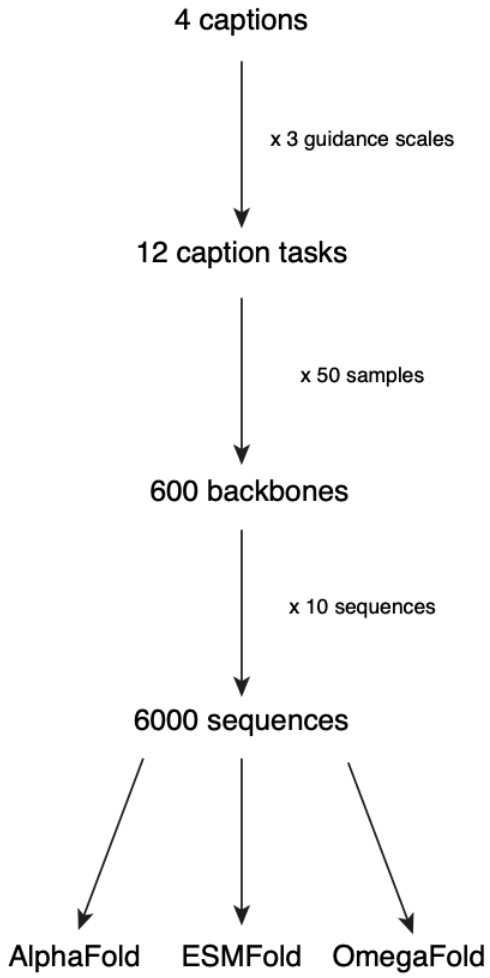
Class-conditioned samples can refold *in silico*

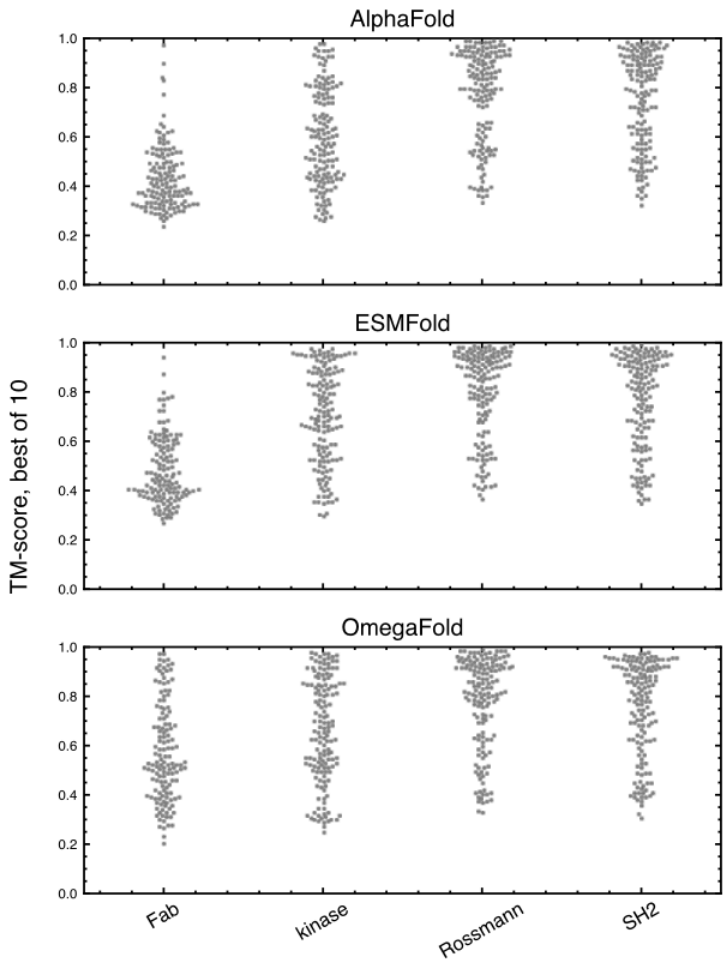


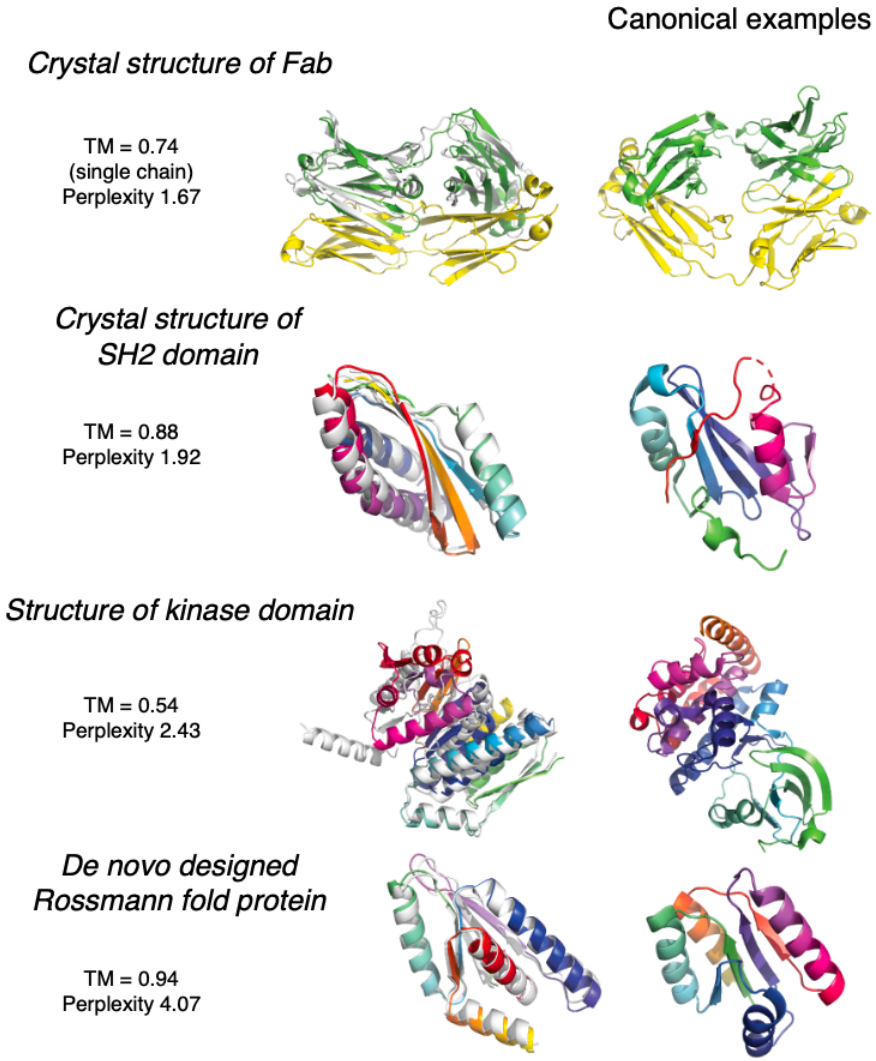




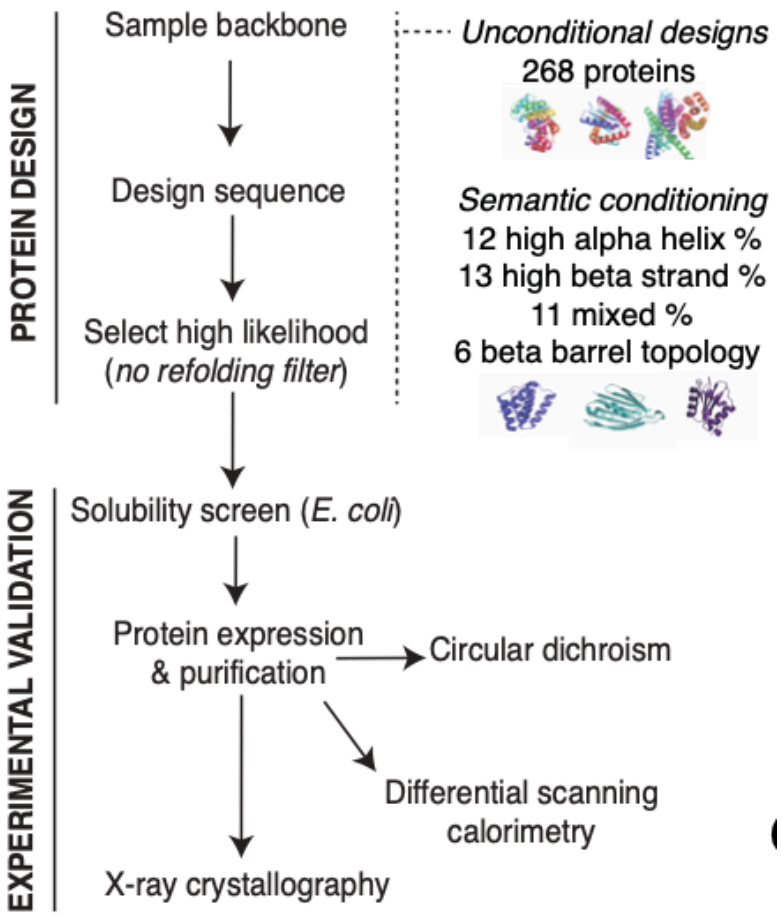
Text-conditioned samples can refold *in silico*



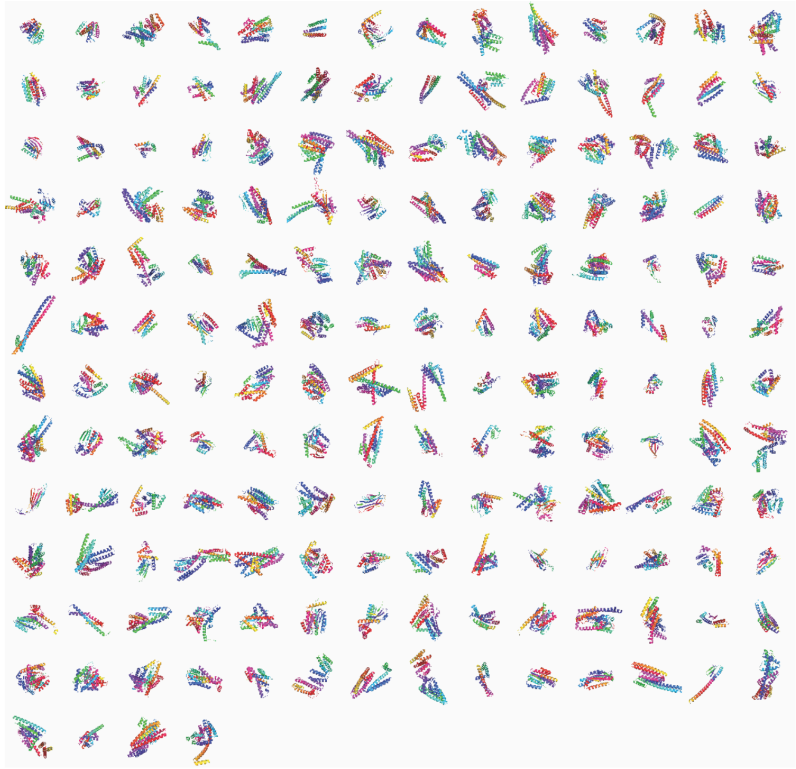




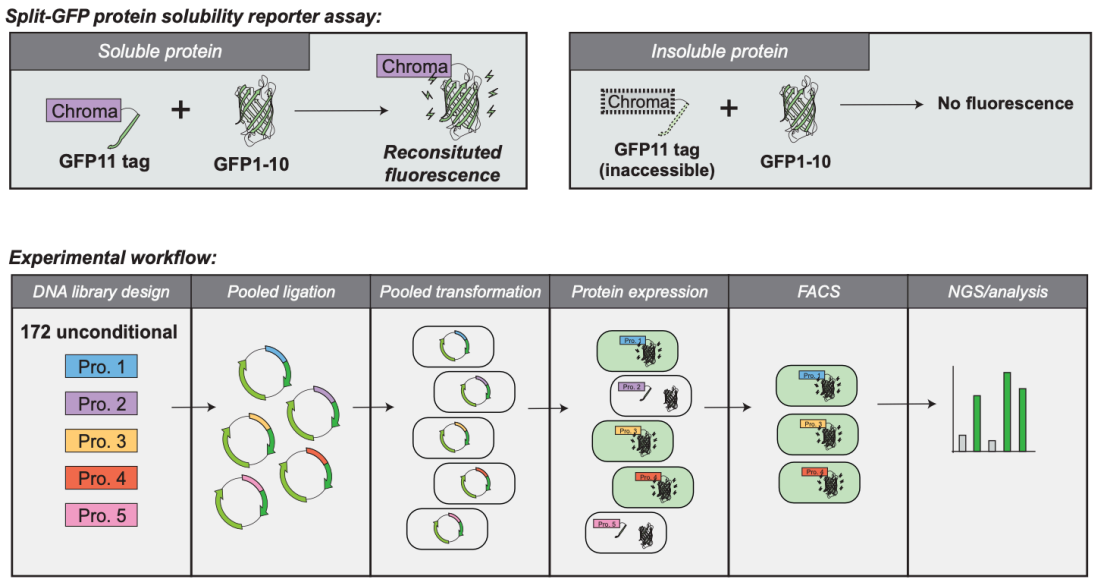
We designed an experiment intentionally to test **sufficiency**

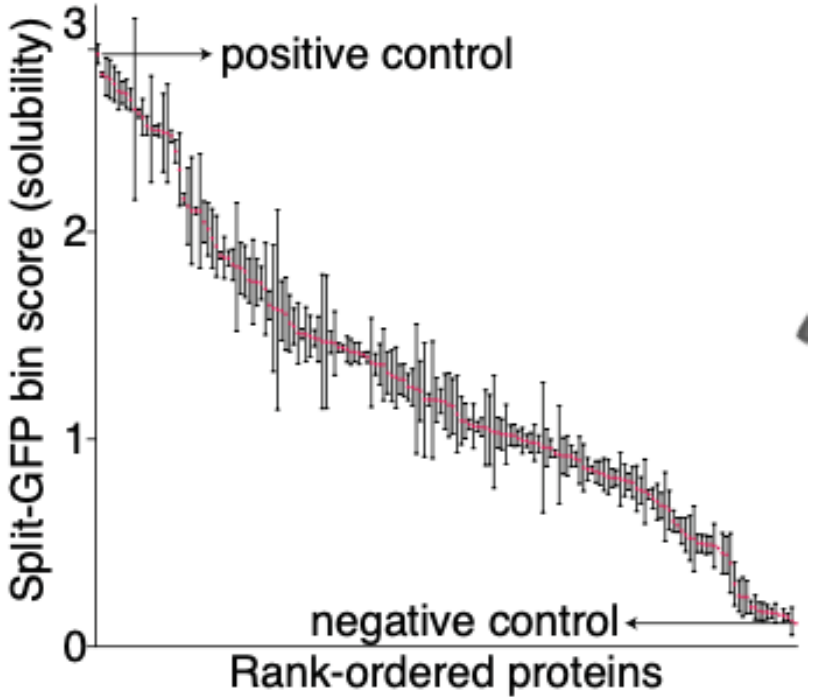


- Only sample using the **base model,** no external ML models or energy function filters

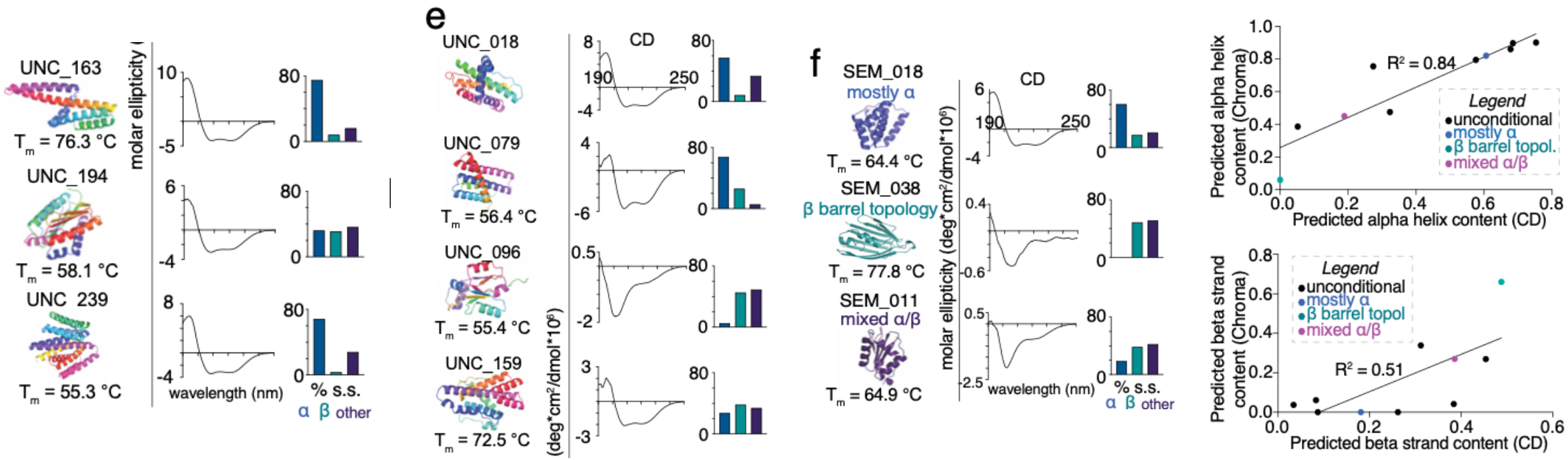


We saw samples from Chroma have favorable **expression**

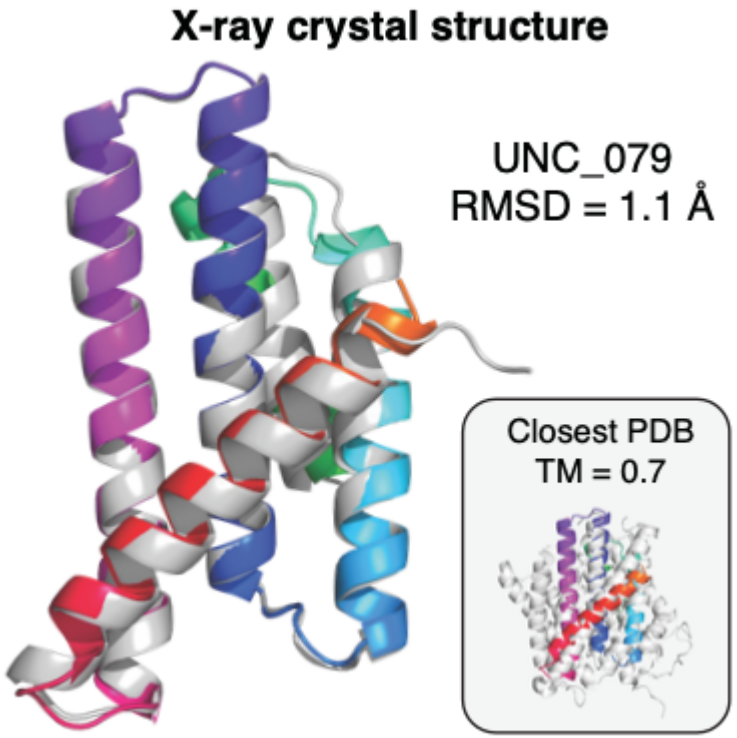


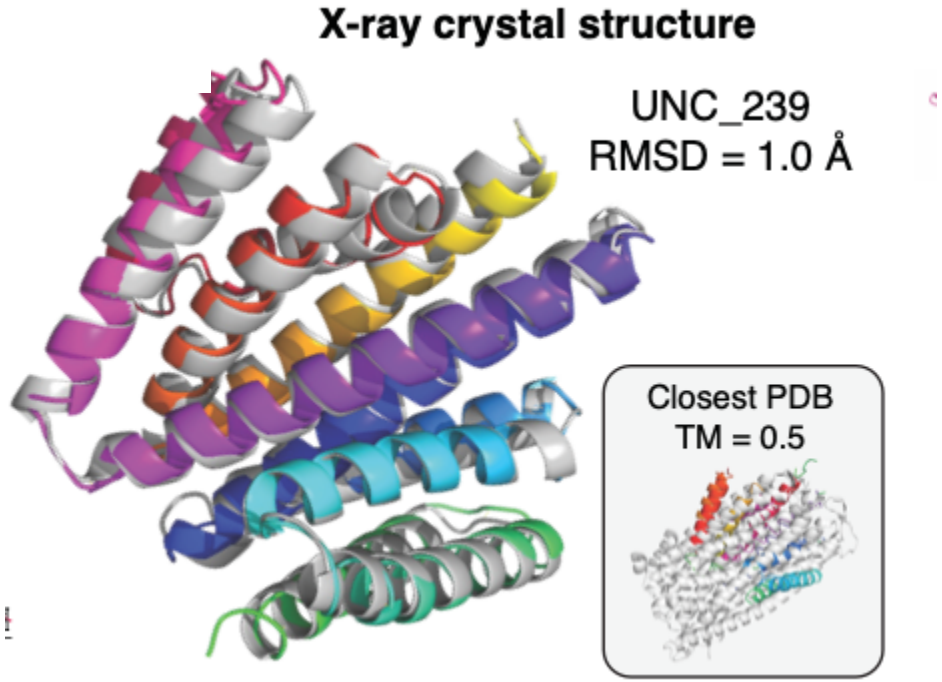


We saw samples from Chroma show evidence of folding and **biophysically behaving** as predicted



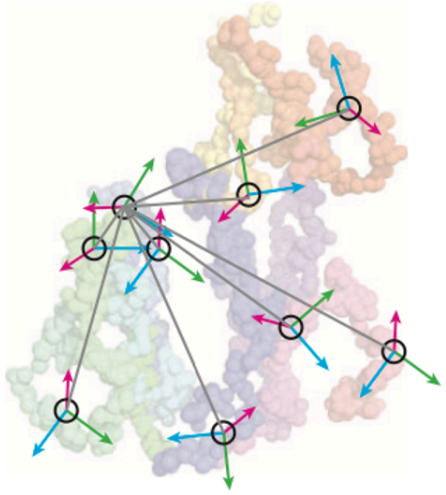
We saw direct samples from Chroma show **atomic agreement** within the width of an atom (X-ray Xtal)



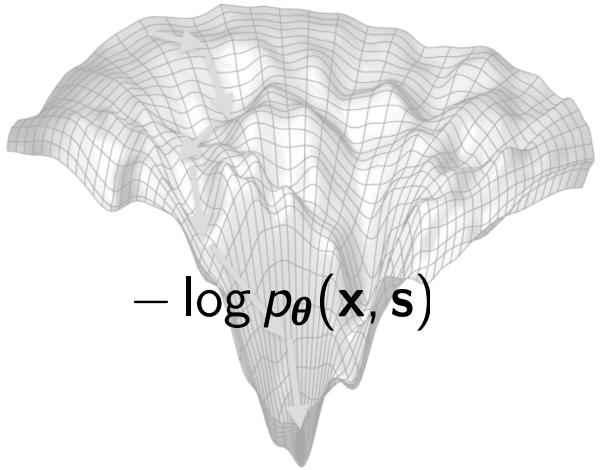


Three properties we find important for a protein generative modeling framework

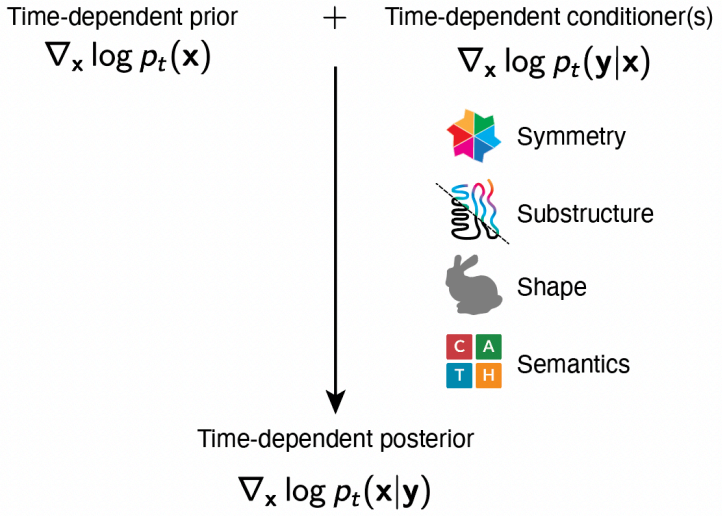
Subquadratic architecture



Likelihood that is useful and *controllable*



Programmable generation



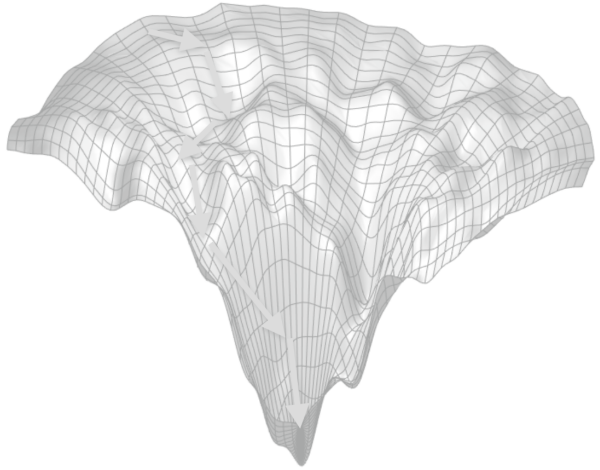
Generative models are influencing many fields of science, algorithms still needed!

Genetics

AQKLYLTHIDAEVEGD
ADRLYMTKIHHQFDGD
ADTLFITEVKQVFEGD
ADRLYMTKIHHTFDGD
ADKLYCTLIHNSFEGD
ADRLYMTKIHHEFEGD

Likelihood estimation

Molecular Physics



Amortizing intractable inference

Molecular Engineering

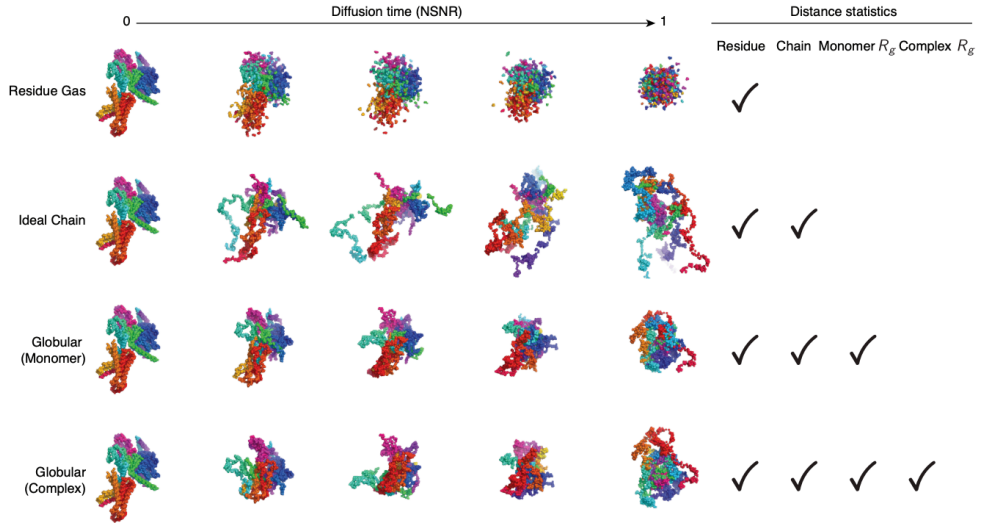
function ⋯⋯▸



Controllable generation

APPENDIX

Globular Covariance model



Linear time matrix-vector products & analytic determinant

Inverse covariance for our diffusion process

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{a^2}\left(b \begin{bmatrix} 1 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 1 \end{bmatrix} + (1-b)^2\mathbf{I} + \begin{bmatrix} b(1-b) & & & \\ & & & \\ & & & \\ & & & b(1-b) \end{bmatrix} \right)$$

Chain springs

Burial springs

boundary condition

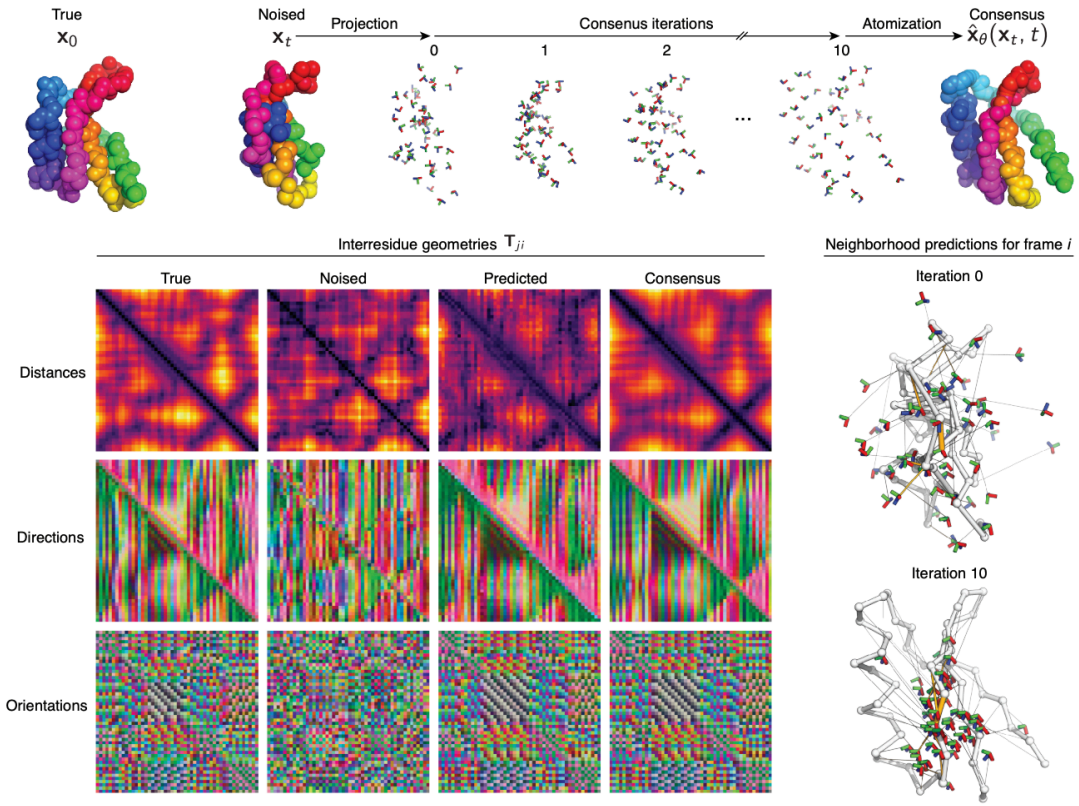
Inter-residue consensus algorithm enables distributed pairwise representations to synethesize a self-consistent atomic structure

Pose is a rotation and translation

$$\mathbf{T} = (\mathbf{O}, \mathbf{t}) \in \mathrm{SE}(3)$$

Every node in the graph predicts the relative poses of all neighbors together with per-edge confidence values.

We update the position of each node to the confidence-weighted average.

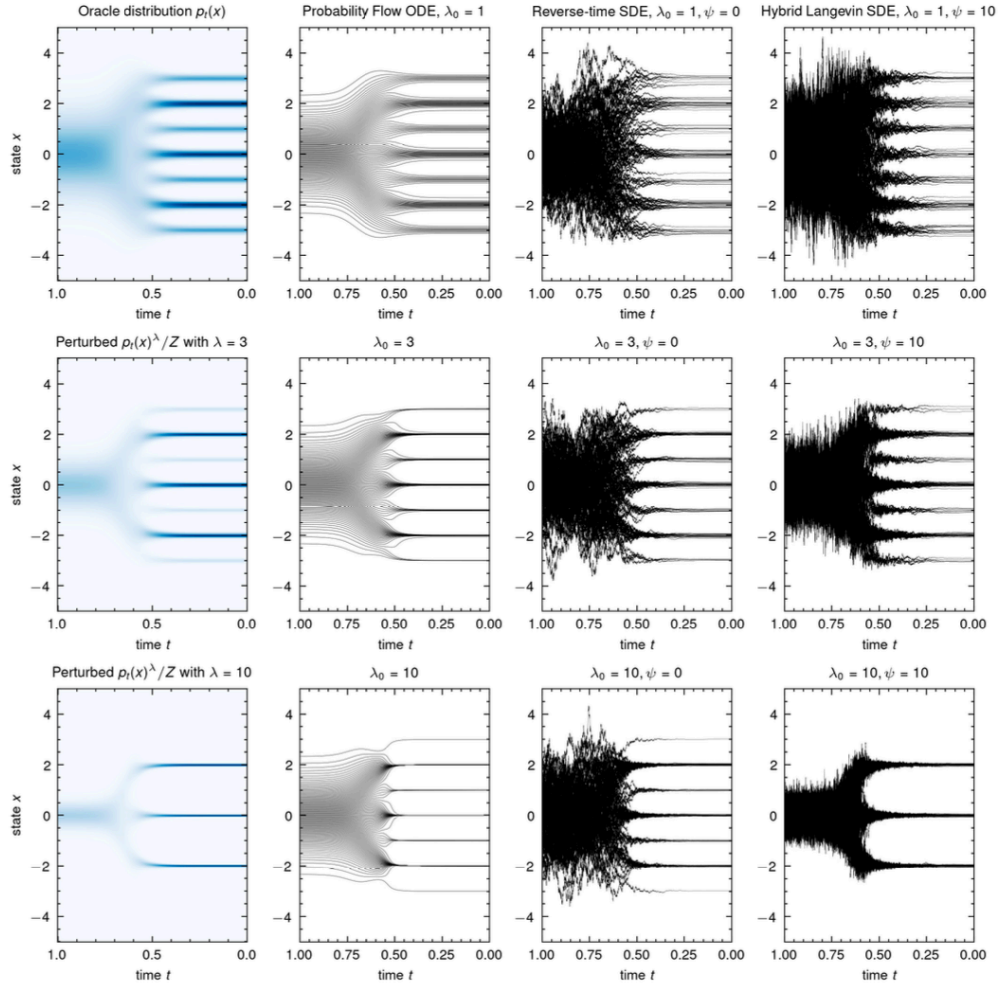


Analytic pose averaging = Weighted average with prediction

$$\mathbf{T}_i^\star = \left(\mathrm{Proj}_{\mathrm{SO}(3)} \left(\sum_j p_{ij} \mathbf{O}_j \mathbf{O}_{ji} \right), \sum_j p_{ij} (\mathbf{O}_j \mathbf{t}_{ji} + \mathbf{t}_j) \right), \quad \text{where } p_{ij} = \frac{w_{ij}}{\sum_j w_{ij}}$$

Confidence weights

Low temperature, Langevin-enriched “Hybrid SDE” enables low temperature and other perturbed sampling with re-equilibriation



Diffusion in … whitened space

$$\mathrm{d}\mathbf{z} = h_t \mathbf{z}\,\mathrm{d}t + g_t\,\mathrm{d}\mathbf{w}$$

… correlated space

$$\begin{aligned}\mathrm{d}\mathbf{x} &= \mathbf{R}\mathrm{d}\mathbf{z} \\ &= h_t \mathbf{R}\mathbf{z}\,\mathrm{d}t + g_t \mathbf{R}\,\mathrm{d}\mathbf{w} \\ &= h_t \mathbf{x}\,\mathrm{d}t + g_t \mathbf{R}\,\mathrm{d}\mathbf{w}.\end{aligned}$$

Reverse diffusion SDE (Song et al 2021)

$$\mathrm{d}\mathbf{x} = h_t\,\mathbf{x} - g_t^2\,\mathbf{R}\mathbf{R}^\intercal \nabla_{\mathbf{x}} \log p_t(\mathbf{x})\,\mathrm{d}t + g_t\,\mathbf{R}\,\mathrm{d}\bar{\mathbf{w}}.$$

Langevin Dynamics at inverse temperate λ_0

$$\mathrm{d}\mathbf{x} = -g_t^2 \frac{\psi}{2} \mathbf{R}\mathbf{R}^\intercal \lambda_0 \nabla_{\mathbf{x}} \log p_t(\mathbf{x})\,\mathrm{d}t + g_t \sqrt{\psi}\,\mathbf{R}\,\mathrm{d}\bar{\mathbf{w}}.$$

Hybrid SDE = sum of Reverse SDE + Langevin

$$\mathrm{d}\mathbf{x} = h_t\,\mathbf{x} - g_t^2 \left(\lambda_t + \frac{\lambda_0 \psi}{2}\right) \mathbf{R}\mathbf{R}^\intercal \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)\,\mathrm{d}t + g_t \sqrt{(1+\psi)}\,\mathbf{R}\,\mathrm{d}\bar{\mathbf{w}}$$

Time dependent annealing for reverse SDE based on Gaussian case

$$\lambda_t = \frac{(1-\alpha_t^2)\sigma_{\text{prior}}^2 + \alpha_t^2 \sigma_{\text{data}}^2}{(1-\alpha_t^2)\sigma_{\text{prior}}^2 + \alpha_t^2 \sigma_{\text{data}}^2/\lambda_0} \approx \frac{\lambda_0}{\alpha_t^2 + (1-\alpha_t^2)\lambda_0}$$