

# Why generate proteins?

All UniProt sequences: **~2e8 sequences**; << All medium-size proteins: **20<sup>400</sup>~2.6e520**

Most sequences are useless, just as “most” images → we need to sample “good ones”

Diffusion/flow models are good for images → let’s try to use them for proteins, which are also high-dimensional objects with complicated correlations, but with protein-specific tweaks

Article | [Open access](#) | Published: 15 November 2023

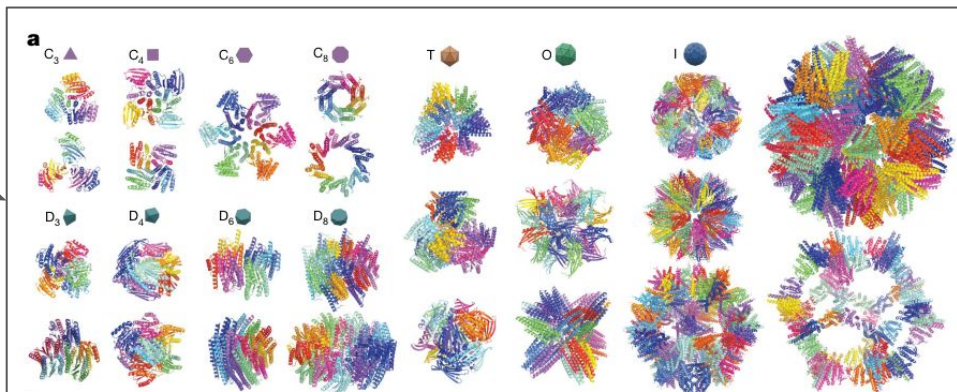
## **Illuminating protein space with a programmable generative model**

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

[Nature](#) **623**, 1070–1078 (2023) | [Cite this article](#)

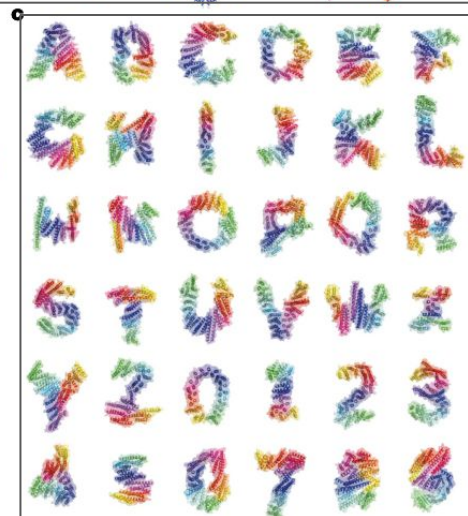
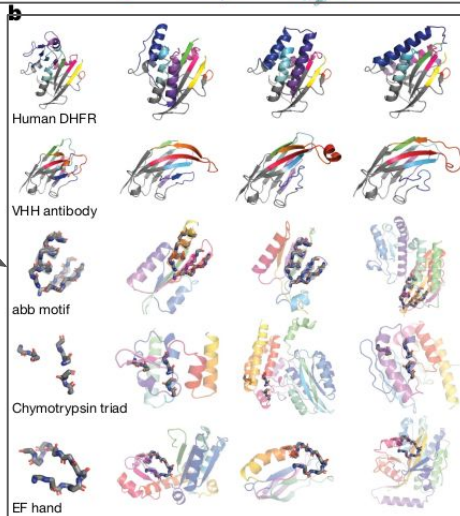
# Chroma: Controllable (bio)physics-informed generation of proteins

Can control  
symmetry



Is it just  
picture-making?

Can build on  
motifs

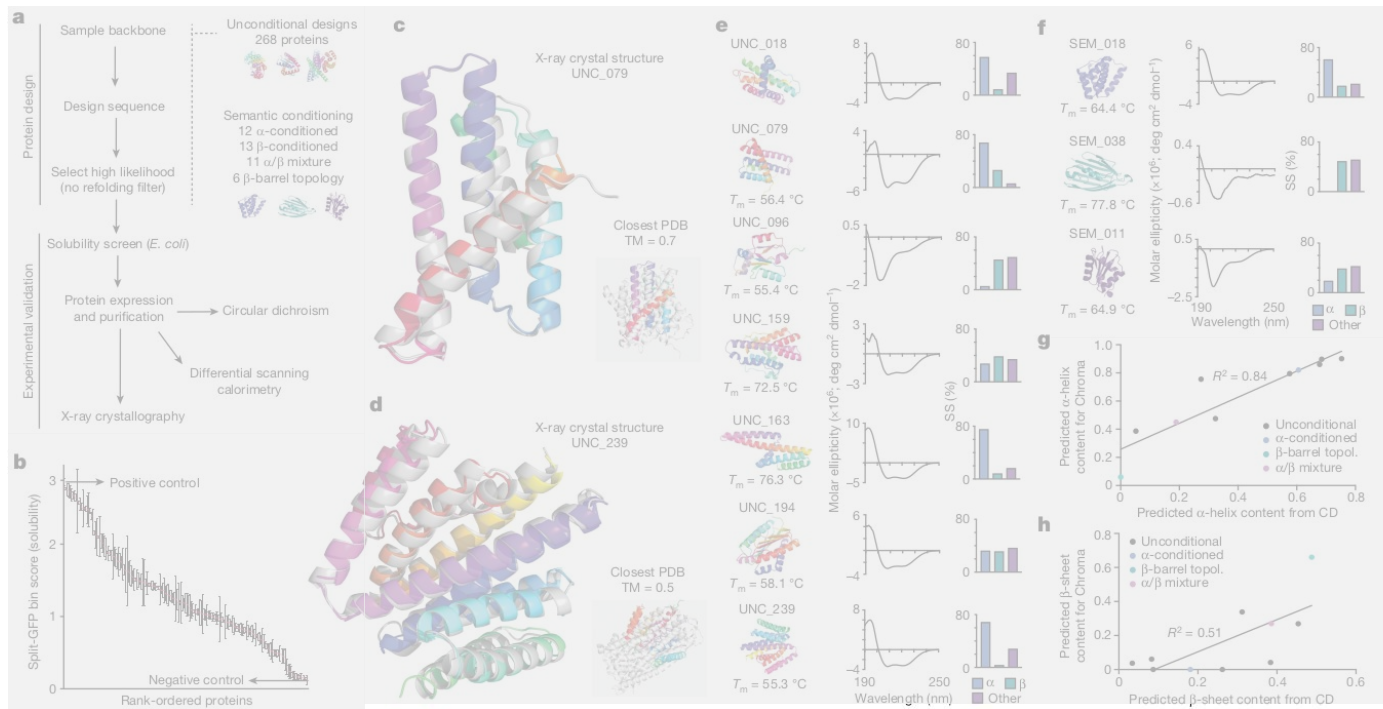


Can control  
shape

# Is it just picture-making? - no, seems to work in experiments

Generate  
172 uncond.  
42 cond.  
\*no in-silico  
filtering

Select  
17  
ones in  
solubility

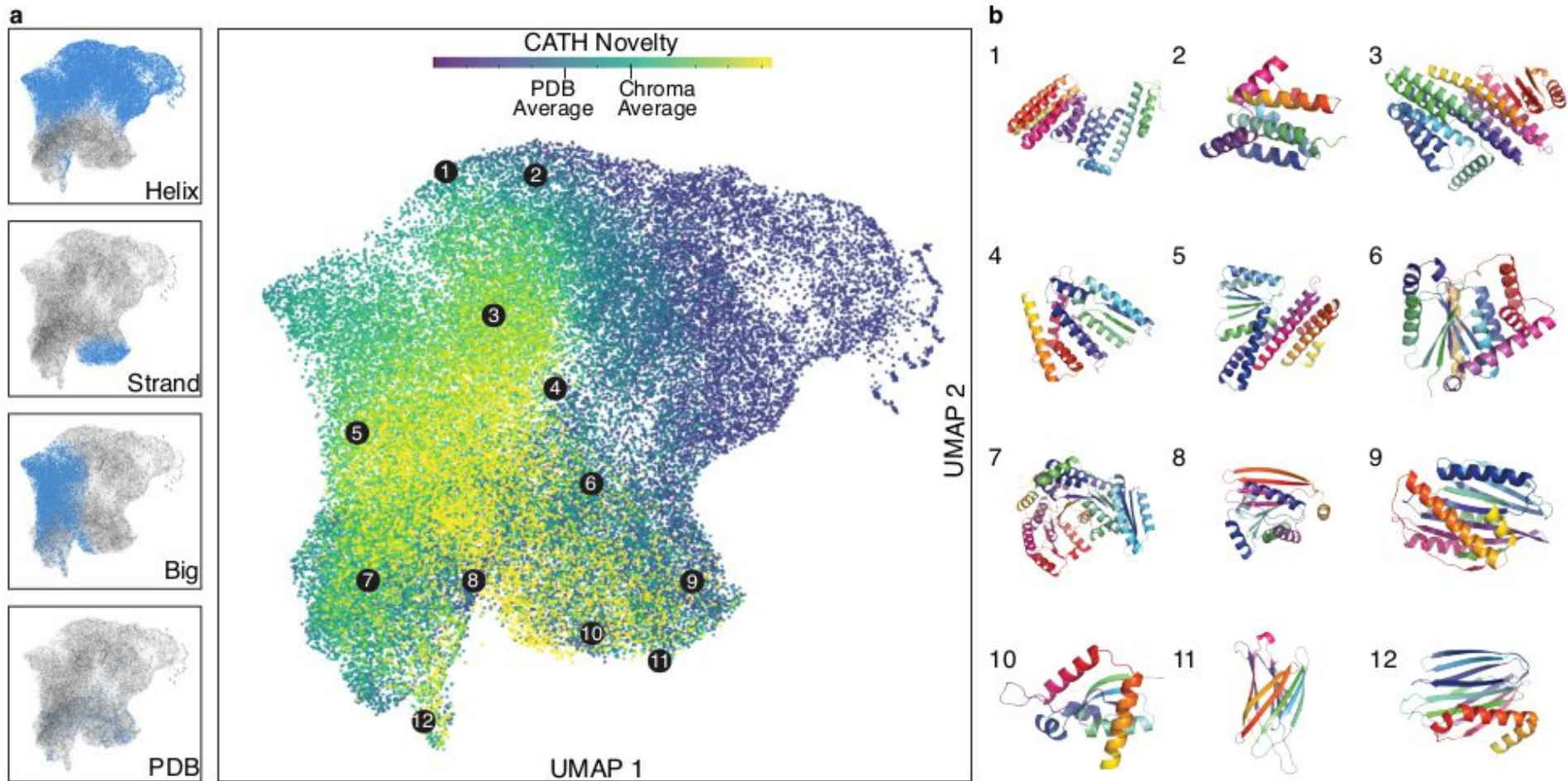


Select 10  
with  
more  
a/b  
structure

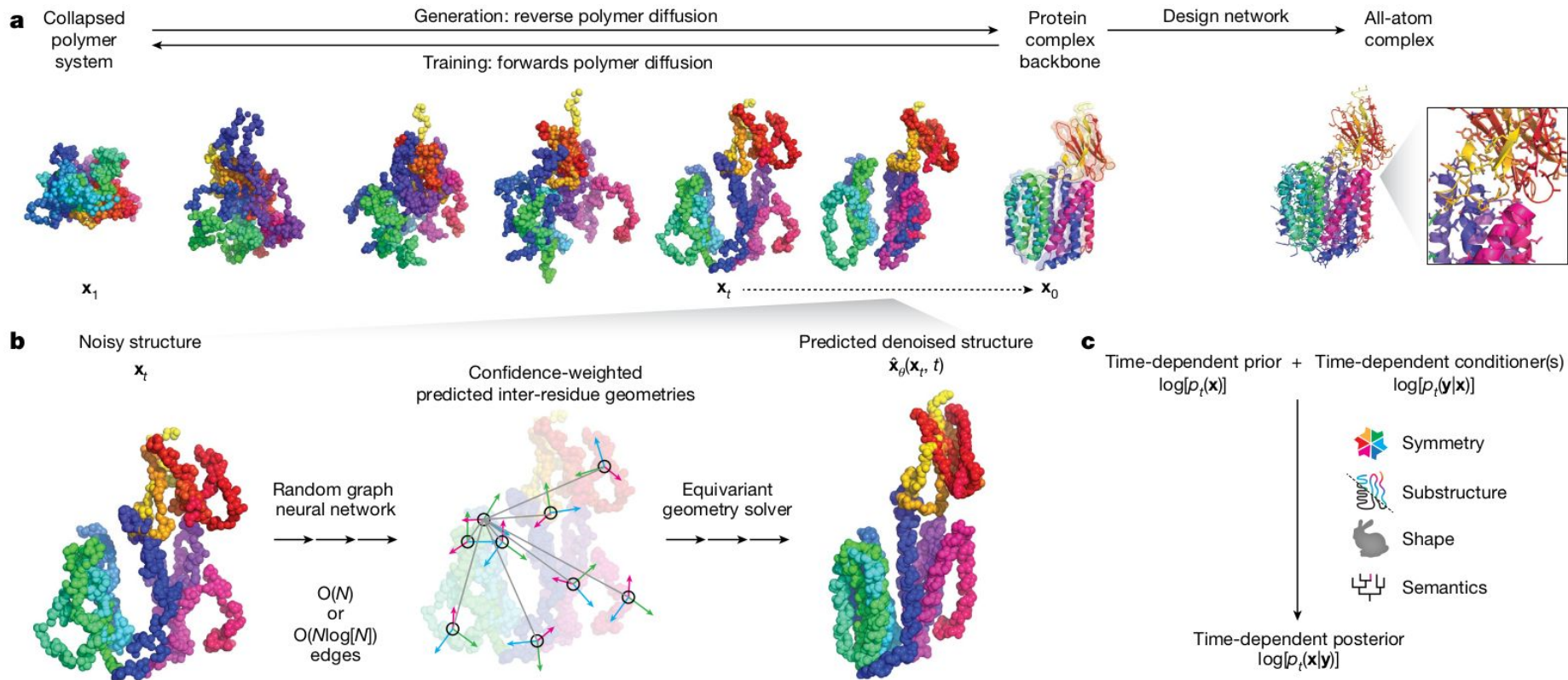
Structures match within  
RMSD~1Å  
for 2/2 crystallized ones

Conditioning (on a/b content)  
works reasonably

# Diversity of (unconditioned) samples: PDB-wide or better



# The pipeline overview



# How they get there

- 1) The dataset
- 2) Protein-specific tricks: domain knowledge helps (reverse bitter lesson?)
  - a) Protein-structure-inspired separation of degrees-of-freedom (DOFs)
  - b) Protein-structure-inspired loss for protein substructure alignment
  - c) Protein-structure-based correlations in the initial noise
  - d) Protein-structure-inspired selection of important DOFs
  - e) Physics-inspired approach to sampling more “low-energy” structures
  - f) Physics-inspired approach to AA orientation
  - g) Physics-inspired side-chains and simple sequence distributions
- 3) Ablation study (which parts are important)
- 4) Supported conditioning modes

# How they get there

## 1) The dataset

- 2) Protein-specific tricks: domain knowledge helps (reverse bitter lesson?)
  - a) Protein-structure-inspired separation of degrees-of-freedom (DOFs)
  - b) Protein-structure-inspired loss for protein substructure alignment
  - c) Protein-structure-based correlations in the initial noise
  - d) Protein-structure-inspired selection of important DOFs
  - e) Physics-inspired approach to sampling more “low-energy” structures
  - f) Physics-inspired approach to AA orientation
  - g) Physics-inspired side-chains and simple sequence distributions
- 3) Ablation study (which parts are important)
- 4) Supported conditioning modes

## The dataset: 30k PDBs, <2.6Å resolution, 80/10/10 split

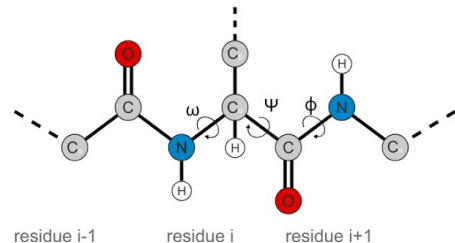
**Dataset** We constructed a dataset of 28,819 protein complex structures from the Protein Data Bank circa March 20, 2022 (Supplementary Appendix [H](#)). These complexes were filtered for X-ray crystal structures at resolution  $\leq 2.6$  Å and were then redundancy-reduced via general sequence clustering at 50% identity followed by re-enrichment of 1,726 highly-variable antibody systems with clustering at 10% sequence identity. We split these data into 80%/10%/10% training/validation/test components based on a graph-based annotation overlap reduction procedure.

# How they get there

- 1) The dataset
- 2) Protein-specific tricks: domain knowledge helps (reverse bitter lesson?)
  - a) Protein-structure-inspired separation of degrees-of-freedom (DOFs)**
  - b) Protein-structure-inspired loss for protein substructure alignment
  - c) Protein-structure-based correlations in the initial noise
  - d) Protein-structure-inspired selection of important DOFs
  - e) Physics-inspired approach to sampling more “low-energy” structures
  - f) Physics-inspired approach to AA orientation
  - g) Physics-inspired side-chains and simple sequence distributions
- 3) Ablation study (which parts are important)
- 4) Supported conditioning modes

# Protein-specific factorization of the structure

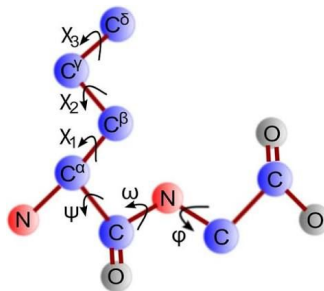
$$\log p_{\theta}(\mathbf{x}, \mathbf{s}, \chi) = \underbrace{\log p_{\theta}(\mathbf{x})}_{\text{backbone likelihood}} + \underbrace{\log p_{\theta}(\mathbf{s}|\mathbf{x})}_{\text{sequence likelihood}} + \underbrace{\log p_{\theta}(\chi|\mathbf{x}, \mathbf{s})}_{\text{side-chain likelihood}},$$



$\mathbf{x} \in \mathbb{R}^{4N \times 3}$  represents the backbone heavy atom coordinates (i.e., N, C $\alpha$ , C, and O)

$\mathbf{s} \in \llbracket 20 \rrbracket^N$  represents the discrete sequences over all residues

$\chi \in (-\pi, \pi]^{4N}$  represents the side-chain



diffusion modeling to estimate  $\log p_{\theta}(\mathbf{x})$

discretely supported distributions to estimate  $\log p_{\theta}(\chi, \mathbf{s}|\mathbf{x})$

# How they get there

- 1) The dataset
- 2) Protein-specific tricks: domain knowledge helps (reverse bitter lesson?)
  - a) Protein-structure-inspired separation of degrees-of-freedom (DOFs)
  - b) Protein-structure-inspired loss for protein substructure alignment**
  - c) Protein-structure-based correlations in the initial noise
  - d) Protein-structure-inspired selection of important DOFs
  - e) Physics-inspired approach to sampling more “low-energy” structures
  - f) Physics-inspired approach to AA orientation
  - g) Physics-inspired side-chains and simple sequence distributions
- 3) Ablation study (which parts are important)
- 4) Supported conditioning modes

# Training: ELBO + heuristic

$$\mathbf{x} = F(\mathbf{z}) = \mathbf{R}\mathbf{z} + \mathbf{b} \quad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

$\mathbf{R}$  = variance specific for proteins

$$p(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}(\mathbf{x}; \alpha_t \mathbf{x}_0, \sigma_t^2 \mathbf{R}\mathbf{R}^\top)$$

$$\log p(\mathbf{x}_0) = \log p(\mathbf{z}_0) - \log \det \mathbf{R}$$

$$\geq -\frac{N}{2} \log(2\pi e) - \log \det \mathbf{R} - \frac{1}{2} \int_0^1 \left( \frac{N}{1 + \text{SNR}_t} - \mathbb{E}_{p(\mathbf{x}_t | \mathbf{x}_0)} [\|\mathbf{R}^{-1} (\hat{\mathbf{x}}(\mathbf{x}_t, t) - \mathbf{x}_0)\|_2^2] \right) \text{SNR}'_t \, dt$$

$$= \underbrace{-\frac{1}{2} \log \det(2\pi e \mathbf{R}\mathbf{R}^\top)}_{\text{Entropy of the Gaussian prior}} - \underbrace{\frac{1}{2} \mathbb{E}_{p(\mathbf{x}_t | \mathbf{x}_0) p(t)} \left[ \text{SNR}'_t \left( \frac{N}{1 + \text{SNR}_t} - \|\mathbf{R}^{-1} (\hat{\mathbf{x}}(\mathbf{x}_t, t) - \mathbf{x}_0)\|_2^2 \right) \right]}_{\text{Deviation from Gaussianity (Bound)}}$$

$$\mathcal{L}_{\mathbf{x}}(\mathbf{x}_0; \boldsymbol{\theta}) = -\mathbb{E}_{\mathbf{x}_t \sim p(\mathbf{x}_t | \mathbf{x}_0), t \sim \text{Unif}(0,1)} \left[ \omega^{-2} \text{SNR}'_t \|\hat{\mathbf{x}}_{\boldsymbol{\theta}}(\mathbf{x}_t, t) - \mathbf{x}\|_2^2 \right] \quad \text{ELBO-weighted MSE}$$

$$\mathcal{D}_{\text{substructure}}(\mathbf{x}, \mathbf{x}') = \sum_{\mathcal{M}_i \in \{\mathcal{M}_i\}} \min_{\mathbf{T} \in \text{SE}(3)} \|\mathbf{x}^{\mathcal{M}_i} - \mathbf{T} \circ \mathbf{x}'^{\mathcal{M}_i}\|_2^2, \quad \text{Emphasis to protein substructures}$$

$$\mathcal{D}_{\text{distance}}(\mathbf{x}, \mathbf{x}') = \sum_{ij} (D_{ij}^{\text{CA}}(\mathbf{x}) - D_{ij}^{\text{CA}}(\mathbf{x}'))^2. \quad \mathcal{L}_{\mathcal{D}}(\mathbf{x}_0; \boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x}_t, \mathbf{x}'_t \sim p(\mathbf{x}_t | \mathbf{x}_0), t \sim \text{Unif}(0,1)} \left[ \frac{1}{p(\text{SSNR}_t)} \frac{\mathcal{D}(\hat{\mathbf{x}}(\mathbf{x}_t, t); \mathbf{x}_0)}{\mathcal{D}(\mathbf{x}'_t; \mathbf{x}_0)} \right]$$

$$\mathcal{L}_{\text{transform}}(\mathbf{x}_0; \boldsymbol{\theta}) = \sum_{ij \in \mathcal{G}(\mathbf{x})} \left\| \mathbf{t}_{ij}(\mathbf{x}_0) - \hat{\mathbf{t}}_{ij}^{\boldsymbol{\theta}}(\mathbf{x}_t) \right\|_2^2 + \left\| \mathbf{R}_{ij}(\mathbf{x}_0) - \hat{\mathbf{R}}_{ij}^{\boldsymbol{\theta}}(\mathbf{x}_t) \right\|_2^2 \quad \text{Inter-residue geometry}$$

# How they get there

- 1) The dataset
- 2) Protein-specific tricks: domain knowledge helps (reverse bitter lesson?)
  - a) Protein-structure-inspired separation of degrees-of-freedom (DOFs)
  - b) Protein-structure-inspired loss for protein substructure alignment
  - c) Protein-structure-based correlations in the initial noise**
  - d) Protein-structure-inspired selection of important DOFs
  - e) Physics-inspired approach to sampling more “low-energy” structures
  - f) Physics-inspired approach to AA orientation
  - g) Physics-inspired side-chains and simple sequence distributions
- 3) Ablation study (which parts are important)
- 4) Supported conditioning modes

Atom correlations: previous ideas, e.g. gas of AAs

AlphaFold2 (effectively):

$$\mathbf{R}_{\text{gas}} = \begin{bmatrix} \mathbf{R}_{\text{residue}} & & & \\ & \mathbf{R}_{\text{residue}} & & \\ & & \ddots & \\ & & & \mathbf{R}_{\text{residue}} \end{bmatrix}$$

# Atom correlations: incorporating known polymer statistics

Independent random walk (AAs):  $\mathbf{r}_{ij} \sim \mathcal{N}(0, a^2|i-j|)$

Correlated random walk:  $x_i = az_i + bx_{i-1}$

$$\text{Var}(x_1) = a^2 v^2 = \text{Var}(x_i) = \text{Var}(x_{i-1})$$

$$\mathbf{R} = a \begin{bmatrix} vb^0 & & & & \\ vb^1 & b^0 & & & \\ vb^2 & b^1 & b^0 & & \\ \vdots & & \ddots & \ddots & \\ vb^{N-2} & & & b^1 & b^0 \\ vb^{N-1} & \dots & b^2 & b^1 & b^0 \end{bmatrix}$$

$$v = \frac{1}{\sqrt{1-b^2}}$$

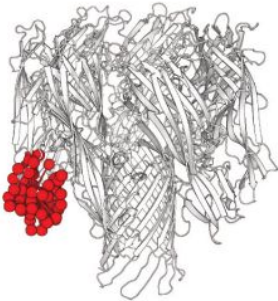
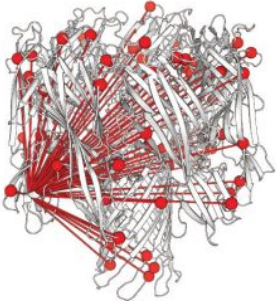
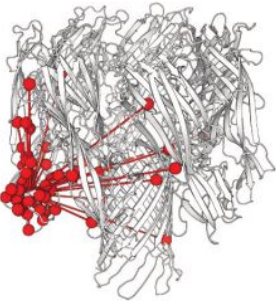
$$R_g = rN^v \quad (rN^v)^2 = \frac{Na^2}{6b + N(1-b^2)}$$

$$b_{\text{effective}}(N_{\text{atoms}}, a, r, v) = \frac{3}{N_{\text{atoms}}} + N_{\text{atoms}}^{-v} \sqrt{N_{\text{atoms}}^{2(v-1)} (N_{\text{atoms}}^2 + 9) - \frac{a^2}{r^2}},$$

# How they get there

- 1) The dataset
- 2) Protein-specific tricks: domain knowledge helps (reverse bitter lesson?)
  - a) Protein-structure-inspired separation of degrees-of-freedom (DOFs)
  - b) Protein-structure-inspired loss for protein substructure alignment
  - c) Protein-structure-based correlations in the initial noise
  - d) Protein-structure-inspired selection of important DOFs**
  - e) Physics-inspired approach to sampling more “low-energy” structures
  - f) Physics-inspired approach to AA orientation
  - g) Physics-inspired side-chains and simple sequence distributions
- 3) Ablation study (which parts are important)
- 4) Supported conditioning modes

# Which AA pairs are considered? Because “all” are too many

| Deterministic graph                                                                                                           | Random graphs                                                                     |                                                                                    |                                                                                     | Mixed graph                                                                         |
|-------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|
| k-NN                                                                                                                          | Uniform                                                                           | Exponential                                                                        | Inverse cubic                                                                       | 20 k-NN + 40 Inverse Cubic                                                          |
|                                              |  |  |  |  |
| Edge propensity<br>$p(\mathcal{E}_{ij} \in \mathcal{G}(\mathbf{x})   D_{ij}(\mathbf{x})) \propto$                             | constant                                                                          | $\exp\left(-\frac{D_{ij}(\mathbf{x})}{l}\right)$                                   | $\frac{1}{D_{ij}(\mathbf{x})^3}$                                                    |                                                                                     |
| Marginal distance propensity (uniform grid)<br>$p(D_{ij}(\mathbf{x})   \mathcal{E}_{ij} \in \mathcal{G}(\mathbf{x})) \propto$ | $D_{ij}(\mathbf{x})^2$                                                            | $D_{ij}(\mathbf{x})^2 \exp\left(-\frac{D_{ij}(\mathbf{x})}{l}\right)$              | $\frac{1}{D_{ij}(\mathbf{x})}$                                                      |                                                                                     |
| Long-range attachment                                                                                                         | ✓                                                                                 | ✓                                                                                  | ✓                                                                                   | ✓                                                                                   |
| Non-vanishing local attachment                                                                                                |                                                                                   | ✓                                                                                  | ✓                                                                                   | ✓                                                                                   |
| Monotonic decreasing distance propensity                                                                                      |                                                                                   |                                                                                    | ✓                                                                                   | ✓                                                                                   |

# How they get there

- 1) The dataset
- 2) Protein-specific tricks: domain knowledge helps (reverse bitter lesson?)
  - a) Protein-structure-inspired separation of degrees-of-freedom (DOFs)
  - b) Protein-structure-inspired loss for protein substructure alignment
  - c) Protein-structure-based correlations in the initial noise
  - d) Protein-structure-inspired selection of important DOFs
  - e) Physics-inspired approach to sampling more “low-energy” structures**
  - f) Physics-inspired approach to AA orientation
  - g) Physics-inspired side-chains and simple sequence distributions
- 3) Ablation study (which parts are important)
- 4) Supported conditioning modes

Sampling: uses the trained denoiser  $\mathbf{x}_0(\mathbf{x}_t, t)$

$$\mathbf{x} = F(\mathbf{z}) = \mathbf{R}\mathbf{z} + \mathbf{b} \quad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

Variance-preserving:

$$\sigma_t^2 = 1 - \alpha_t^2 \quad \frac{\alpha'_t}{\alpha_t} \triangleq -\frac{1}{2}\beta_t$$

$$d\mathbf{x} = -\frac{1}{2}\beta_t \mathbf{x} dt + \sqrt{\beta_t} \mathbf{R} d\mathbf{w}$$



$$d\mathbf{x} = h_t \mathbf{x} dt + g_t \mathbf{R} d\mathbf{w}$$

$$h_t = \frac{\alpha'_t}{\alpha_t}, \quad g_t = \sqrt{2\sigma_t^2 \left( \frac{\sigma'_t}{\sigma_t} - \frac{\alpha'_t}{\alpha_t} \right)},$$

$$p(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}(\mathbf{x}; \alpha_t \mathbf{x}_0, \sigma_t^2 \mathbf{R} \mathbf{R}^\top)$$

SDE: 
$$d\mathbf{x} = \left( -\frac{1}{2} \mathbf{x} - \mathbf{R} \mathbf{R}^\top \nabla_{\mathbf{x}} \log p_t(\mathbf{x}) \right) \beta_t dt + \sqrt{\beta_t} \mathbf{R} d\bar{\mathbf{w}} \quad \text{Tweedie relationship}$$

ODE: 
$$\frac{d\mathbf{x}}{dt} = -\frac{\beta_t}{2} (\mathbf{x} + \mathbf{R} \mathbf{R}^\top \nabla_{\mathbf{x}} \log p_t(\mathbf{x}))$$

$$\nabla_{\mathbf{x}} \log p_t(\mathbf{x}) = (\sigma_t^2 \mathbf{R} \mathbf{R}^\top)^{-1} (\alpha_t \hat{\mathbf{x}}_\theta(\mathbf{x}, t) - \mathbf{x})$$

Inference-time steering

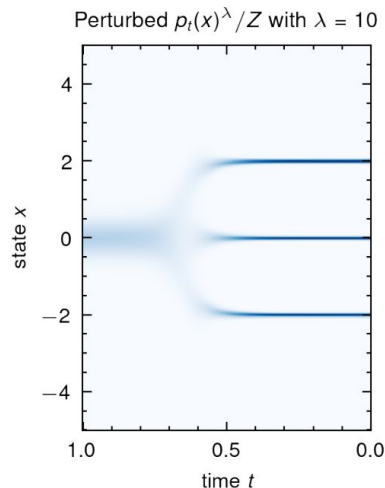
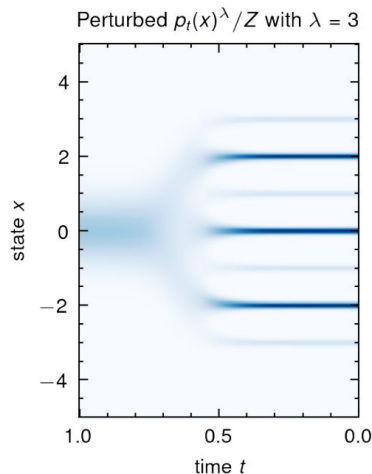
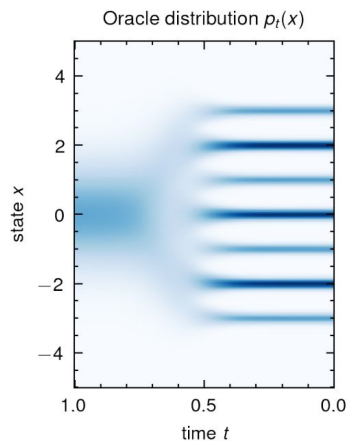
$$d\mathbf{x} = \left( -\frac{1}{2} \mathbf{x} - \mathbf{R} \mathbf{R}^\top (\nabla_{\mathbf{x}} \log p_t(\mathbf{x}) + \nabla_{\mathbf{x}} \log p_t(\mathbf{y} | \mathbf{x})) \right) \beta_t dt + \sqrt{\beta_t} \mathbf{R} d\bar{\mathbf{w}}$$

$$\frac{d\mathbf{x}}{dt} = -\frac{\beta_t}{2} (\mathbf{x} + \mathbf{R} \mathbf{R}^\top (\nabla_{\mathbf{x}} \log p_t(\mathbf{x}) + \nabla_{\mathbf{x}} \log p_t(\mathbf{y} | \mathbf{x})))$$

Classifier

# Low-temperature sampling

$$p_\lambda(\mathbf{x}) = \frac{1}{Z} p(\mathbf{x})^\lambda \quad \lambda \rightarrow \infty$$



What (naive) approach does not work: constant score rescaling

$$d\mathbf{x} = \left( -\frac{1}{2} \mathbf{x} - \lambda_0 \mathbf{R} \mathbf{R}^\top \nabla_{\mathbf{x}} \log p_t(\mathbf{x}) \right) \beta_t dt + \sqrt{\beta_t} \mathbf{R} d\bar{\mathbf{w}}$$

# Working approach 1 (probably motivated the naive approach): Near-equilibrium overdamped langevin sampling

Not good: 
$$d\mathbf{x} = \left( -\frac{1}{2} \mathbf{x} - \lambda_0 \mathbf{R} \mathbf{R}^\top \nabla_{\mathbf{x}} \log p_t(\mathbf{x}) \right) \beta_t dt + \sqrt{\beta_t} \mathbf{R} d\bar{\mathbf{w}}$$

Good: 
$$d\mathbf{x} = -\frac{\beta_t \psi}{2} \lambda_0 \mathbf{R} \mathbf{R}^\top \nabla_{\mathbf{x}} \log p_t(\mathbf{x}) dt + \sqrt{\beta_t \psi} \mathbf{R} d\bar{\mathbf{w}}$$

Why good works: **Langevin Monte Carlo** [\[edit\]](#)

The [overdamped Langevin equation](#) gives

$$d\mathbf{x}_t = -\frac{D}{k_B T} \nabla_{\mathbf{x}} U(\mathbf{x}_t) dt + \sqrt{2D} dW_t$$

## Working approach 2: Time-dependent score-scaling

transforming a Gaussian data distribution  $\mathcal{N}(\mathbf{x}_0; \boldsymbol{\mu}_{\text{data}}, \boldsymbol{\sigma}_{\text{data}}^2)$  to a Gaussian prior  $\mathcal{N}(\mathbf{x}_1; 0, \boldsymbol{\sigma}_{\text{prior}}^2)$

$$p_t(\mathbf{x}) = \mathcal{N}\left(\mathbf{x}; \alpha_t \boldsymbol{\mu}_{\text{data}}, \alpha_t^2 \boldsymbol{\sigma}_{\text{data}}^2 + (1 - \alpha_t^2) \boldsymbol{\sigma}_{\text{prior}}^2\right)$$

$$\mathbf{s}_t \triangleq \nabla_{\mathbf{x}} \log p_t(\mathbf{x}) = \frac{\alpha_t \boldsymbol{\mu}_{\text{data}} - \mathbf{x}}{\alpha_t^2 \boldsymbol{\sigma}_{\text{data}}^2 + (1 - \alpha_t^2) \boldsymbol{\sigma}_{\text{prior}}^2}$$

transitions to  $\frac{1}{Z_{\text{lambda}0}} p_0(\mathbf{x})^{\lambda_0}$

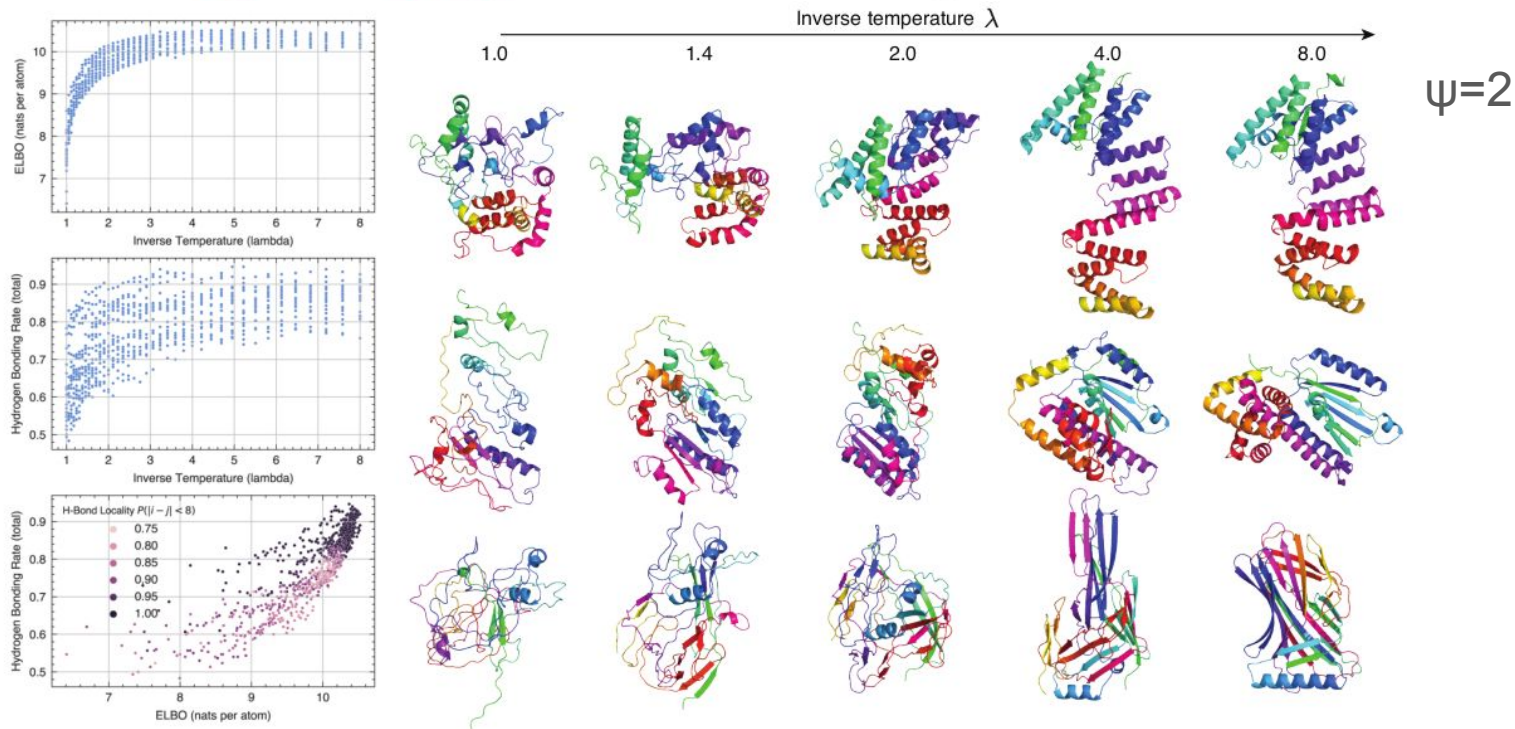
$$\mathbf{s}_t^{\text{perturb}} = \mathbf{s}_t \frac{(1 - \alpha_t^2) \boldsymbol{\sigma}_{\text{prior}}^2 + \alpha_t^2 \boldsymbol{\sigma}_{\text{data}}^2}{(1 - \alpha_t^2) \boldsymbol{\sigma}_{\text{prior}}^2 + \alpha_t^2 \boldsymbol{\sigma}_{\text{data}}^2 / \lambda_0} = \mathbf{s}_t \lambda_t$$

**Good:**

$$d\mathbf{x} = \left( -\frac{1}{2} \mathbf{x} - \lambda_t \mathbf{R} \mathbf{R}^T \nabla_{\mathbf{x}} \log p_t(\mathbf{x}) \right) \beta_t dt + \sqrt{\beta_t} \mathbf{R} d\bar{\mathbf{w}}$$

They use both (again, physics-informed) and engineer the mix

$$d\mathbf{x} = \left( -\frac{1}{2} \mathbf{x} - \left( \lambda_t + \frac{\lambda_0 \psi}{2} \right) \mathbf{R} \mathbf{R}^\top \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t) \right) \beta_t dt + \sqrt{\beta_t (1 + \psi)} \mathbf{R} d\bar{\mathbf{w}}$$

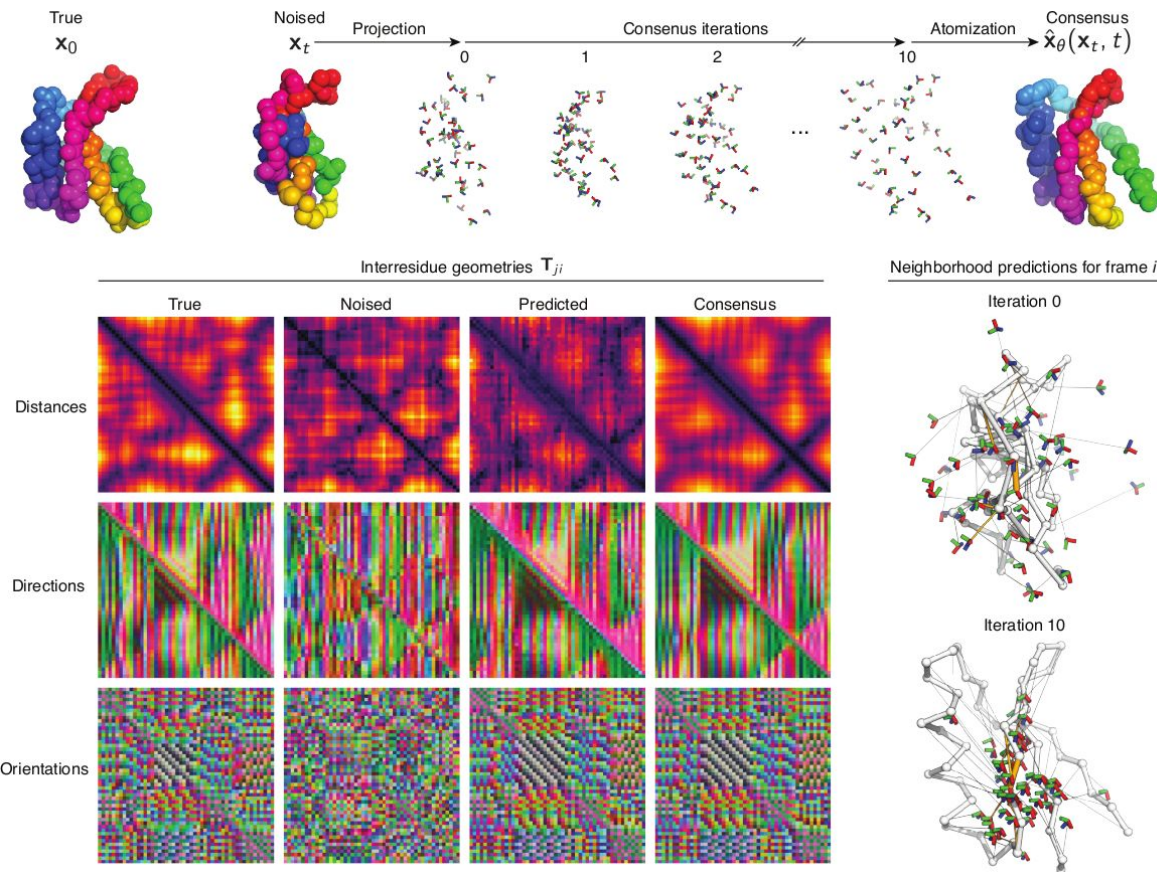


# How they get there

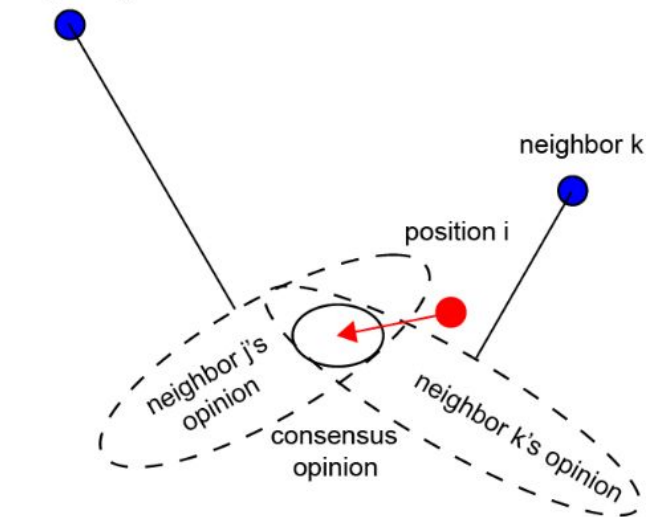
- 1) The dataset
- 2) Protein-specific tricks: domain knowledge helps (reverse bitter lesson?)
  - a) Protein-structure-inspired separation of degrees-of-freedom (DOFs)
  - b) Protein-structure-inspired loss for protein substructure alignment
  - c) Protein-structure-based correlations in the initial noise
  - d) Protein-structure-inspired selection of important DOFs
  - e) Physics-inspired approach to sampling more “low-energy” structures
  - f) Physics-inspired approach to AA orientation**
  - g) Physics-inspired side-chains and simple sequence distributions
- 3) Ablation study (which parts are important)
- 4) Supported conditioning modes

# Denoiser representation

$$\nabla_{\mathbf{x}} \log p_t(\mathbf{x}) = (\sigma_t^2 \mathbf{R} \mathbf{R}^\top)^{-1} (\alpha_t \hat{\mathbf{x}}_\theta(\mathbf{x}, t) - \mathbf{x})$$



neighbor j



# How they get there

- 1) The dataset
- 2) Protein-specific tricks: domain knowledge helps (reverse bitter lesson?)
  - a) Protein-structure-inspired separation of degrees-of-freedom (DOFs)
  - b) Protein-structure-inspired loss for protein substructure alignment
  - c) Protein-structure-based correlations in the initial noise
  - d) Protein-structure-inspired selection of important DOFs
  - e) Physics-inspired approach to sampling more “low-energy” structures
  - f) Physics-inspired approach to AA orientation
  - g) Physics-inspired side-chains and simple sequence distributions**
- 3) Ablation study (which parts are important)
- 4) Supported conditioning modes

## Sequence and side-chains: simple/empirical distributions

$$\log p(\mathbf{x}, \mathbf{s}, \boldsymbol{\chi}) = \underbrace{\log p(\mathbf{x})}_{\text{backbone likelihood}} + \underbrace{\log p(\mathbf{s}|\mathbf{x})}_{\text{sequence likelihood}} + \underbrace{\log p(\boldsymbol{\chi}|\mathbf{x}, \mathbf{s})}_{\text{side-chain likelihood}}$$

$$p_{\theta}(\mathbf{s}|\mathbf{x}) = \frac{1}{Z(\mathbf{x}, \boldsymbol{\theta})} \exp \left( - \sum_i \mathbf{h}_i(s_i; \mathbf{x}) - \sum_{i < j} \mathbf{J}_{ij}(s_i, s_j; \mathbf{x}) \right) \quad \text{+MCMC sampling}$$

$$p_{\theta}(\boldsymbol{\chi}|\mathbf{s}, \mathbf{x}) = \prod_i p_{\theta}(\chi_{\pi_i} | \chi_{\pi_{i-1}}, \dots, \chi_{\pi_1}, \mathbf{s}, \mathbf{x}),$$

where the conditional joint distributions  $p_{\theta}(\chi_{\pi_i} | \chi_{\pi_{i-1}}, \dots, \chi_{\pi_1}, \mathbf{s}, \mathbf{x})$  at each residue locally factorize as up to four discrete, sequential decisions as in [113]. We model these with empirical histograms for each angular degree of freedom binned at 36 bins, i.e. with  $10^\circ$  angular resolu-

# How they get there

- 1) The dataset
- 2) Protein-specific tricks: domain knowledge helps (reverse bitter lesson?)
  - a) Protein-structure-inspired separation of degrees-of-freedom (DOFs)
  - b) Protein-structure-inspired loss for protein substructure alignment
  - c) Protein-structure-based correlations in the initial noise
  - d) Protein-structure-inspired selection of important DOFs
  - e) Physics-inspired approach to sampling more “low-energy” structures
  - f) Physics-inspired approach to AA orientation
  - g) Physics-inspired side-chains and simple sequence distributions
- 3) Ablation study (which parts are important)**
- 4) Supported conditioning modes

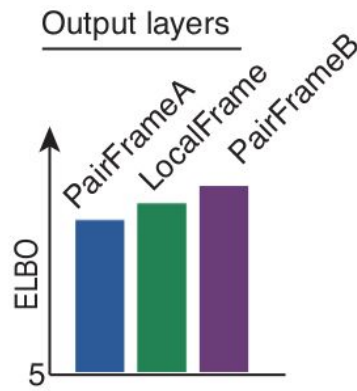
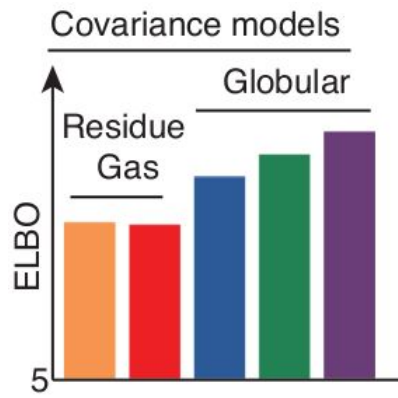
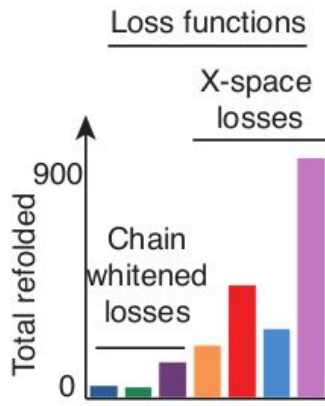
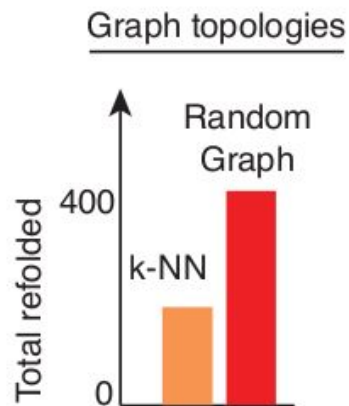
# Ablation studies (how important the components are)

Far-connections if  
atom-pair graphs:  
Pretty important

Loss in X-space  
(not z-space):  
Pretty important

Rg-statistics:  
somewhat important

Rg-statistics:  
somewhat important



\*1) Heuristic losses help refolding (19% vs 2%)

\*2)  $\lambda=10, \psi=2$  low-Temp sampling helps a lot

# How they get there

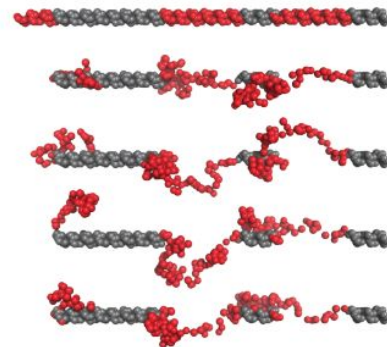
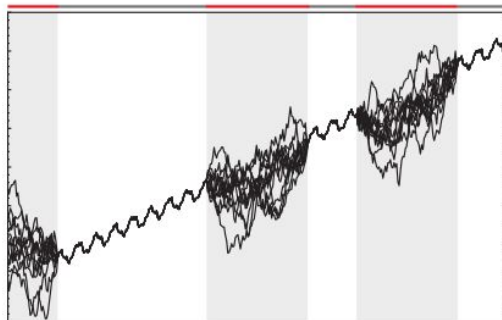
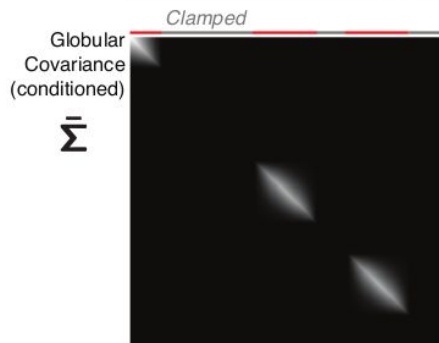
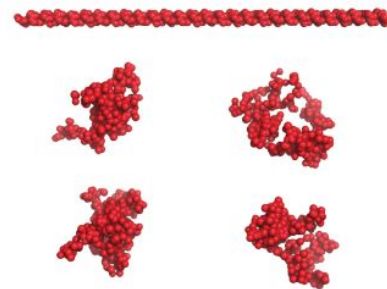
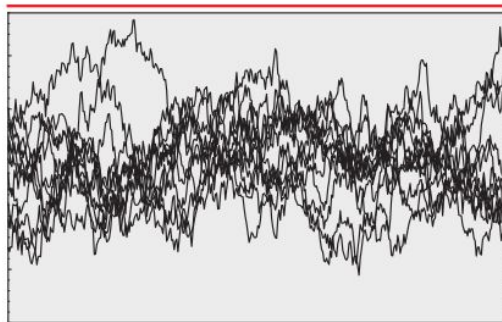
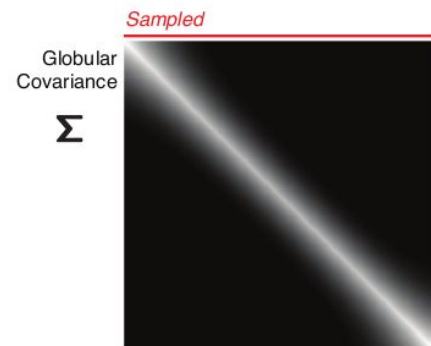
- 1) The dataset
- 2) Protein-specific tricks: domain knowledge helps (reverse bitter lesson?)
  - a) Protein-structure-inspired separation of degrees-of-freedom (DOFs)
  - b) Protein-structure-inspired loss for protein substructure alignment
  - c) Protein-structure-based correlations in the initial noise
  - d) Protein-structure-inspired selection of important DOFs
  - e) Physics-inspired approach to sampling more “low-energy” structures
  - f) Physics-inspired approach to AA orientation
  - g) Physics-inspired side-chains and simple sequence distributions
- 3) Ablation study (which parts are important)
- 4) **Supported conditioning modes**

# Inference-time guidance

$$\begin{aligned}\nabla_{\mathbf{x}} \log p_t(\mathbf{x}|\mathbf{y}) &= \nabla_{\mathbf{x}} \log \frac{p_t(\mathbf{x})p_t(\mathbf{y}|\mathbf{x})}{p_t(\mathbf{y})} \\ &= \nabla_{\mathbf{x}} \log p_t(\mathbf{x}) + \nabla_{\mathbf{x}} \log p_t(\mathbf{y}|\mathbf{x}) - \cancel{\nabla_{\mathbf{x}} \log p_t(\mathbf{y})}\end{aligned}$$

Backbone structure conditioning:

$$\log p(D_0^{ij} < C|\mathbf{x}_t, t) = \log \left( 1 - Q_{\frac{3}{2}} \left( \frac{\sqrt{\alpha_t} d_t^{ij}}{\sigma_t \sigma_{ij}}, \frac{C}{\sigma_t \sigma_{ij}} \right) \right)$$



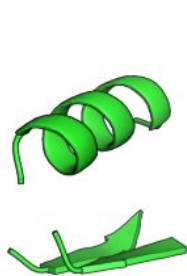
# Inference-time guidance

$$\begin{aligned}\nabla_{\mathbf{x}} \log p_t(\mathbf{x}|\mathbf{y}) &= \nabla_{\mathbf{x}} \log \frac{p_t(\mathbf{x})p_t(\mathbf{y}|\mathbf{x})}{p_t(\mathbf{y})} \\ &= \nabla_{\mathbf{x}} \log p_t(\mathbf{x}) + \nabla_{\mathbf{x}} \log p_t(\mathbf{y}|\mathbf{x}) - \cancel{\nabla_{\mathbf{x}} \log p_t(\mathbf{y})}\end{aligned}$$

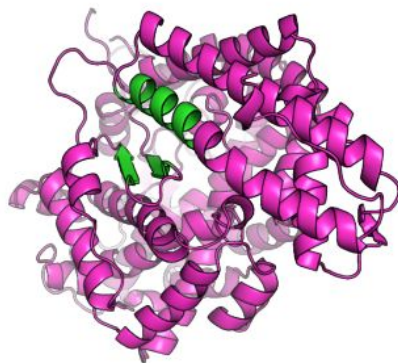
Motif conditioning:

$$\rho(\mathbf{x}_t) = \min_{\pi \in \Pi} \min_{\mathbf{T} \in \text{SE}(3)} \frac{\|\mathbf{x}_t^{\mathcal{M}} - \mathbf{T} \circ \hat{\mathbf{x}}_{\theta}(\mathbf{x}_t, t)^{\mathcal{M}_{\pi}}\|_2}{\sqrt{|\mathcal{M}_{\pi}|}}$$

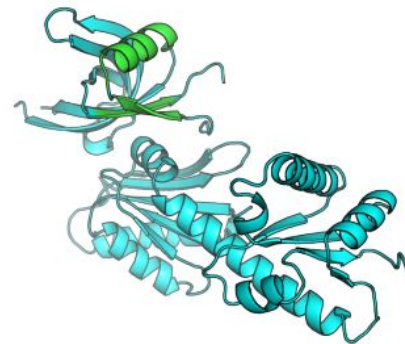
$$U_f(\mathbf{x}_t, U, t) = U + \eta \log \left( 1 + e^{\zeta[\rho(\mathbf{x}_t) - \rho_{\max}]} \right)$$



(a)



(b)



(c)

Supplementary Figure 26: **Motifs can occur in entirely unrelated structural contexts.** **a**, An example motif composed of three disjoint segments. **b**, PDB entry 3NXQ harbors the motif with a backbone RMSD of 0.45 Å. **c**, PDB entry 3OBW harbors the motif with a backbone RMSD of 0.64 Å.

# Inference-time guidance

Symmetry conditioning:

$$\begin{aligned}\nabla_{\mathbf{x}} \log p_t(\mathbf{x}|\mathbf{y}) &= \nabla_{\mathbf{x}} \log \frac{p_t(\mathbf{x})p_t(\mathbf{y}|\mathbf{x})}{p_t(\mathbf{y})} \\ &= \nabla_{\mathbf{x}} \log p_t(\mathbf{x}) + \nabla_{\mathbf{x}} \log p_t(\mathbf{y}|\mathbf{x}) - \cancel{\nabla_{\mathbf{x}} \log p_t(\mathbf{y})}\end{aligned}$$
$$\left\{ \begin{array}{l} \mathbf{x} = \text{subsample}(\text{symmetrize}(\tilde{\mathbf{x}})) \\ U_f(\mathbf{x}, U, t) = U + U_{R_g}(\tilde{\mathbf{x}}) + U_{R_g}(\text{subsample}(\text{symmetrize}(\tilde{\mathbf{x}}))) \end{array} \right.$$

$C_4$



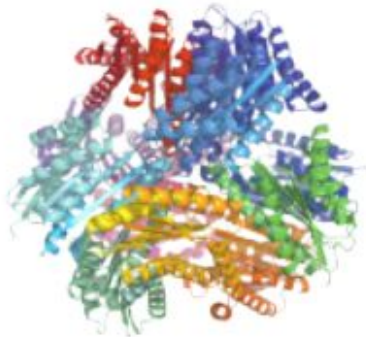
Cyclic

$[C_3, C_5]$

Composed



T



Tetrahedral

# Inference-time guidance

Label conditioning:

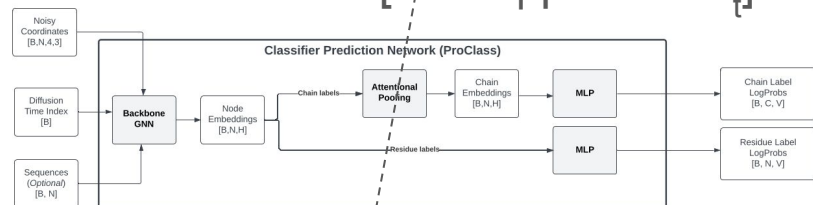
$$\nabla_{\mathbf{x}} \log p_t(\mathbf{x}|\mathbf{y}) = \nabla_{\mathbf{x}} \log \frac{p_t(\mathbf{x})p_t(\mathbf{y}|\mathbf{x})}{p_t(\mathbf{y})}$$

$$= \nabla_{\mathbf{x}} \log p_t(\mathbf{x}) + \nabla_{\mathbf{x}} \log p_t(\mathbf{y}|\mathbf{x}) - \cancel{\nabla_{\mathbf{x}} \log p_t(\mathbf{y})}$$

## S.5 Labels and loss functions

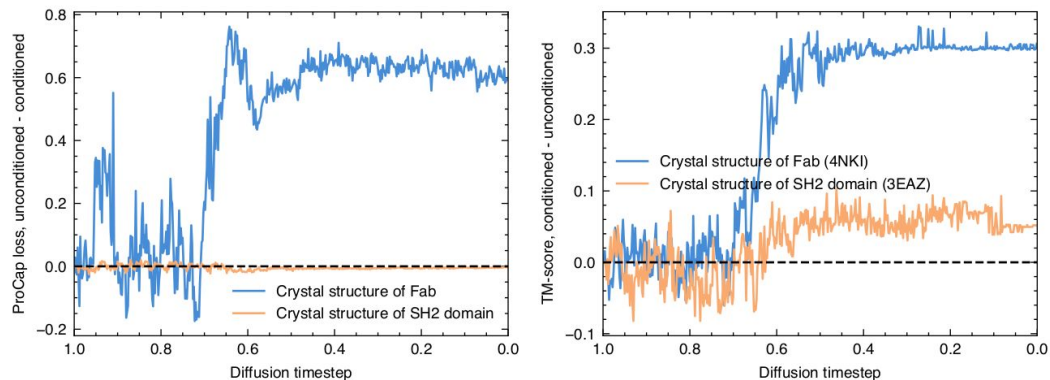
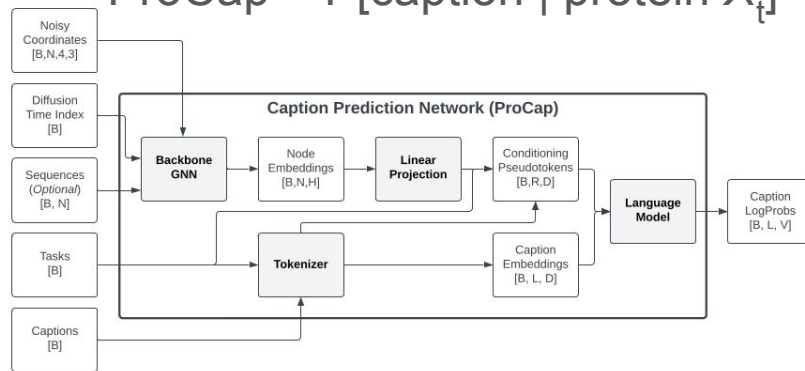
We trained the model to predict the following labels: CATH, PFAM, Funfam, Organism, Secondary Structure, Interfacial Residue. We quantified the loss for each label prediction using cross entropy, and summed all components with equal weights.

$$\text{ProClass} = \text{P}[\text{class} \mid \text{protein } X_t]$$



Natural language (GPT-Neo embeddings) conditioning:

$$\text{ProCap} = \text{P}[\text{caption} \mid \text{protein } X_t]$$



# Workbench

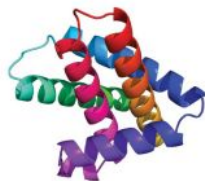
# Chroma: Controllable (bio)physics-informed generation of proteins



C<sub>6</sub>



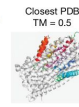
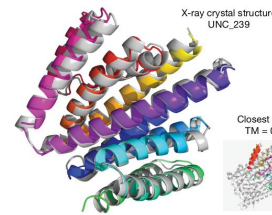
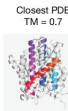
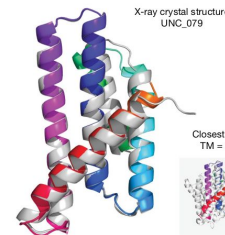
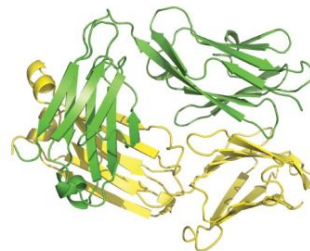
$\alpha$   
C<sub>1</sub>



$\beta$   
C<sub>2</sub>



“Crystal structure of a  
Fab antibody fragment”



Forward diffusion: gaussian, but with biophysically motivated correlations

$$\mathbf{x}_t \sim p(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}; \alpha_t \mathbf{x}_0, (1 - \alpha_t^2) \mathbf{R} \mathbf{R}^\top) \quad d\mathbf{x} = -\frac{1}{2} \beta_t \mathbf{x} dt + \sqrt{\beta_t} d\mathbf{R} \mathbf{w} \quad -\frac{1}{2} \beta_t = \frac{d \log \alpha_t}{dt}$$

$\mathbf{R} \mathbf{R}^\top$  respect the distance statistics of natural proteins

**Diffusion objective** Given this forwards process, we train a neural network  $\hat{\mathbf{x}}_\theta(\mathbf{x}_t, t)$  to predict the optimally denoised structure by optimizing a bound on the likelihood of the noise-free structure  $\mathbf{x}_0$ ,

$$\log p(\mathbf{x}_0) \geq -\frac{1}{2} \log \det(2\pi e \mathbf{R} \mathbf{R}^\top) - \frac{1}{2} \mathbb{E}_{p(\mathbf{x}_t|\mathbf{x}_0)p(t)} \left[ \text{SNR}'_t \left( \frac{N}{1 + \text{SNR}_t} - \|\mathbf{R}^{-1}(\hat{\mathbf{x}}_\theta(\mathbf{x}_t, t) - \mathbf{x}_0)\|_2^2 \right) \right],$$

$$\text{SNR}_t = \frac{\alpha_t^2}{1 - \alpha_t^2}$$

$$d\mathbf{x} = -\frac{1}{2} \beta_t \mathbf{x} - \left( \lambda_t + \frac{\lambda_0 \psi}{2} \right) \beta_t \mathbf{R} \mathbf{R}^\top \nabla_{\mathbf{x}} \log p_t(\mathbf{x}; \theta) dt + \sqrt{\beta_t(1 + \psi)} \mathbf{R} d\bar{\mathbf{w}}$$