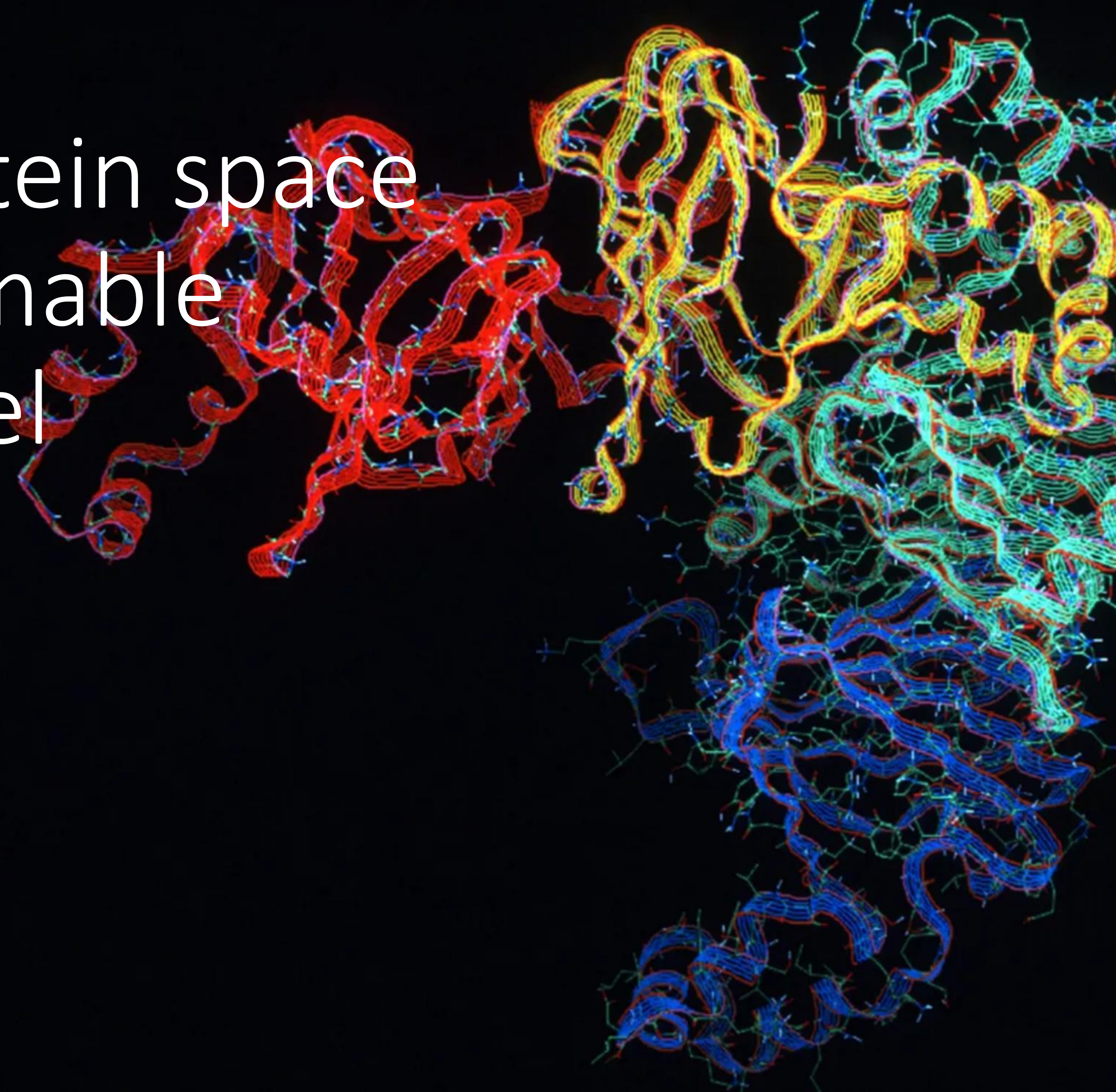


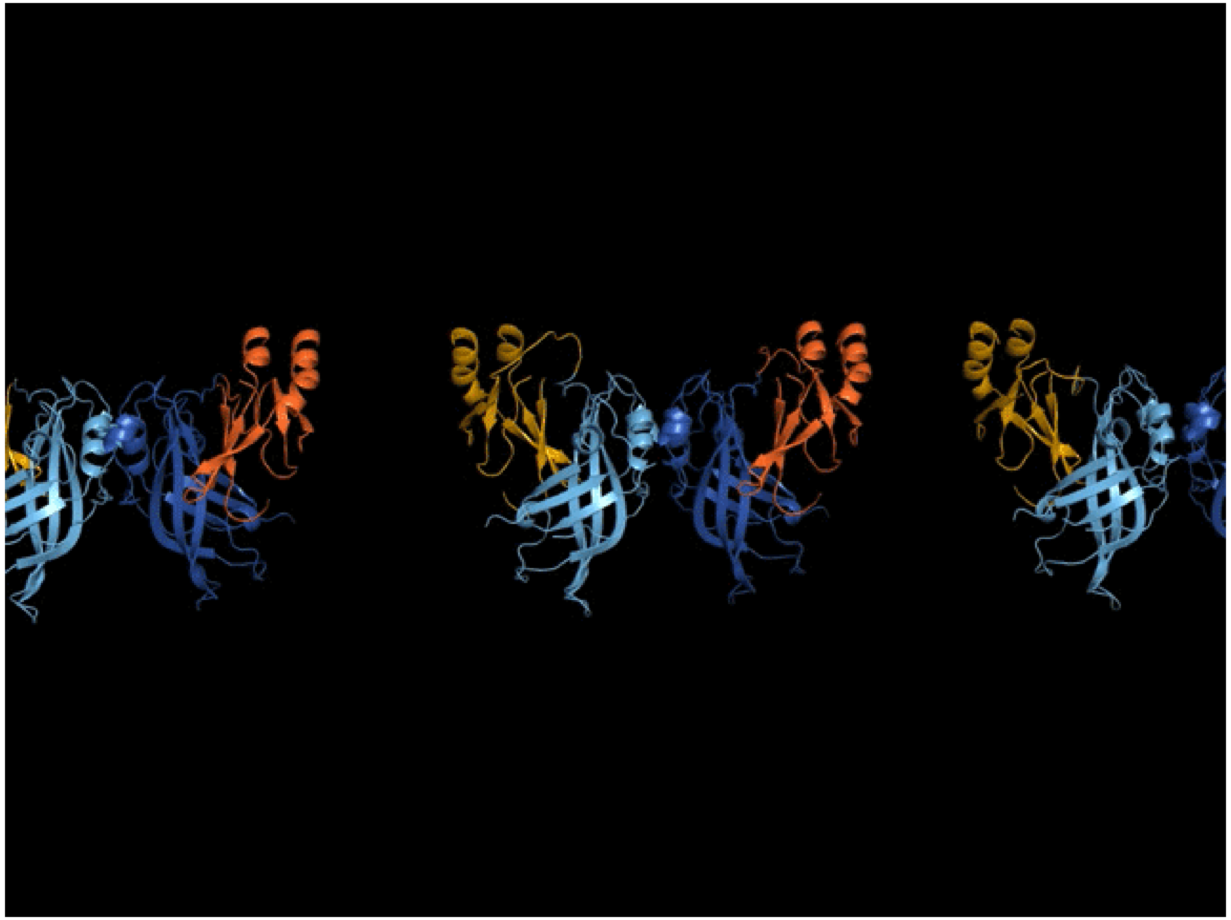
Illuminating protein space with a programmable generative model

J Ingraham, M Baranov, Z Costello, V Frappier,
A Ismail, S Tie, W Wang, V Xue, F Obermeyer,
A Beam, G Grigoryan

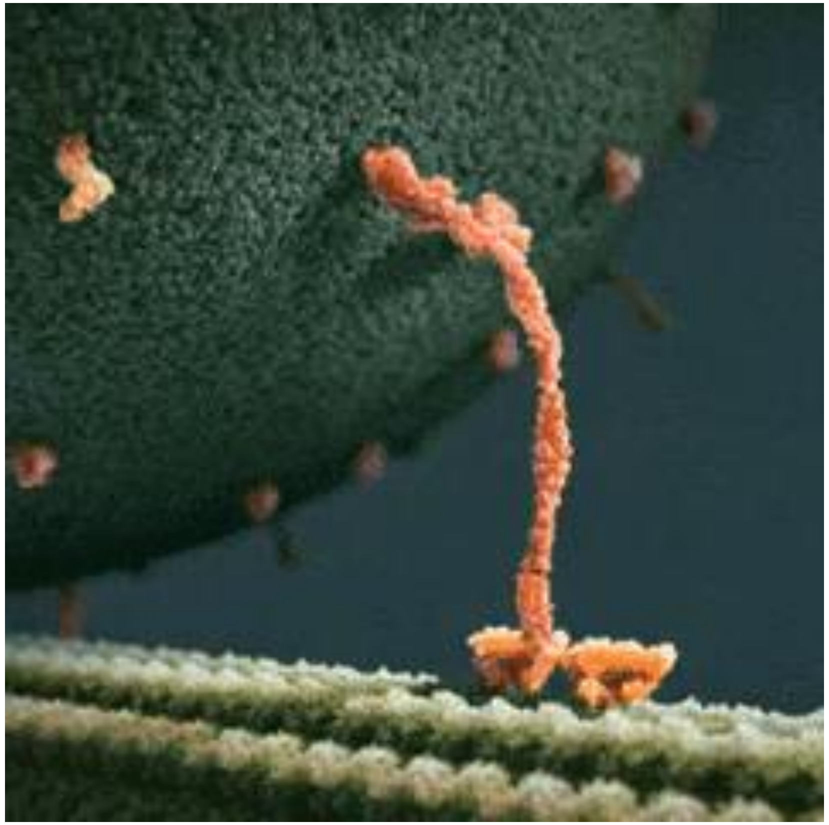
December 1, 2022

Presented by Meret Ackermann

Proteins – the chief actors in cells

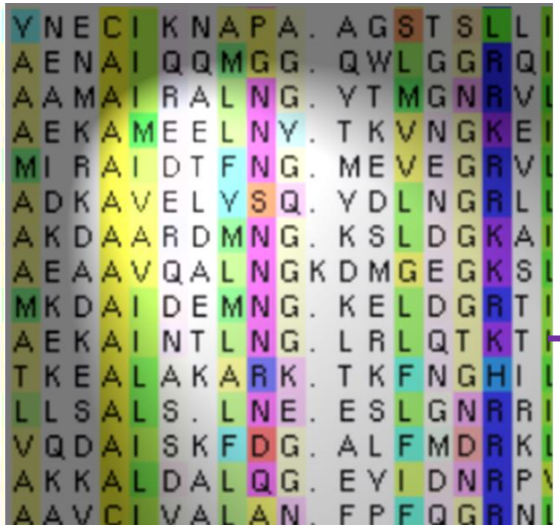


Structure – the key component of function



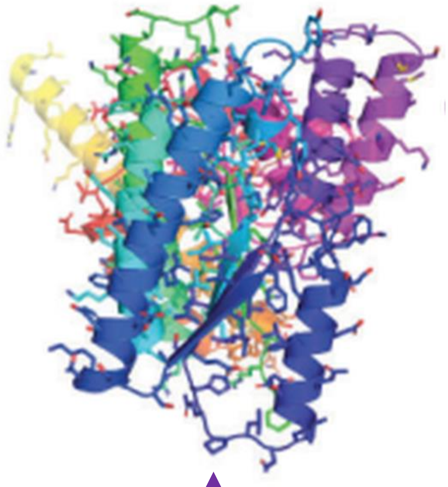
Protein Structure Prediction

Sequence



```
VNECIKNAPA.AGSTSLL
AENAIQQMGG.QWLGGRQ
AAMAIRALNG.YTMGNRV
AEKAMEELNY.TKVNGKE
MIRAIDTFNG.MEVEGRV
ADKAVELYSQ.YDLNGRL
AKDAARDMNG.KSLDGKA
AEAAVQALNGKDMGEGKS
MKDAIDEMNG.KELDGRT
AEKAINTLNG.LRLQTKT
TKEALAKARK.TKFNGHI
LLSALS.LNE.ESLGNRR
VQDAISKFDG.ALFMDRK
AKKALDALQG.EYIDNRP
AAVCIVALAN.FPFQGRN
```

Folding



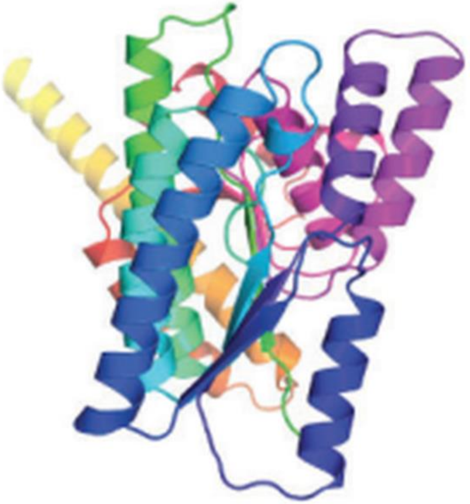
AlphaFold

RosettaFold

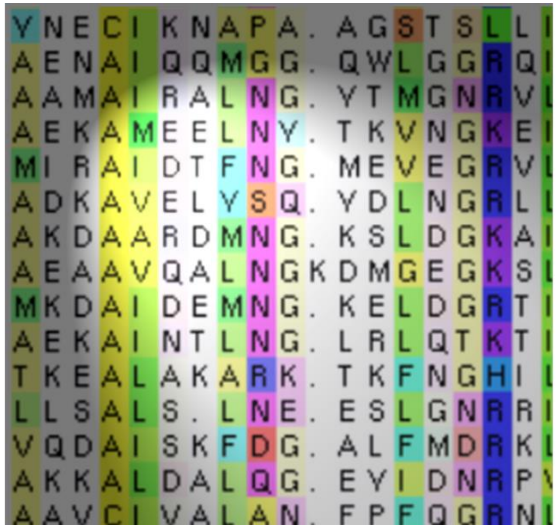
…

Protein Sequence prediction

Structure



Folding

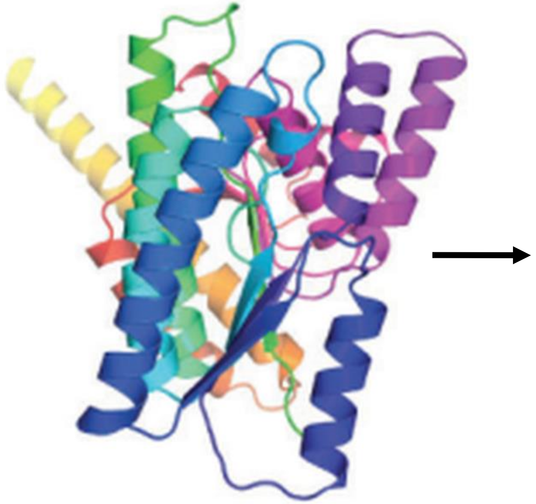


```
Y N E C I K N A P A . A G S T S L L I
A E N A I Q Q M G G . Q W L G G R Q I
A A M A I R A L N G . Y T M G N R V L
A E K A M E E L N Y . T K V N G K E I
M I R A I D T F N G . M E V E G R V L
A D K A V E L Y S Q . Y D L N G R L L
A K D A A R D M N G . K S L D G K A I
A E A A V Q A L N G K D M G E G K S L
M K D A I D E M N G . K E L D G R T I
A E K A I N T L N G . L R L Q T K T I
T K E A L A K A R K . T K F N G H I L
L L S A L S . L N E . E S L G N R R I
V Q D A I S K F D G . A L F M D R K L
A K K A L D A L Q G . E Y I D N R P V
A A V C I V A L A N . F P F Q G R N L
```

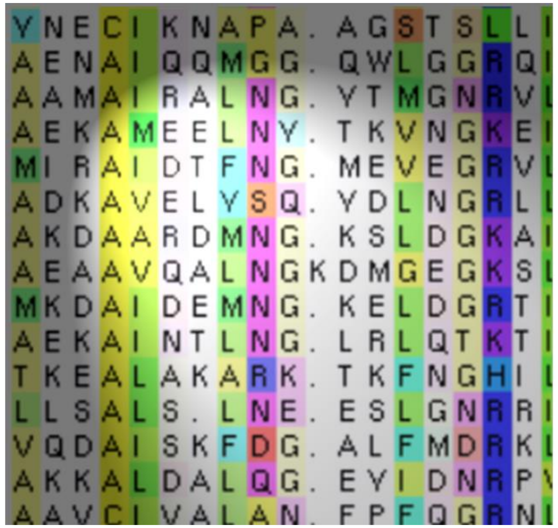
Protein MPNN
GVP-GNN
…

Structure first – Protein Design

Backbone

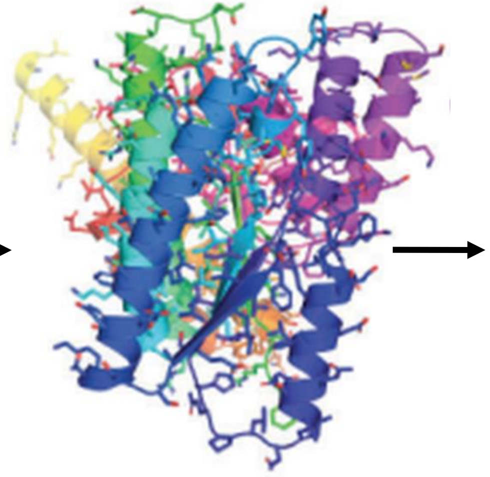


Sequence



```
VNECIKNAPA.AGSTSLLI
AENAIQQMGG.QWLGGRQI
AAMAIRALNG.YTMGNRVL
AEKAMEELNY.TKVNGKEI
MIRAIDTFNG.MEVEGRVL
ADKAVELYSQ.YDLNGRLL
AKDAARDMNG.KSLDGKAI
AEAAVQALNGKDMGEGKSL
MKDAIDEMNG.KELDGRTI
AEKAINTLNG.LRLQTKTI
TKEALAKARK.TKFNGHIL
LLSALS.LNE.ESLGNRRI
VQDAISKFDG.ALFMDRKL
AKKALDALQG.EYIDNRPV
AAVCIVALAN.FPFQGRNL
```

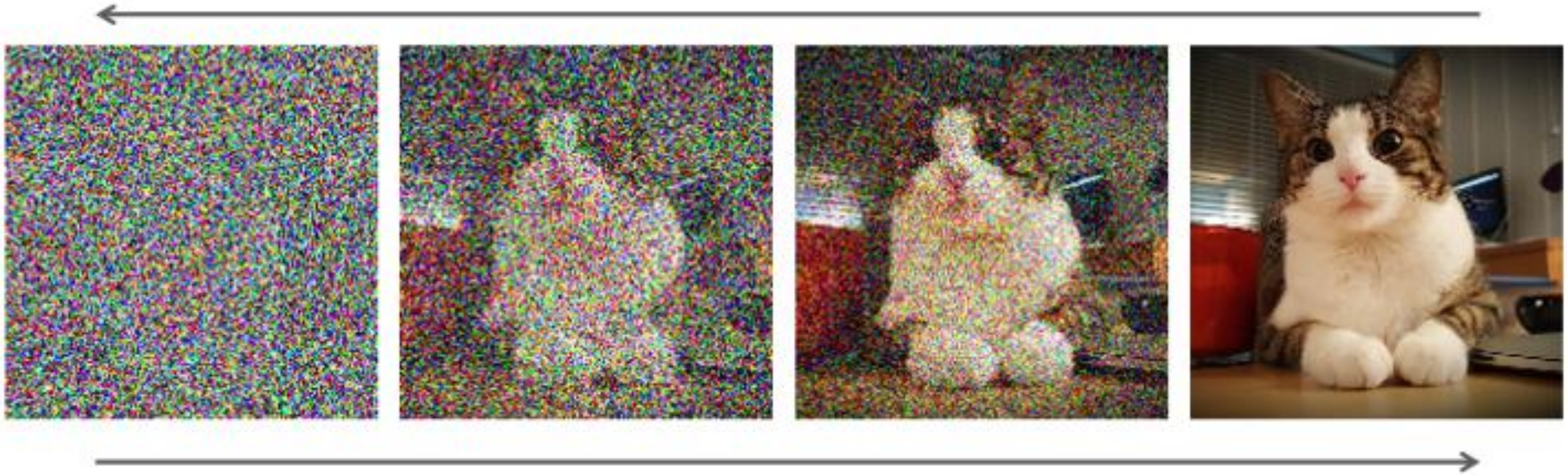
Side Chains



Experimental characterization

Generative diffusion process

Forward SDE. **Training**. Data to Noise.

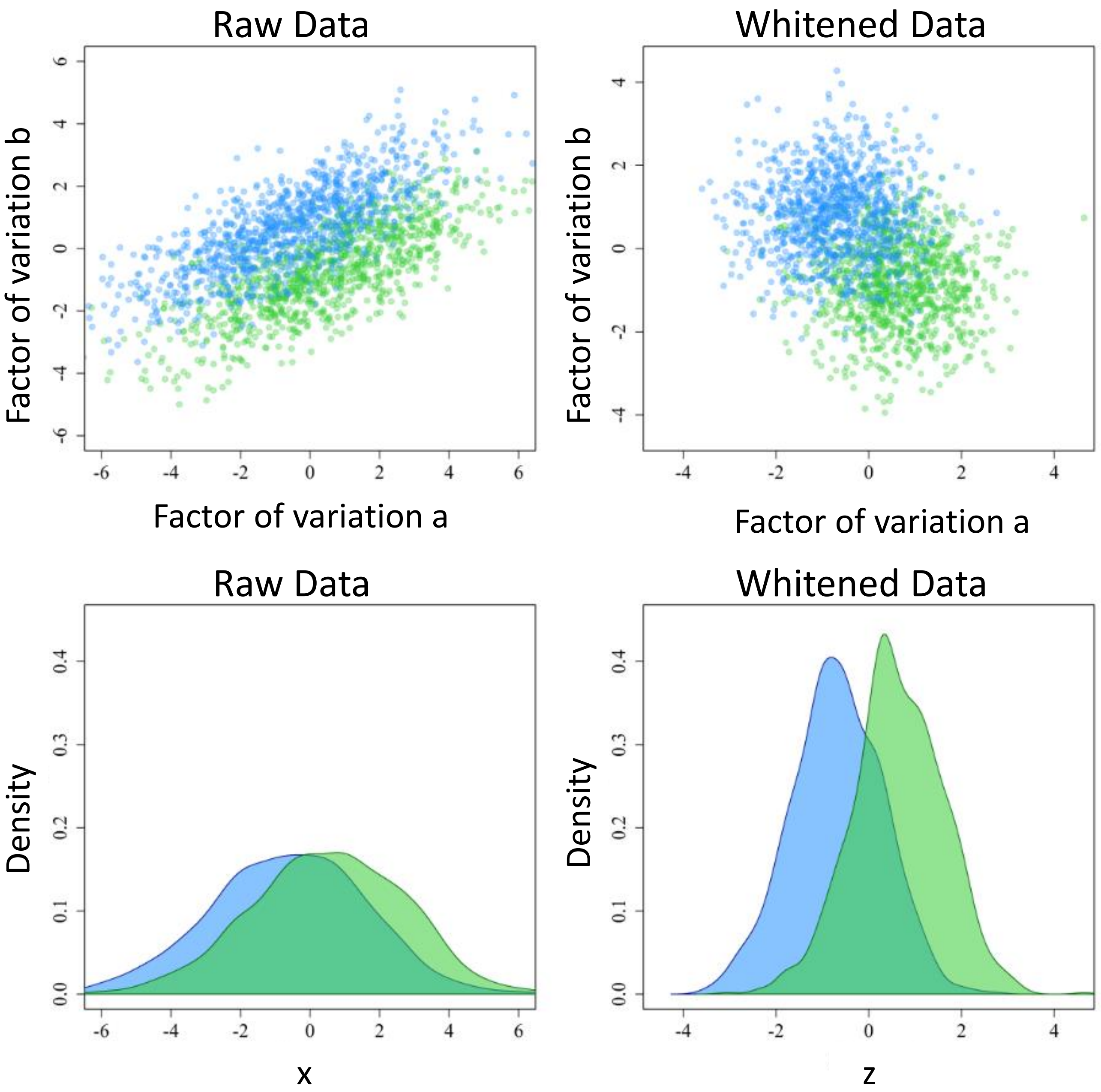


Reverse SDE. **Generation**. Noise to Data

Whitening transformation

Correlated diffusion
as uncorrelated diffusion
in whitened space

$$z = R(x - \mu)$$



Raw Data

Whitened Data

Raw Data

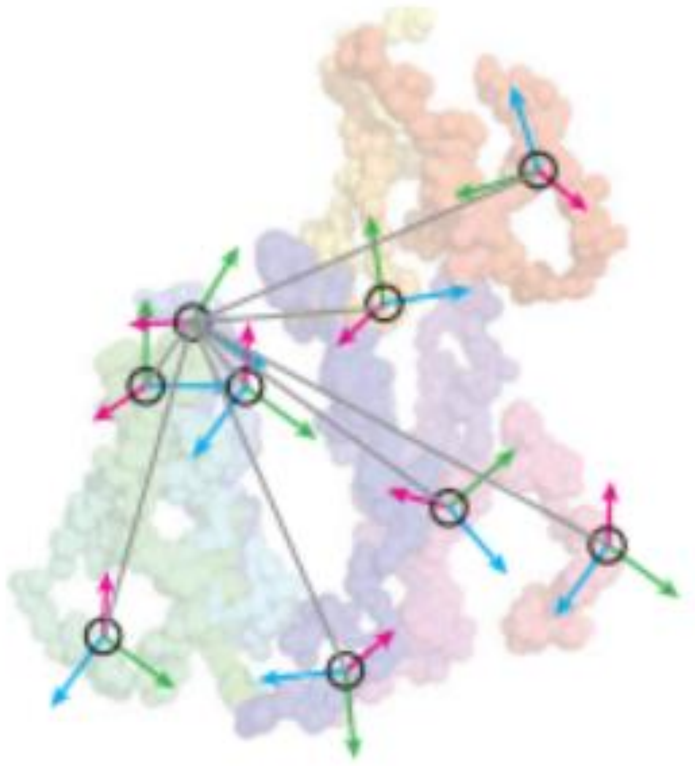
Whitened Data

Correlated forward SDE process

$$d\boldsymbol{x} = \boldsymbol{R}d\boldsymbol{z} = -\frac{\beta_t}{2}\boldsymbol{R}\boldsymbol{z}dt + \sqrt{\beta_t}\boldsymbol{R}dw$$

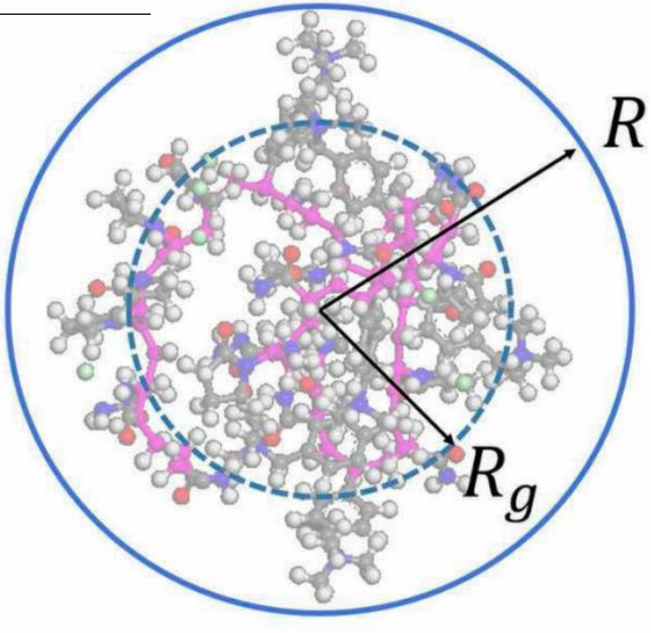
$\mathbf{x}_1$
$\mathbf{x}_t$
$\mathbf{x}_0$

Constraints as a de-whitening transform

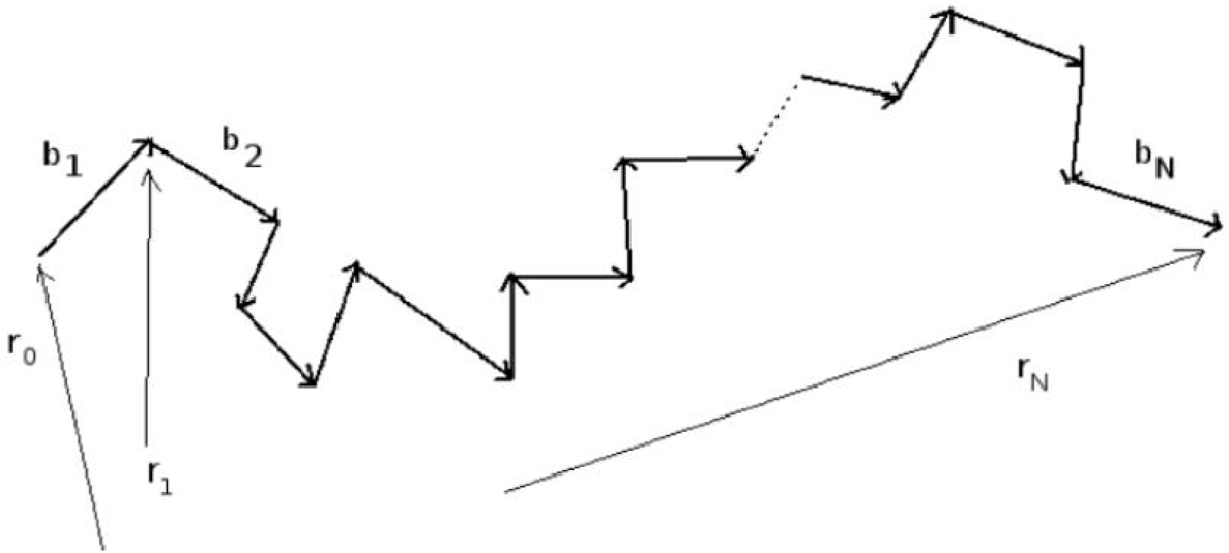


$$F(x) = \sum_{i,j} \boldsymbol{A}_{i,j}\boldsymbol{x}_i\boldsymbol{x}_j$$

$$\mathbb{E}_{p(\boldsymbol{x_t}|\boldsymbol{x_o})}[F(\boldsymbol{x})] = \alpha_t F(\boldsymbol{x}_0) + (1-\alpha_t)\,\mathbb{E}_{p_{model(\boldsymbol{x})}}[F(\boldsymbol{x})]$$



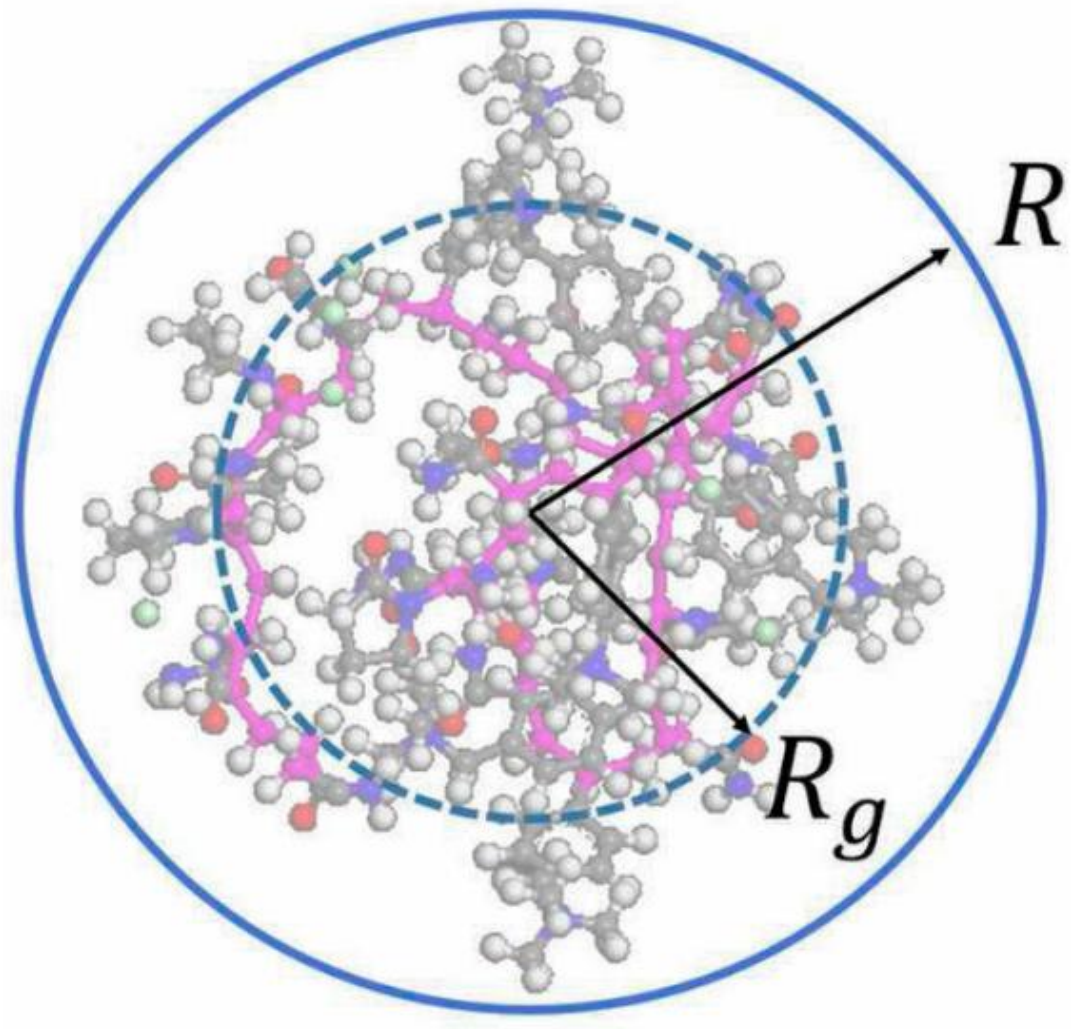
Constraint – Chain Structure



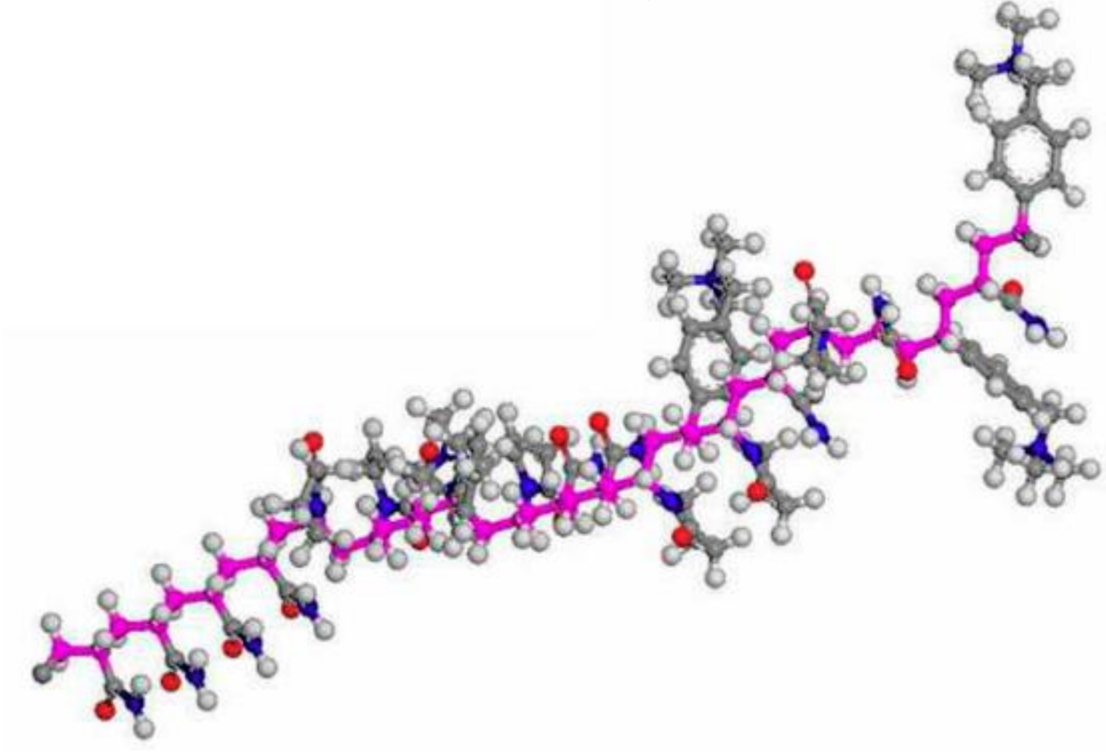
$$\boldsymbol{r_{i,j}} \sim \mathcal{N}(0, \gamma^2 |\mathrm{i} - \mathrm{j}|)$$

$$\mathbb{E}_{p(\boldsymbol{x_t}|\boldsymbol{x_0})}\left[D_{ij}^2(\boldsymbol{x_t})\right] = \alpha_t D_{ij}^2(\boldsymbol{x_0}) + (1 - \alpha_t) 3\gamma^2 |\mathrm{i} - \mathrm{j}|$$

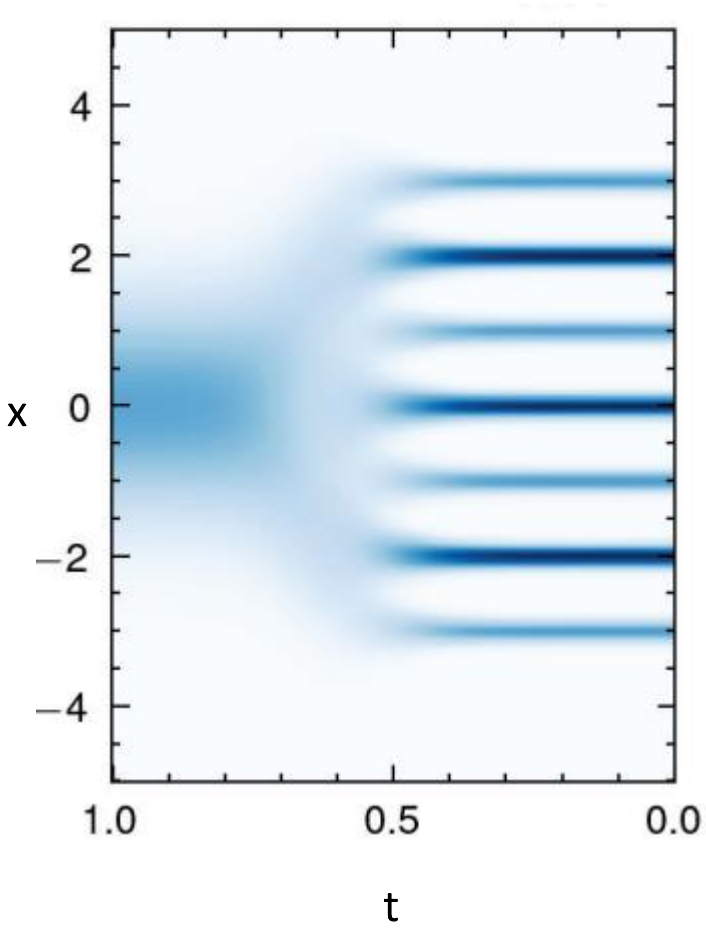
Constraint - Radius of Gyration



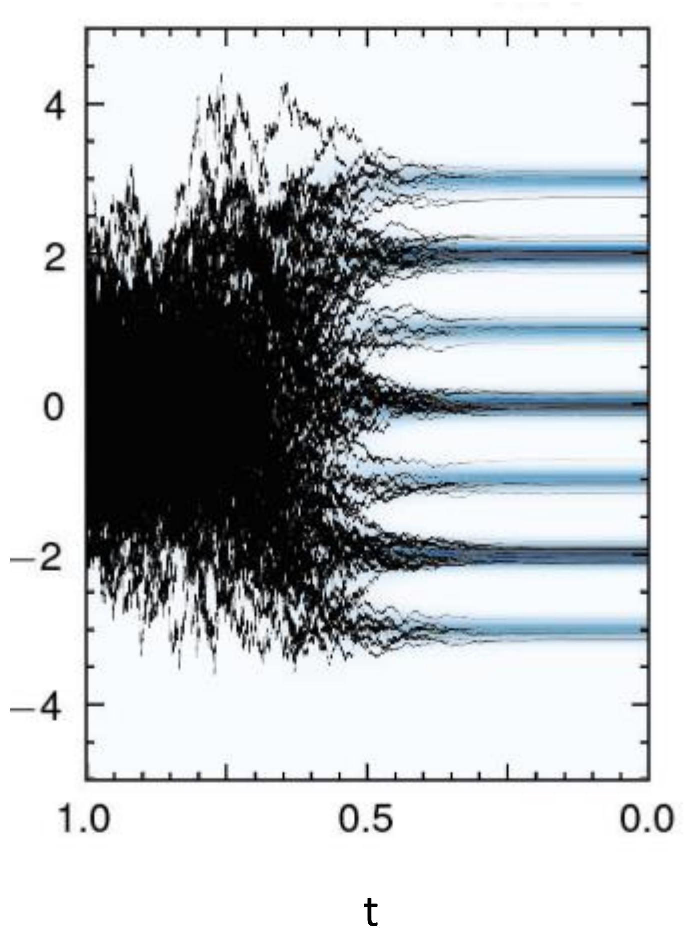
$$R_g^2(x) = \frac{1}{2N^2} \sum_{i,j} D_{i,j}^2(x)$$



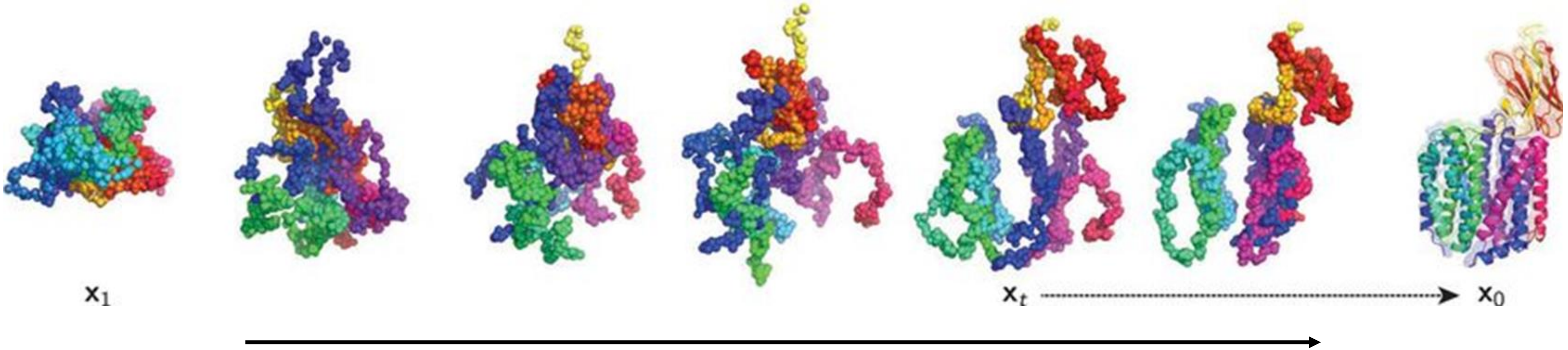
Reverse-time SDE



$$
\begin{gathered}
d\boldsymbol{x} = \\
\left(-\frac{1}{2}\,\boldsymbol{x} - \boldsymbol{R}\boldsymbol{R}^{\boldsymbol{T}}\nabla_x \log \boldsymbol{p}_{\boldsymbol{t}}(\boldsymbol{x})\right)\beta_t dt \\
+\sqrt{\beta_t}\boldsymbol{R}d\widetilde{w}
\end{gathered}
$$

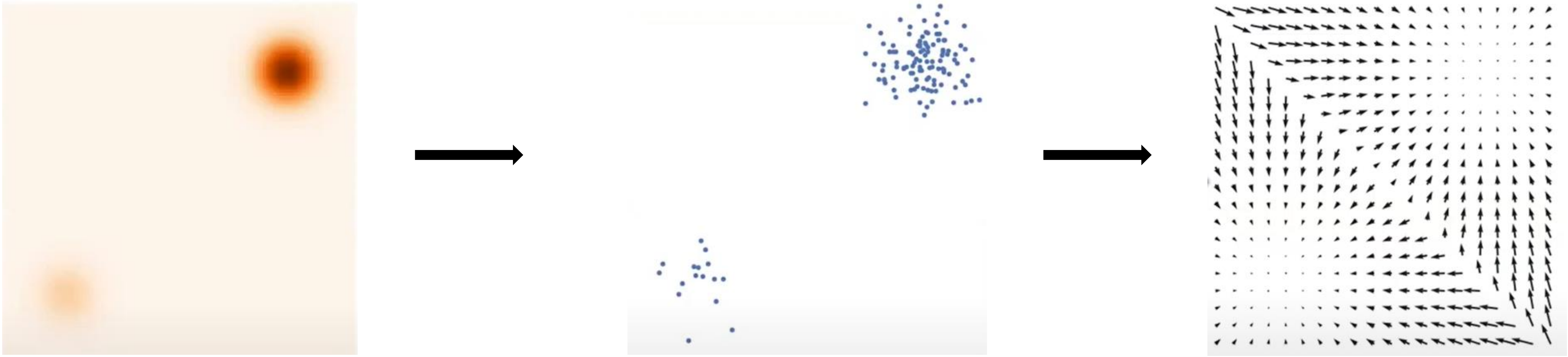


Correlated Reverse-time SDE



$$d\boldsymbol{x} = \left(-\frac{1}{2}\,\boldsymbol{x}\; - \boldsymbol{R}\boldsymbol{R}^{\boldsymbol{T}}\nabla_x \log \boldsymbol{p_t}(\boldsymbol{x})\right)\beta_t dt + \sqrt{\beta_t}\boldsymbol{R}d\widetilde{w}$$

Score Estimation



$p_{data}(x)$

Samples

$\nabla_x \log p_{data}(x)$

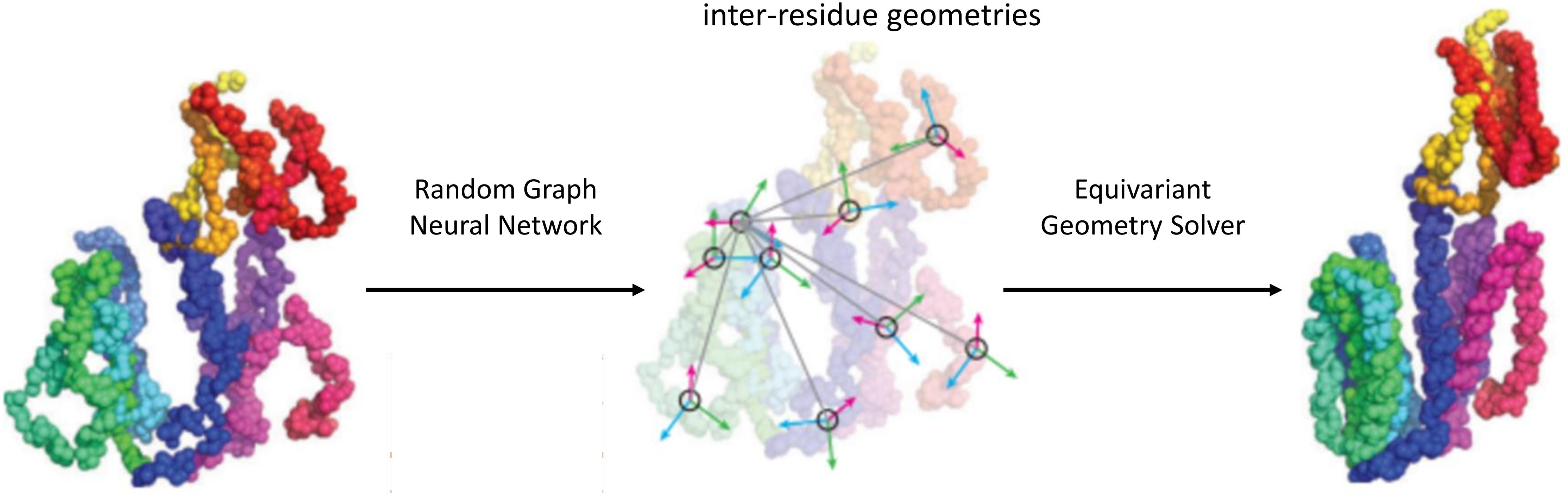
$\sim \nabla_x \log p(x)$

Optimized denoiser

$$\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}) = \left((1-\alpha_t)\boldsymbol{R}\boldsymbol{R}^T\right)^{-1} \left(\sqrt{\alpha_t}\,\hat{\boldsymbol{x}}_\theta(\boldsymbol{x},t) - \boldsymbol{x}\right)$$

$$\mathcal{L}^{\text{reg}}_{\mathbf{x}}(\mathbf{x};\boldsymbol{\theta}) = -\mathbb{E}_{\mathbf{x}_t\sim p(\mathbf{x}_t|\mathbf{x}),t\sim\text{Unif}(0,1)}\left[\frac{\alpha_t\beta_t}{2(1-\alpha_t)^2}\left\|\left(\mathbf{R}^{-1}+\omega\mathbf{I}\right)\left(\hat{\mathbf{x}}_\theta(\mathbf{x}_t,t)-\mathbf{x}\right)\right\|_2^2\right]$$

Optimized denoiser



Noisy structure
$\boldsymbol{x}_t$

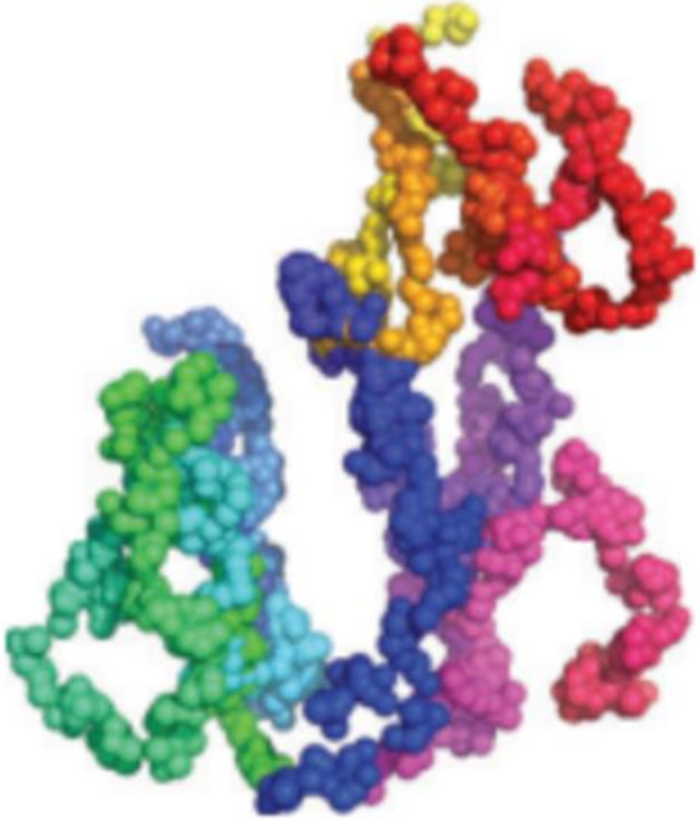
Random Graph
Neural Network

Confidence-weighted predicted
inter-residue geometries

Equivariant
Geometry Solver

Predicted denoised structure
$\hat{\boldsymbol{x}}_\theta(\boldsymbol{x}, t)$

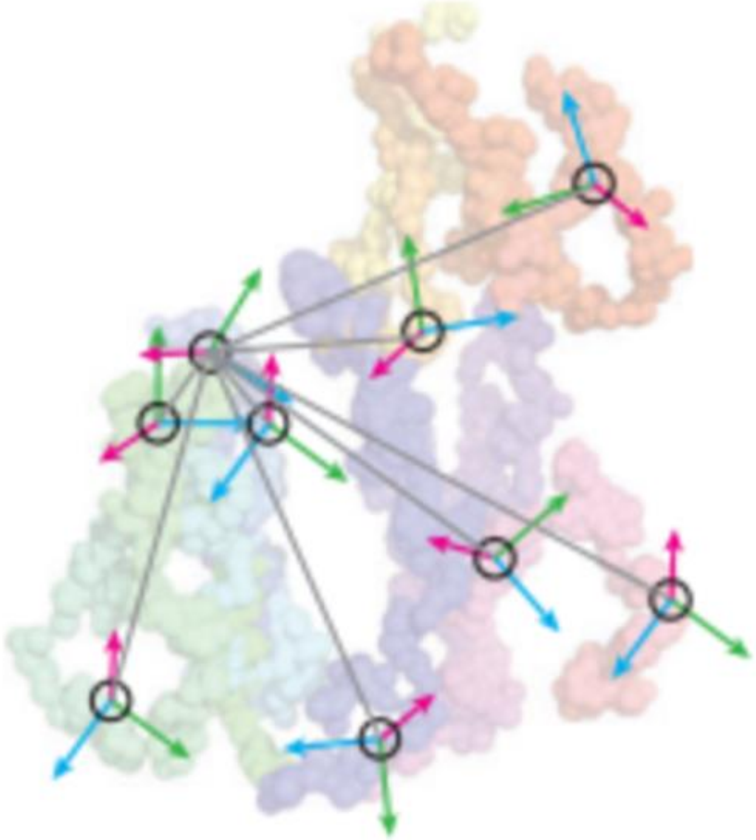
Reduced computational complexity



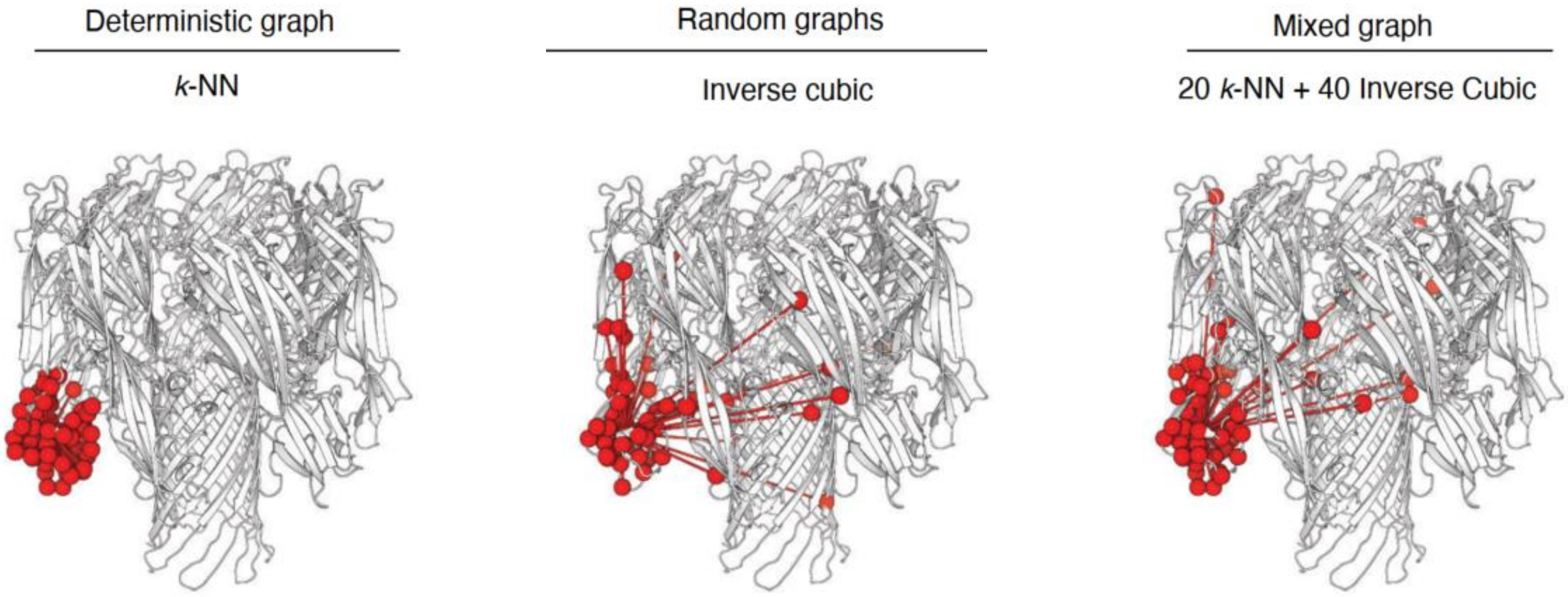
Random Graph Neural Network

→

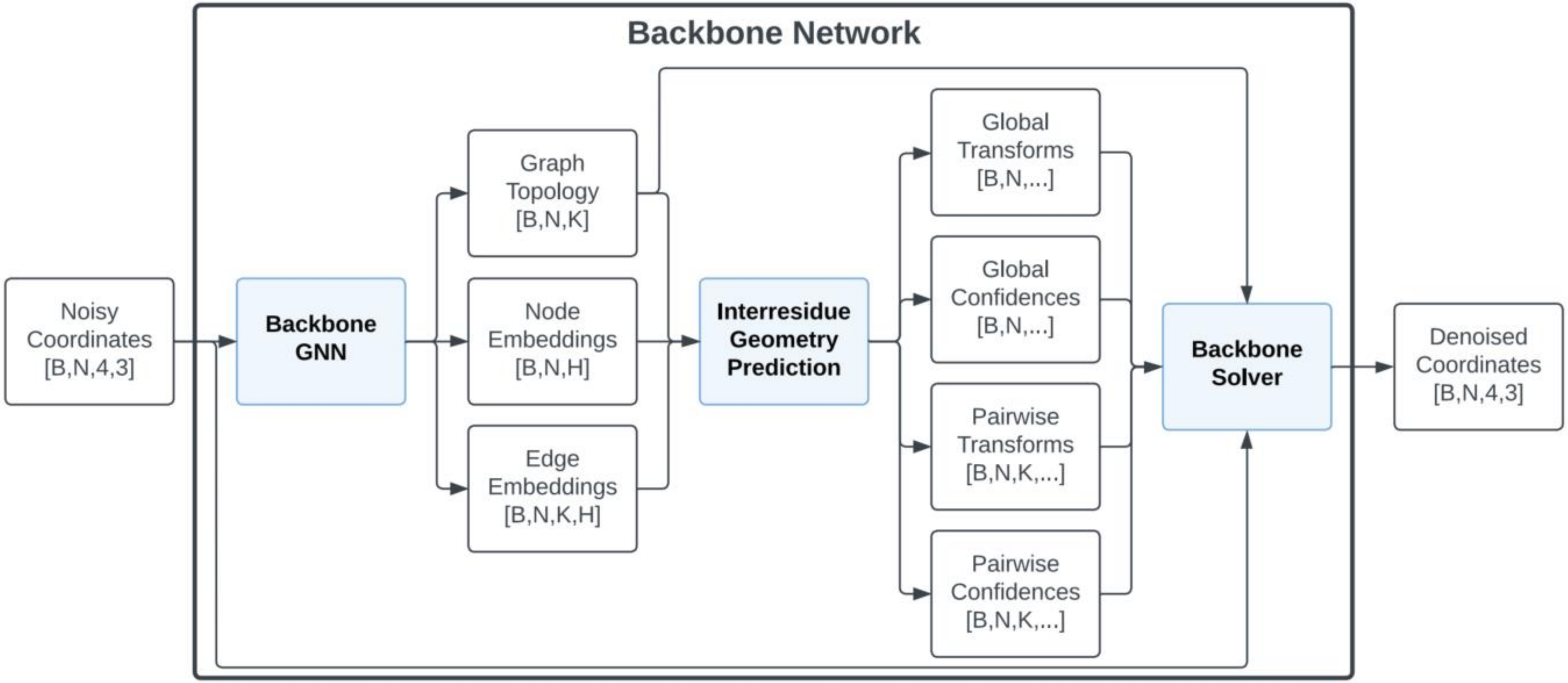
$\boldsymbol{\mathcal{O}}(NlogN)$
or
$\boldsymbol{\mathcal{O}}(N)$
edges



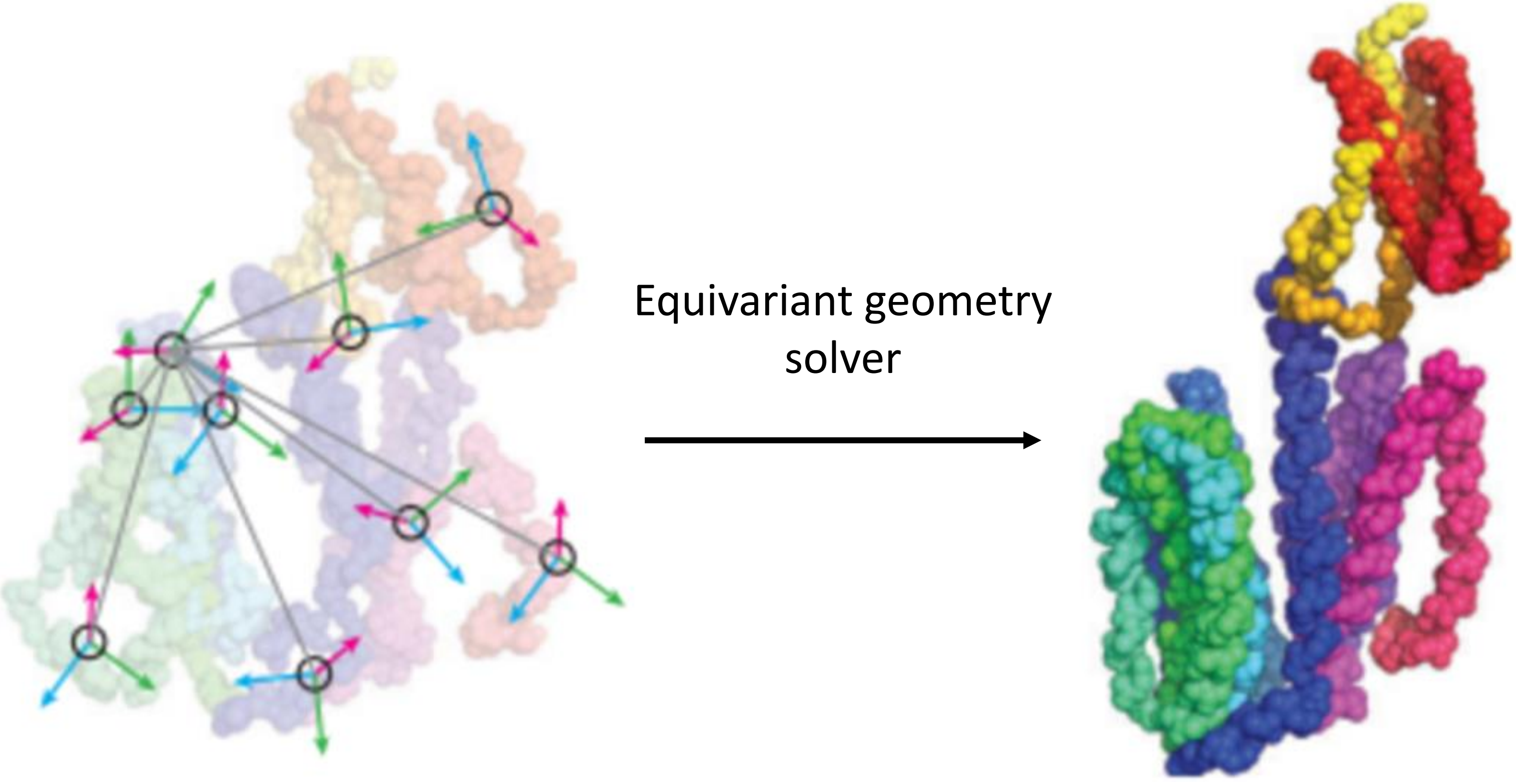
Sub- $\boldsymbol{\mathcal{O}}(N^2)$ scaling - Random edge sampling



Backbone graph neural network

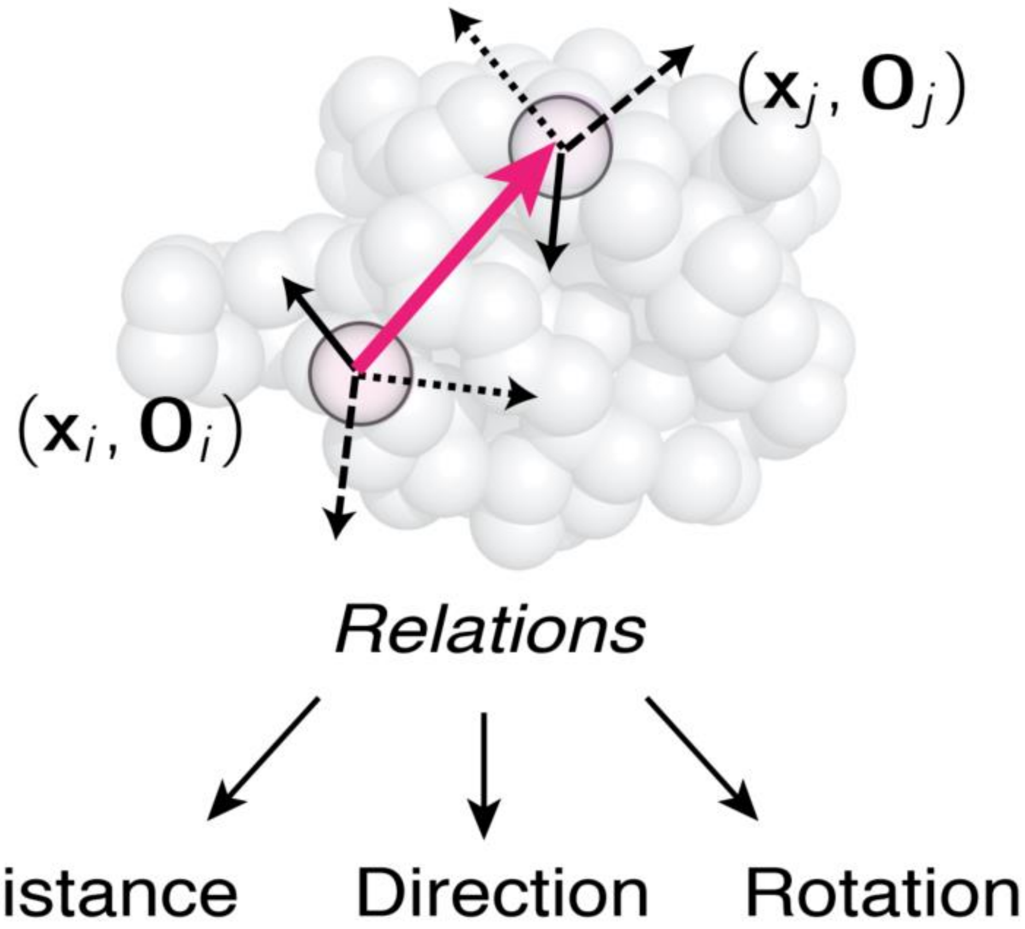

Backbone Network
Noisy Coordinates [B,N,4,3]
Backbone GNN
Graph Topology [B,N,K]
Node Embeddings [B,N,H]
Edge Embeddings [B,N,K,H]
Interresidue Geometry Prediction
Global Transforms [B,N,...]
Global Confidences [B,N,...]
Pairwise Transforms [B,N,K,...]
Pairwise Confidences [B,N,K,...]
Backbone Solver
Denoised Coordinates [B,N,4,3]

Equivariant geometry solver

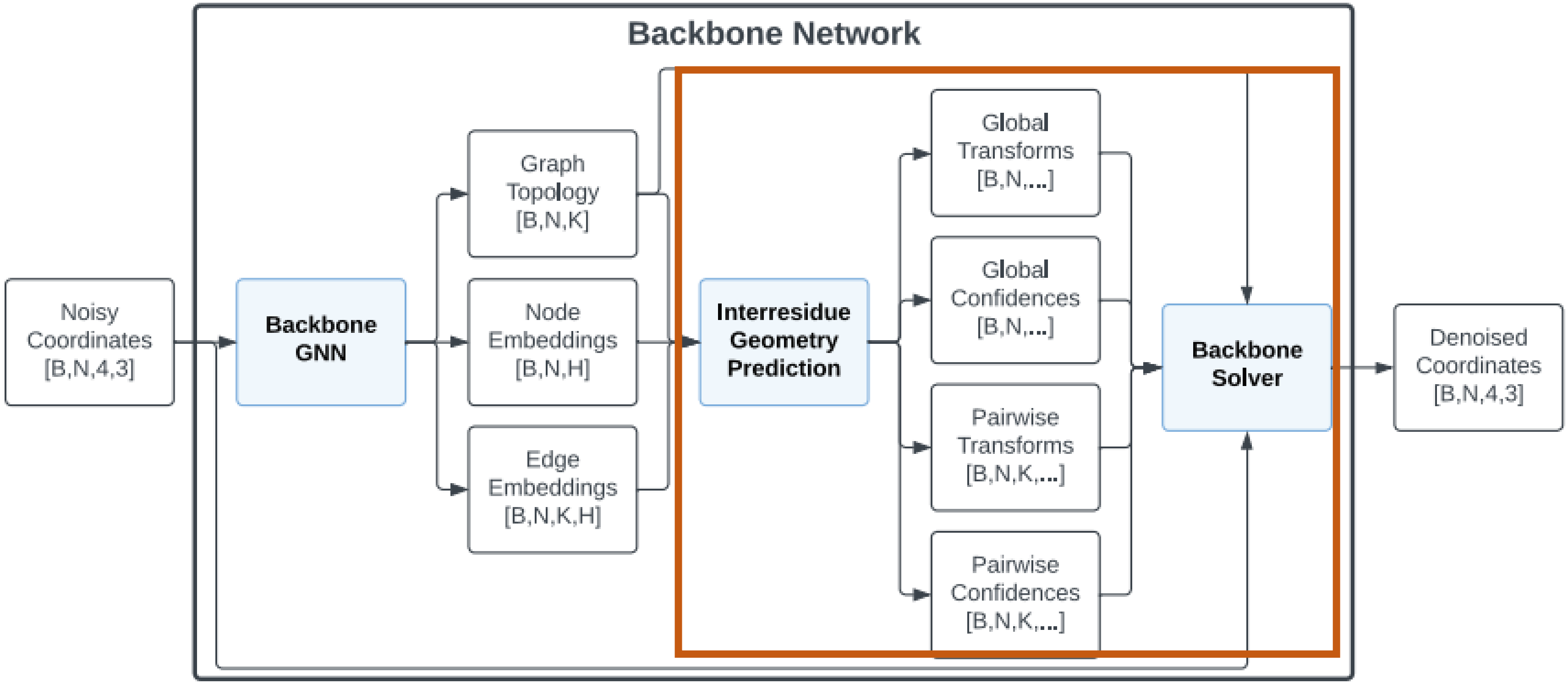


Equivariant geometry solver

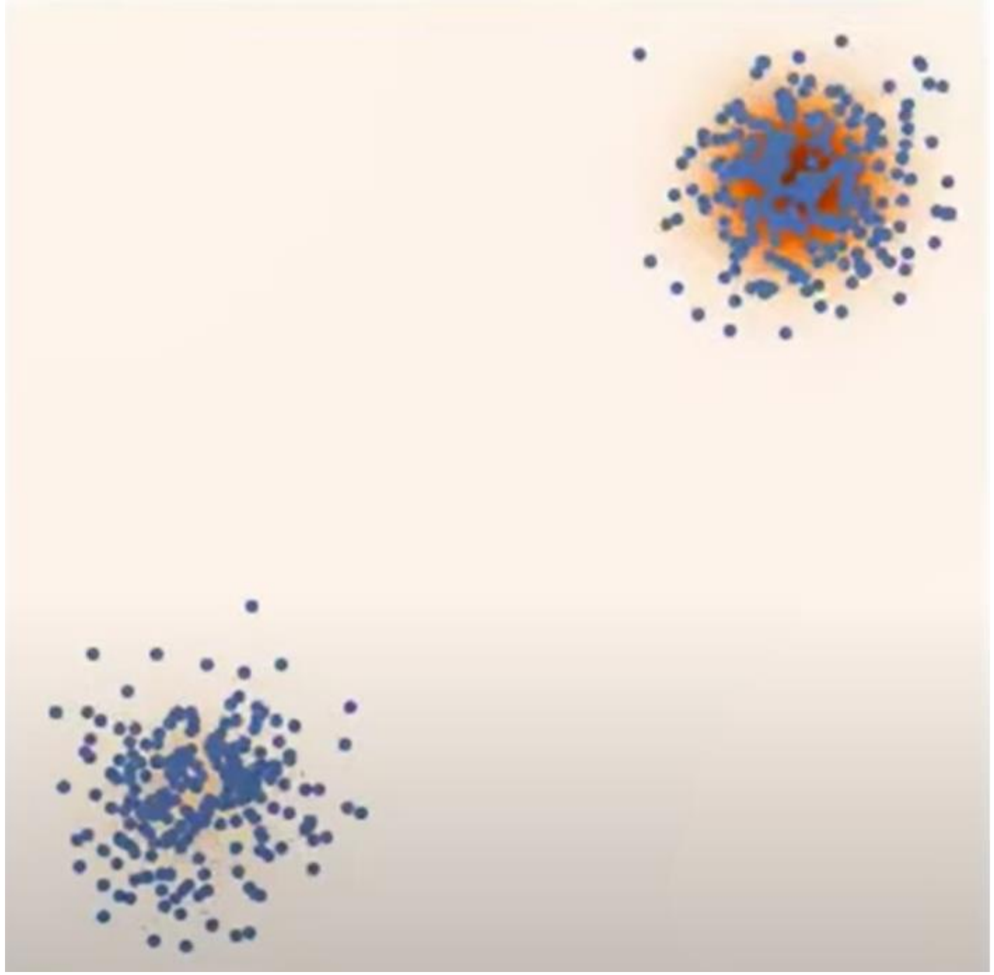
Invariant local frame relations

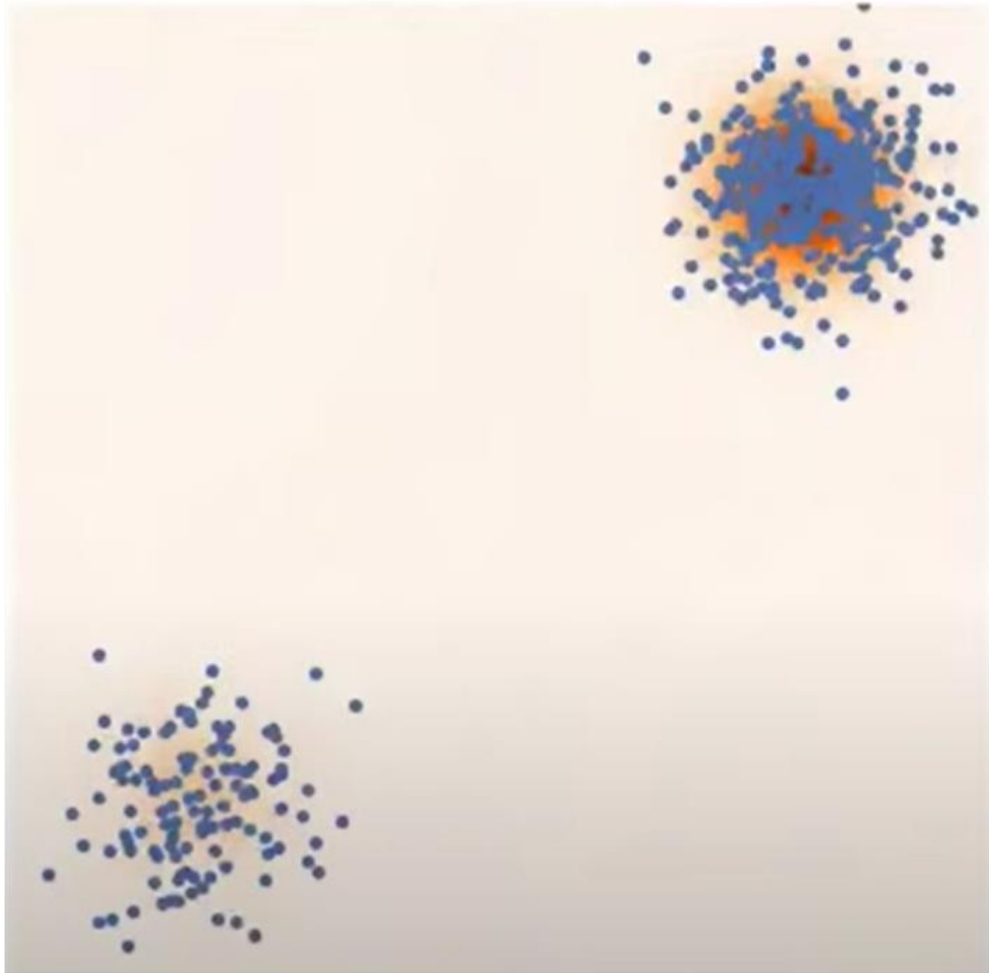

$(\mathbf{x}_j, \mathbf{O}_j)$
$(\mathbf{x}_i, \mathbf{O}_i)$
Relations
Distance
Direction
Rotation

Equivariant geometry solver

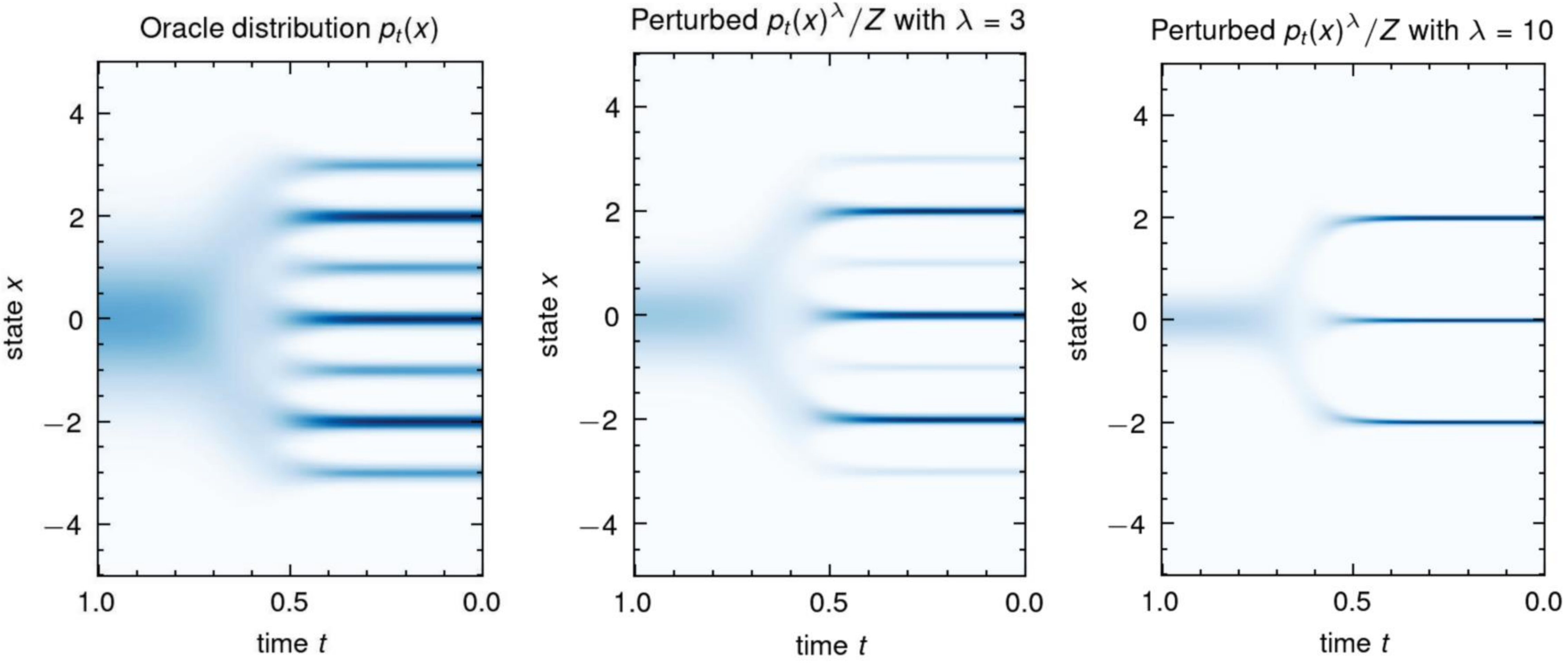

Backbone Network
Noisy Coordinates [B,N,4,3]
Backbone GNN
Graph Topology [B,N,K]
Node Embeddings [B,N,H]
Edge Embeddings [B,N,K,H]
Interresidue Geometry Prediction
Global Transforms [B,N,...]
Global Confidences [B,N,...]
Pairwise Transforms [B,N,K,...]
Pairwise Confidences [B,N,K,...]
Backbone Solver
Denoised Coordinates [B,N,4,3]

Sampling of the backbone - overdispersion



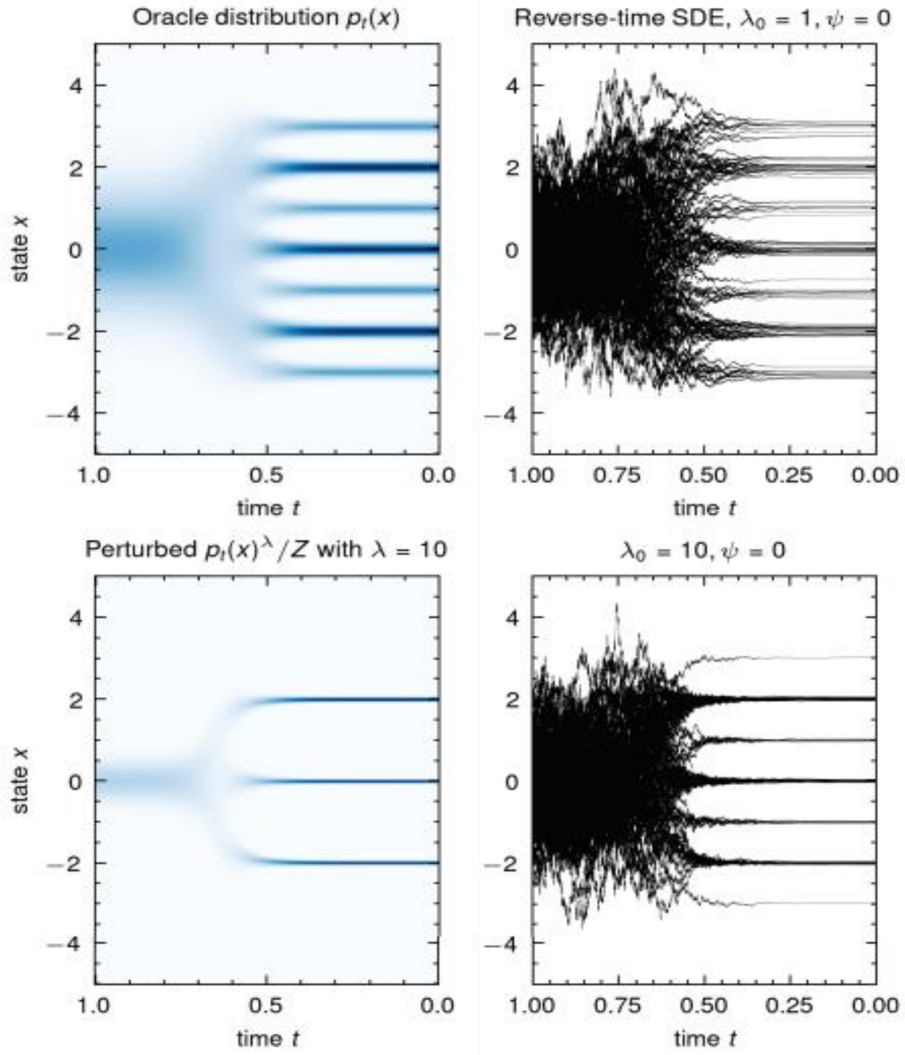


Low temperature sampling – reweight and concentrate

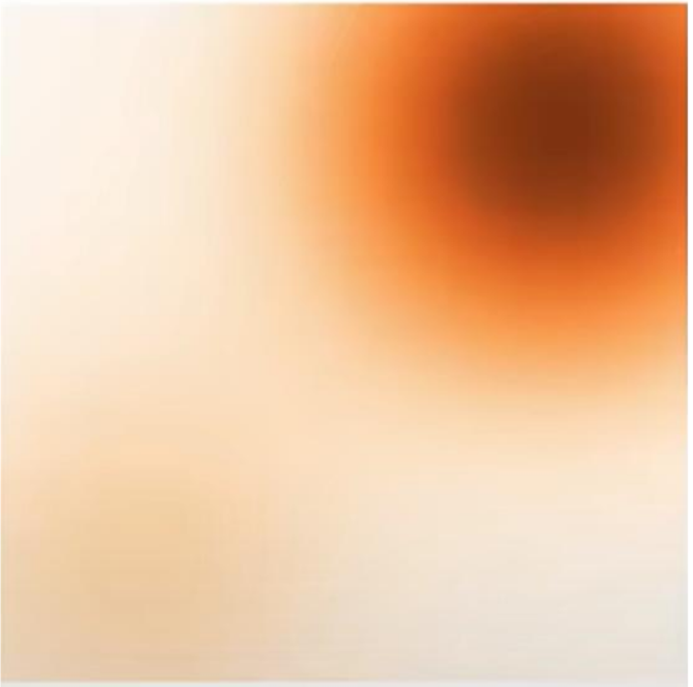


Annealed reverse-time SDE

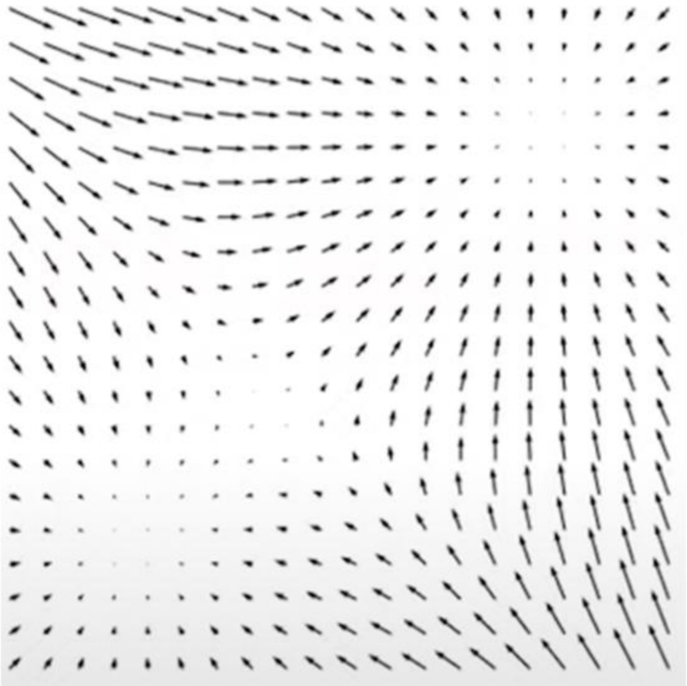
$$\lambda_t \approx \frac{\lambda_0}{\alpha_t + (1 - \alpha_t)\lambda_0}$$



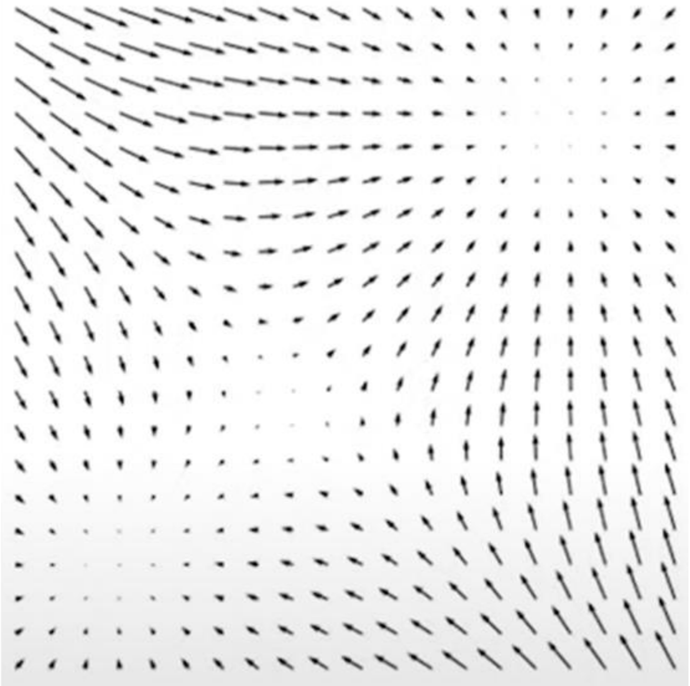
Score evolution



Data Density



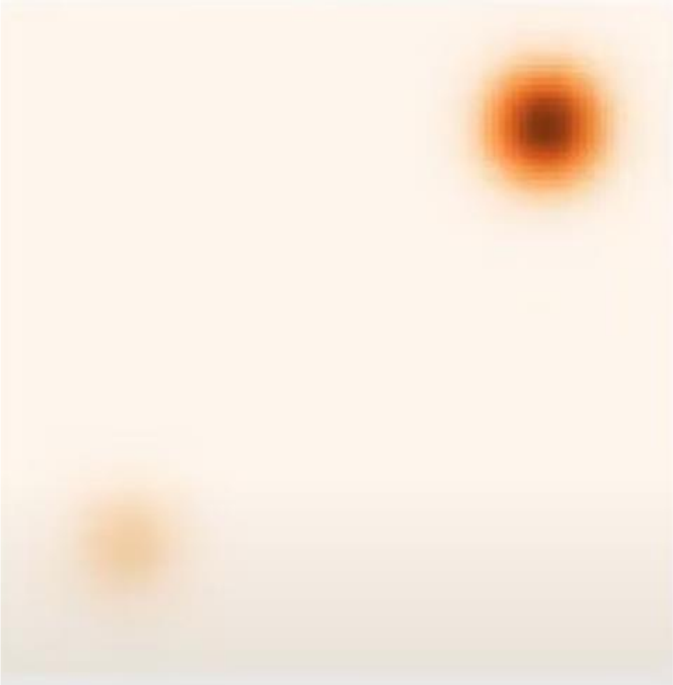
Data Score



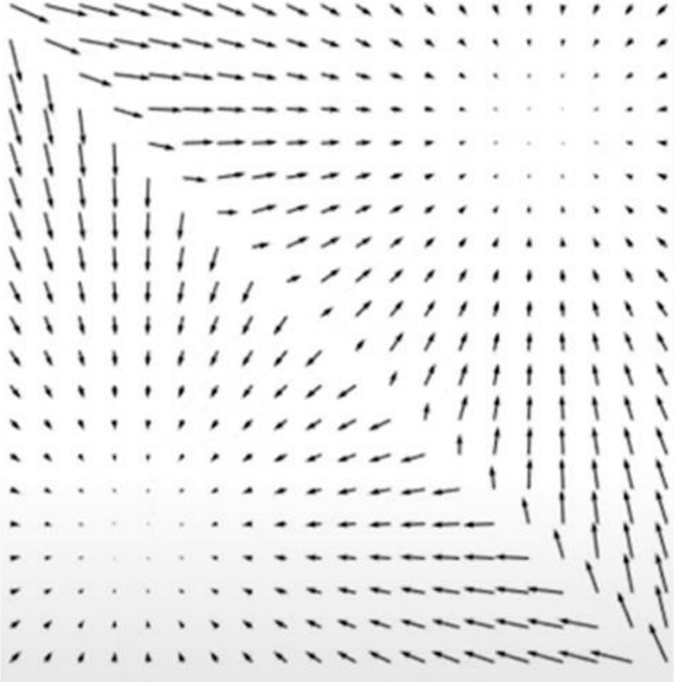
Estimated Score

Score evolution

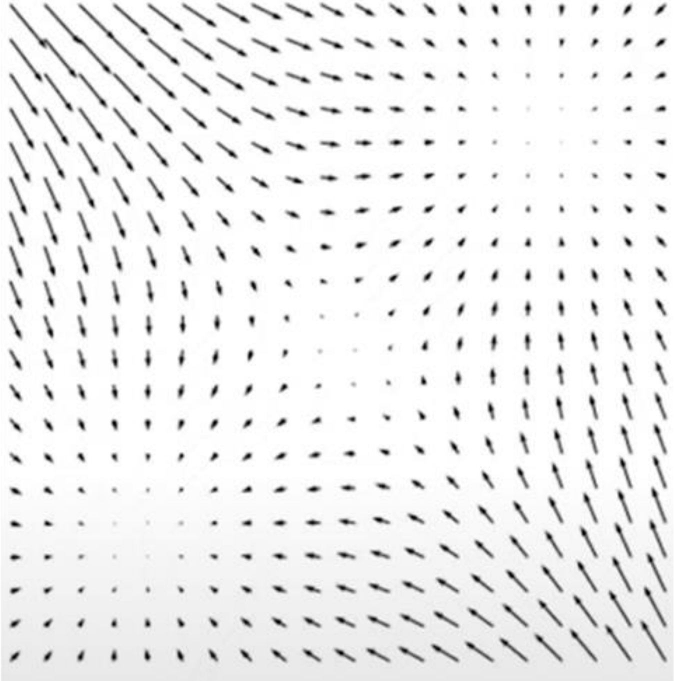
Data Density



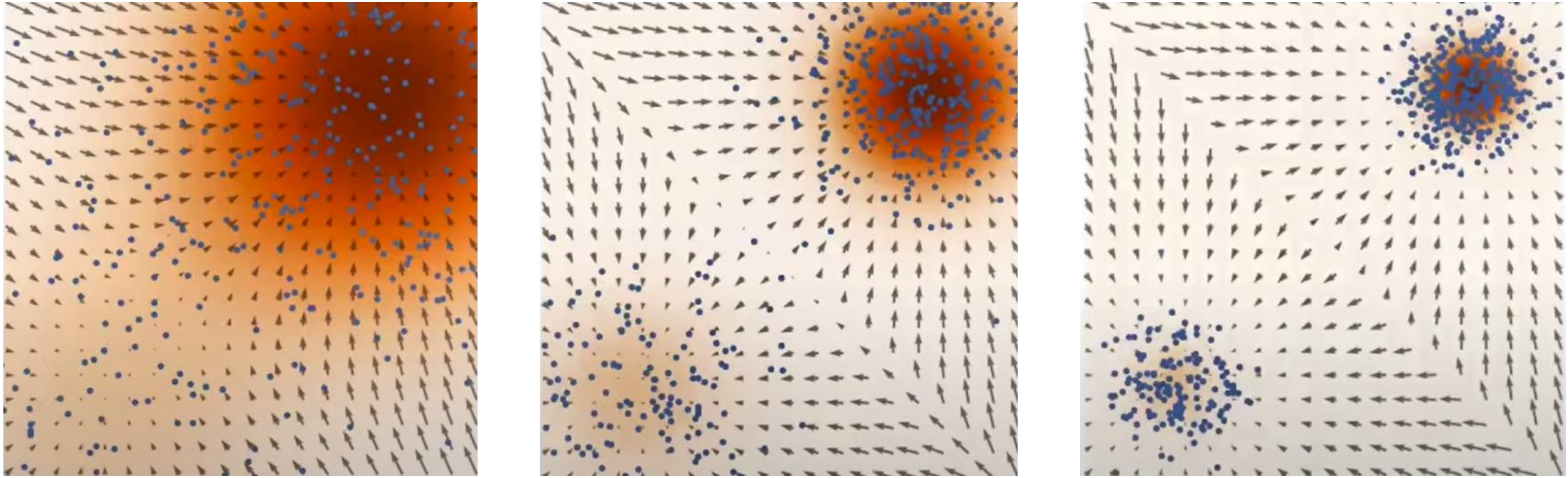
Data Score



Estimated Score

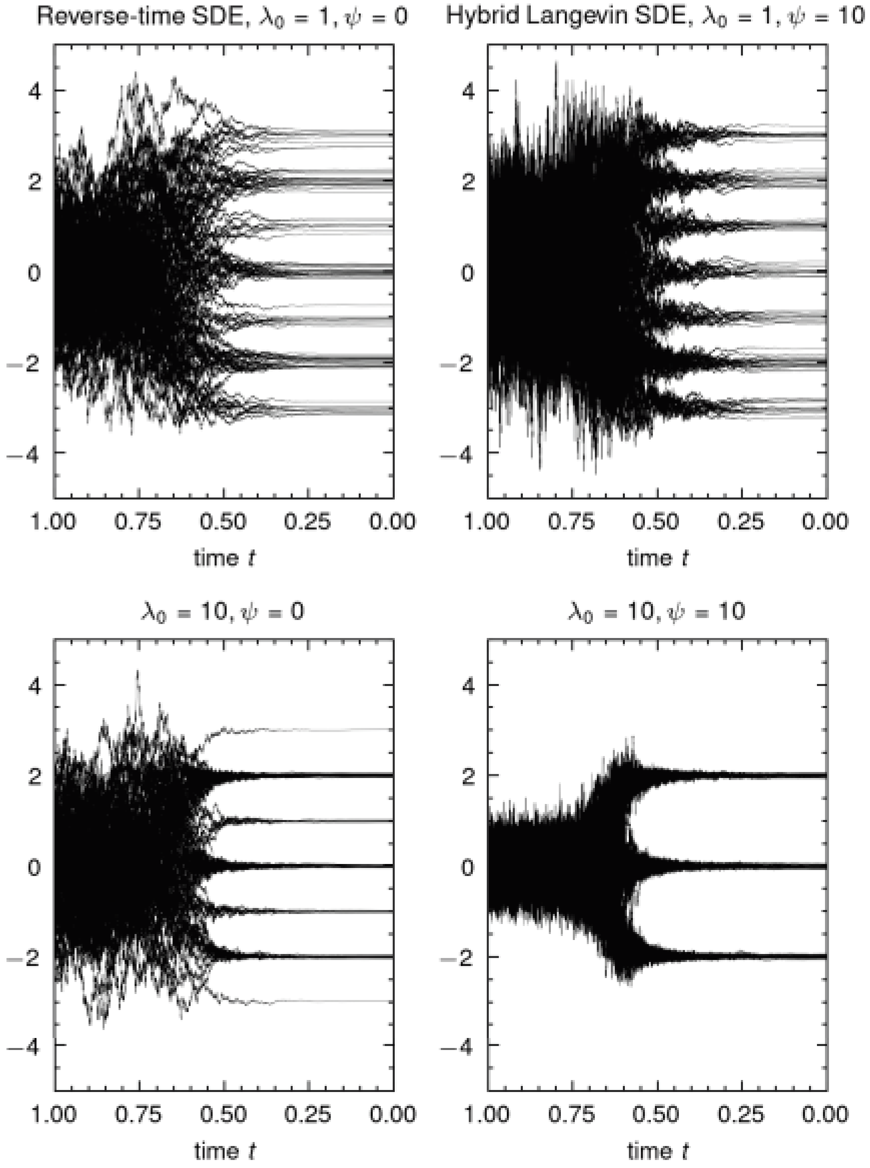


Annealed Langevin dynamics

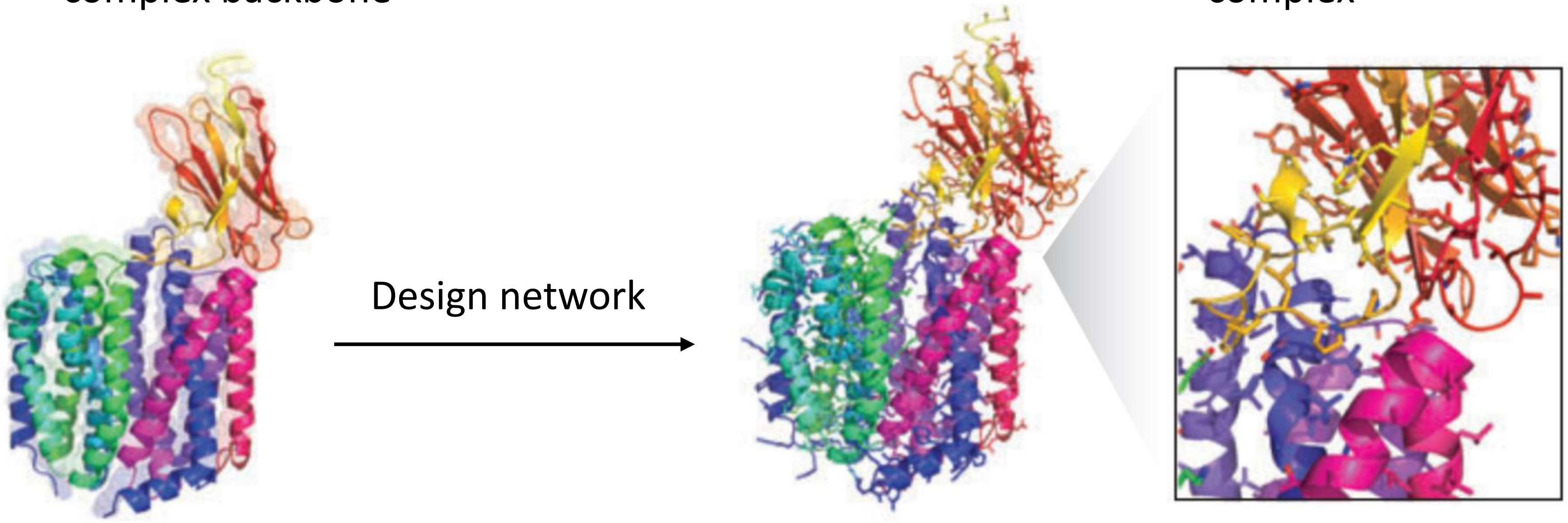


Hybrid Langevin SDE

$$dx = \left(-\frac{1}{2}\boldsymbol{x} - \left(\lambda_t + \frac{\lambda_0 \psi}{2}\right) \frac{\sqrt{\alpha_t}\widehat{\boldsymbol{x}_{\boldsymbol{\theta}}}(\boldsymbol{x}, t) - \boldsymbol{x}}{1 - \alpha_t}\right)\beta_t dt + \sqrt{\beta_t(1 + \psi)}\boldsymbol{R} d\bar{w}$$



From backbone to sequence and heavy atom position

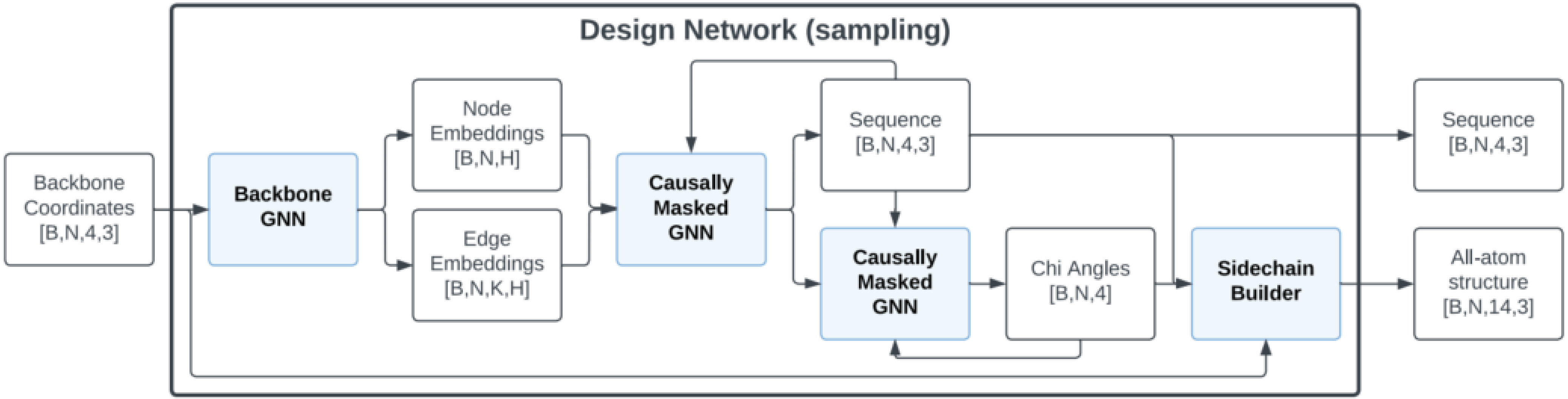


Sampled protein complex backbone

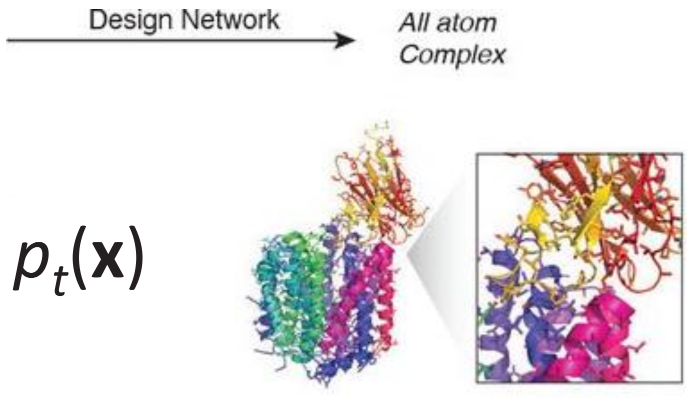
Design network

All atom protein complex

Design Network

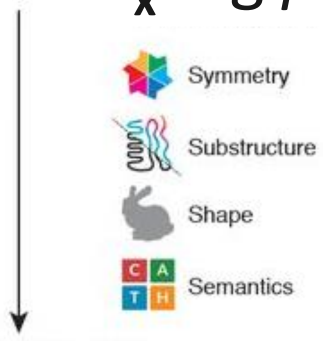


Conditional modeling



$$\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$$

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



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

Conditional modeling

Bayes' rule

$$p(x|y) = \frac{p(x)p(y|x)}{p(y)}$$

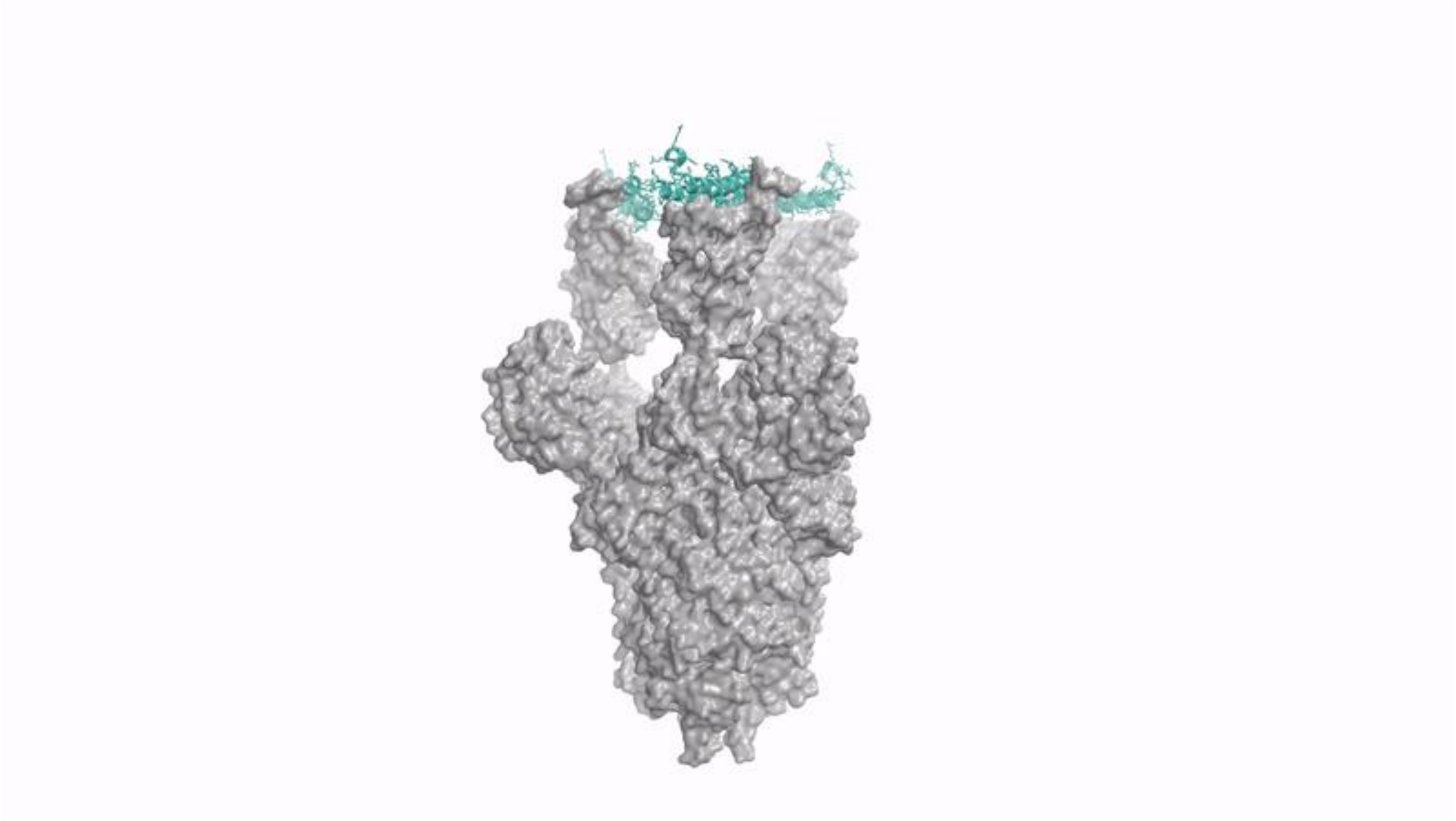
Bayes' rule for score functions

$$\boxed{\nabla x \log p(x)} + \boxed{\nabla x \log p(y|x)} \cancel{- \nabla x \log p(y)}$$

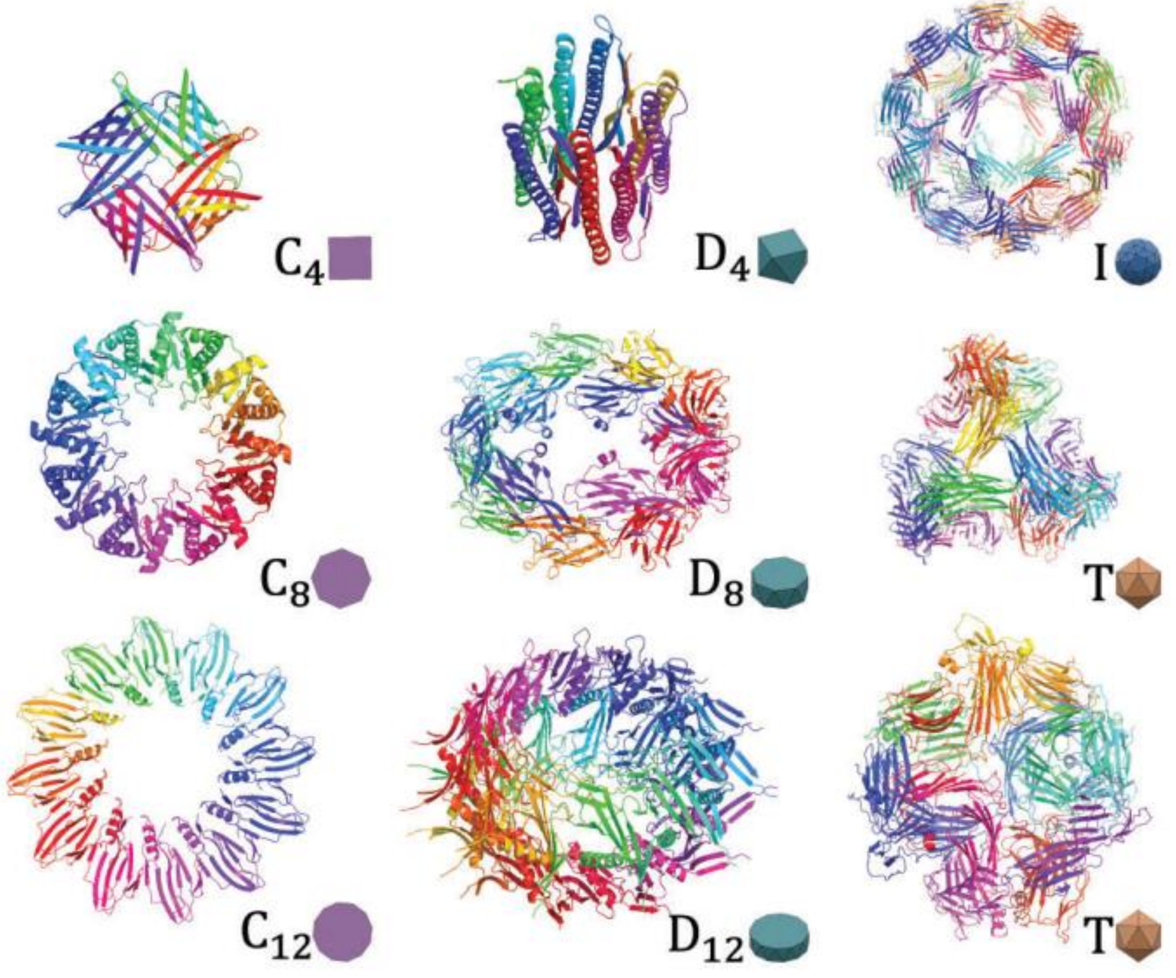
Unconditional score

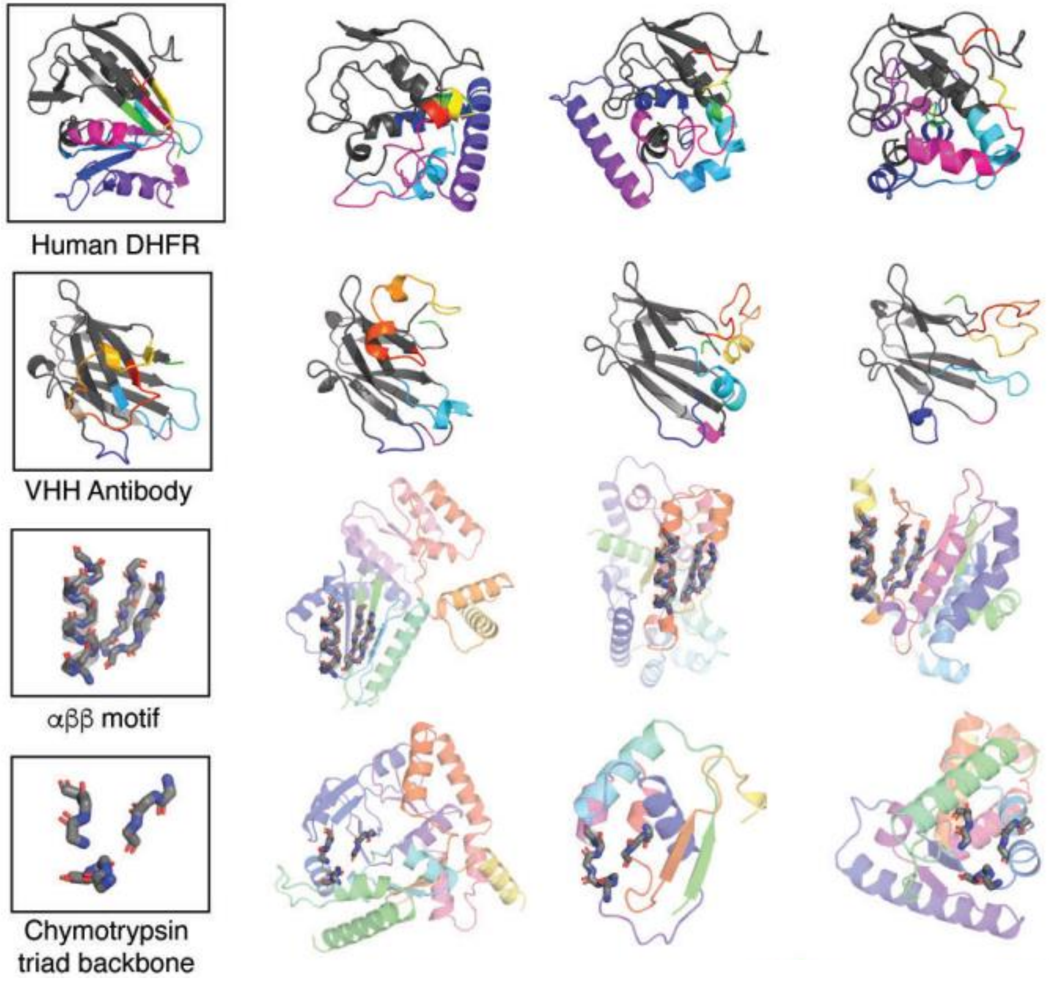
e.g. classifier

Conditional modeling



Symmetry and substructure guided diffusion

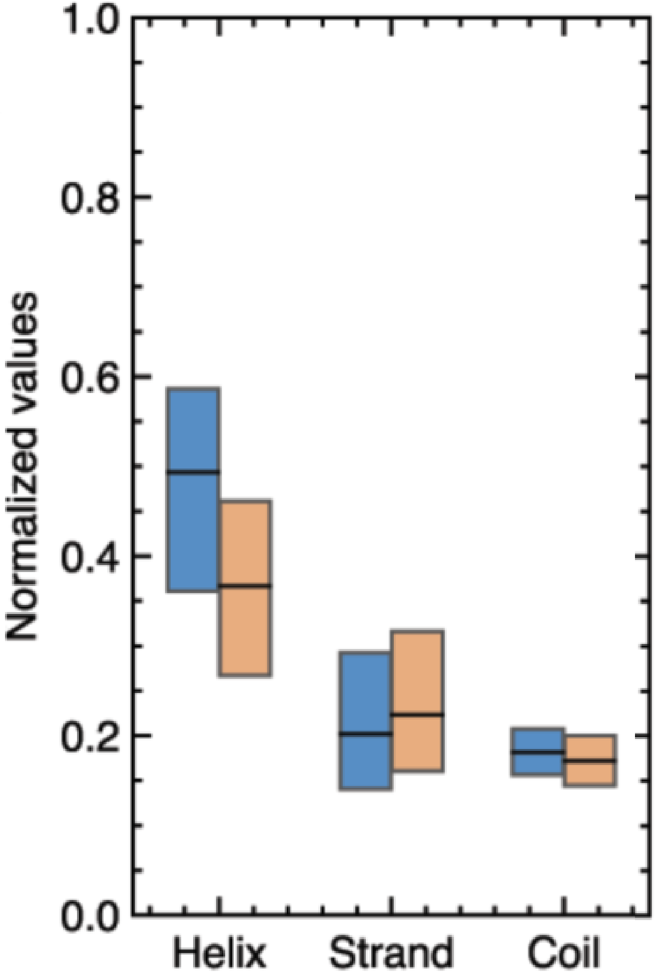


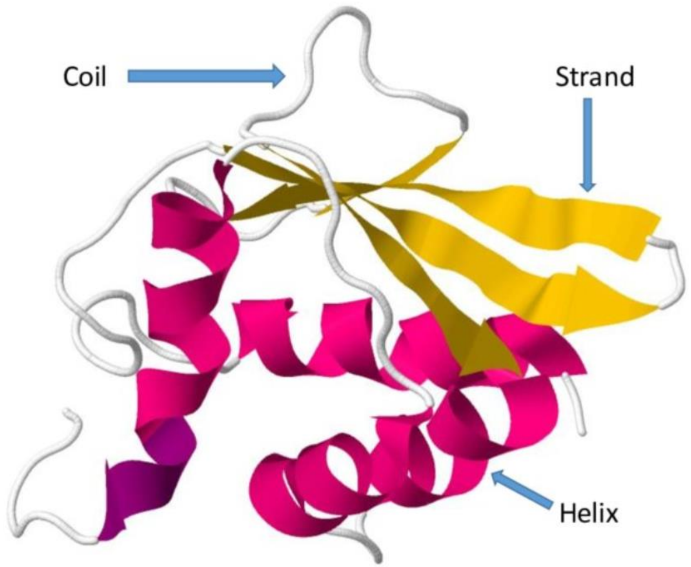


Evaluation

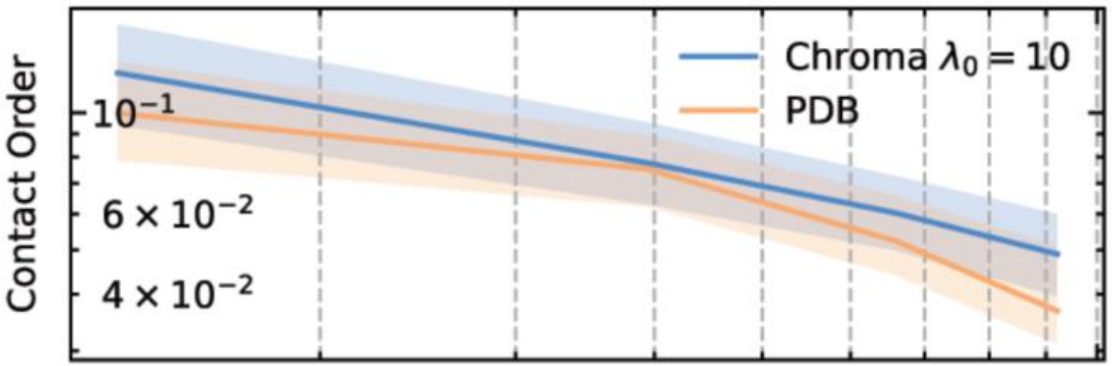
- 50,000 single chains, 10,000 complexes - qualitative
- 10,000 single chain proteins - quantitative
- $\lambda_0 = 10$
- $\psi = 2$
- 200 steps
- Single chain lengths N: $p(N) = 1/N$
- Complex # chain and N = # chain and N of random complex from PDB

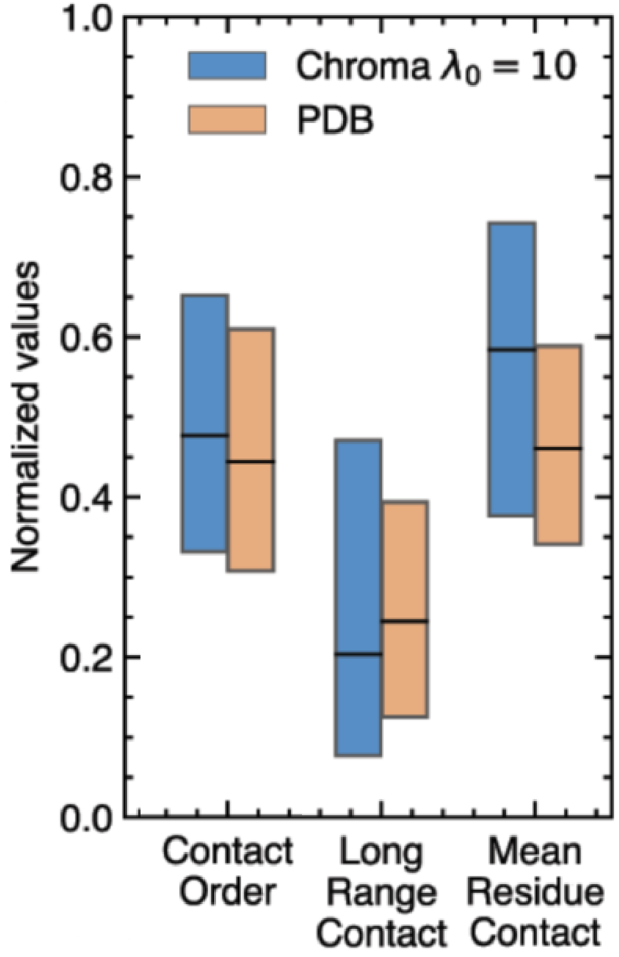
Evaluation – Secondary structurs

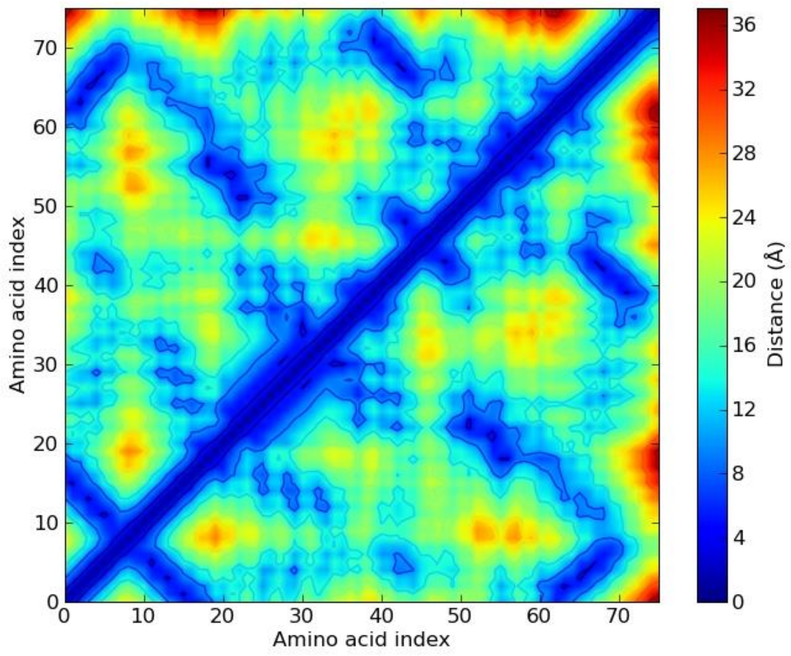




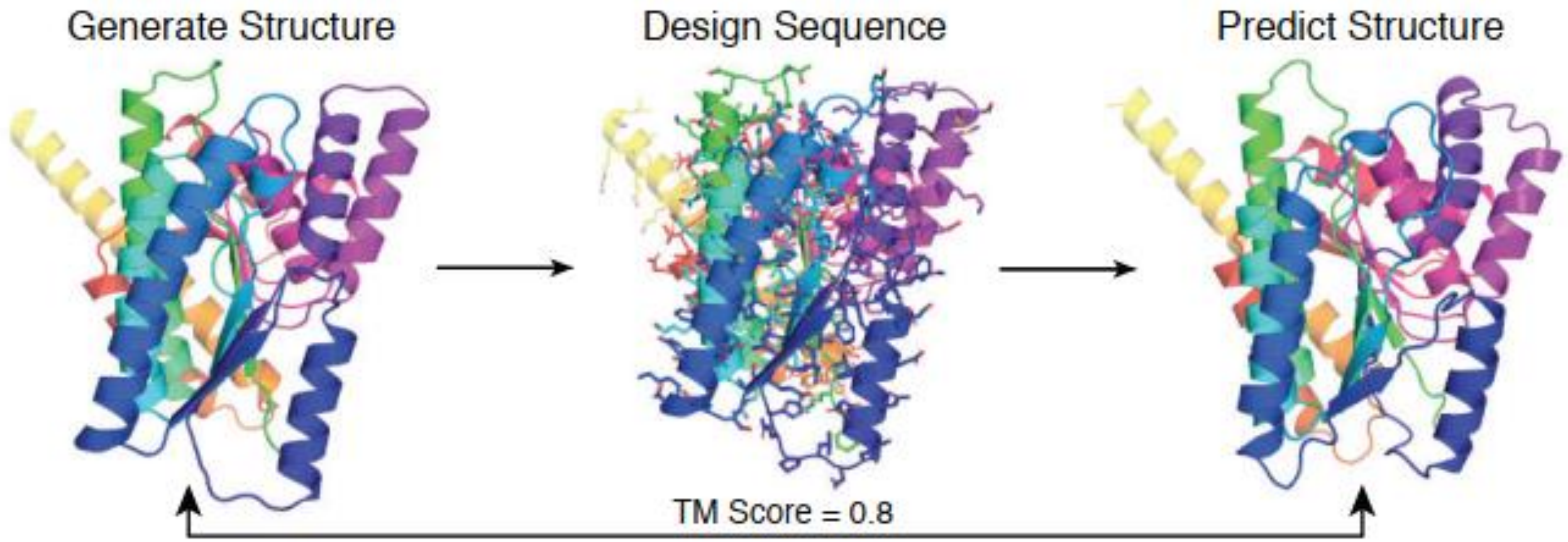
Evaluation – Residue interactions



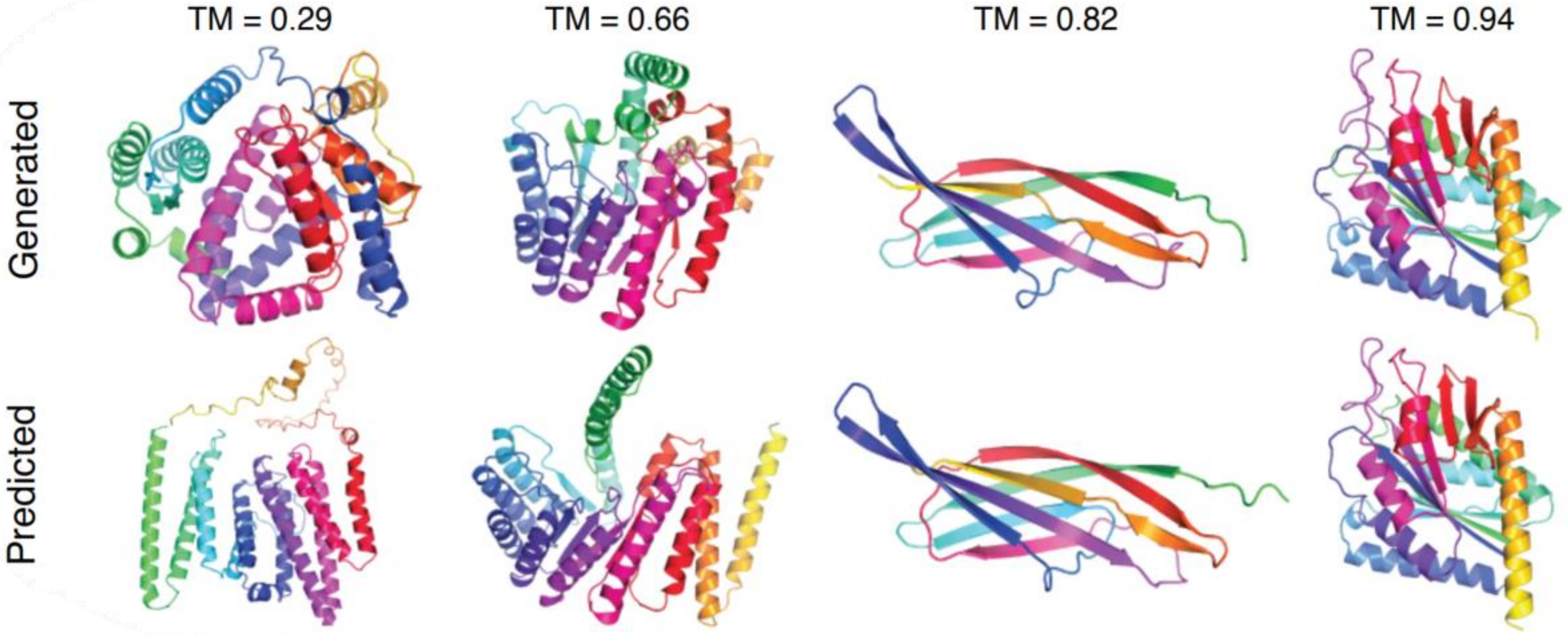




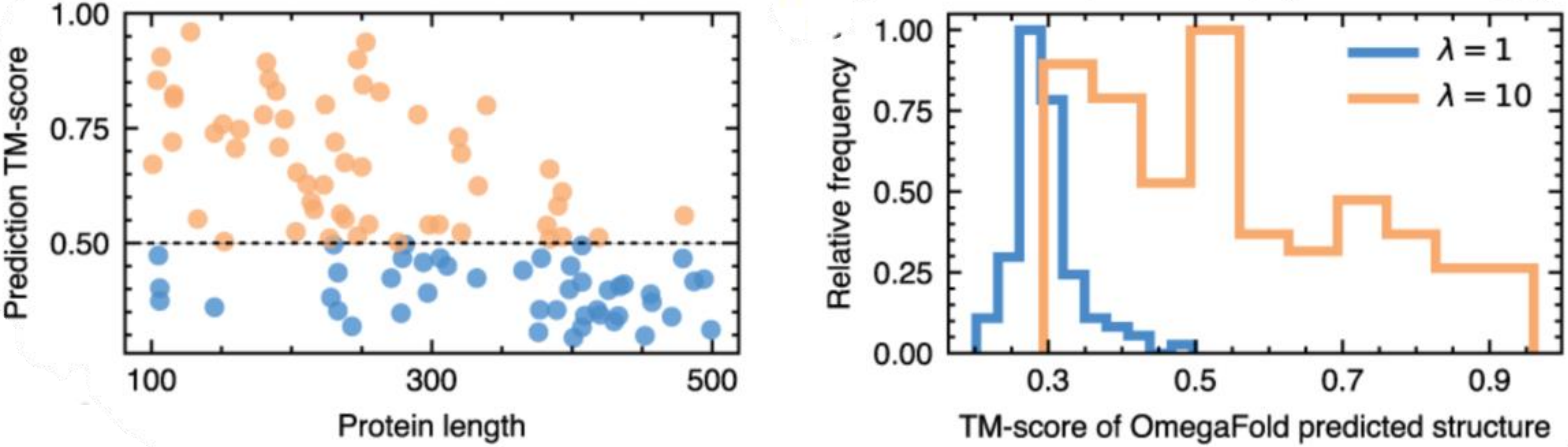
Evaluating Chroma by structure prediction with OmegaFold



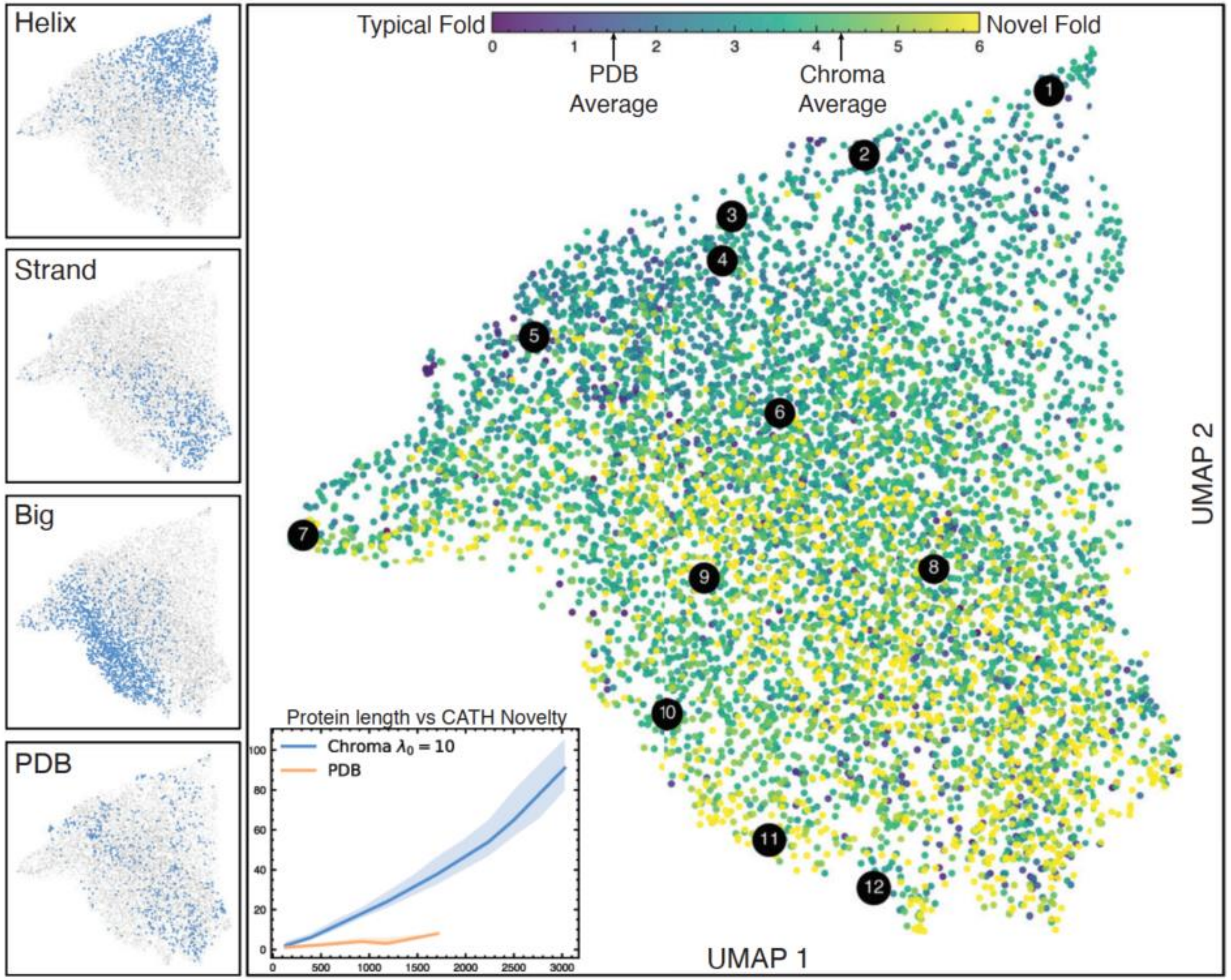
Evaluation - TM-scores



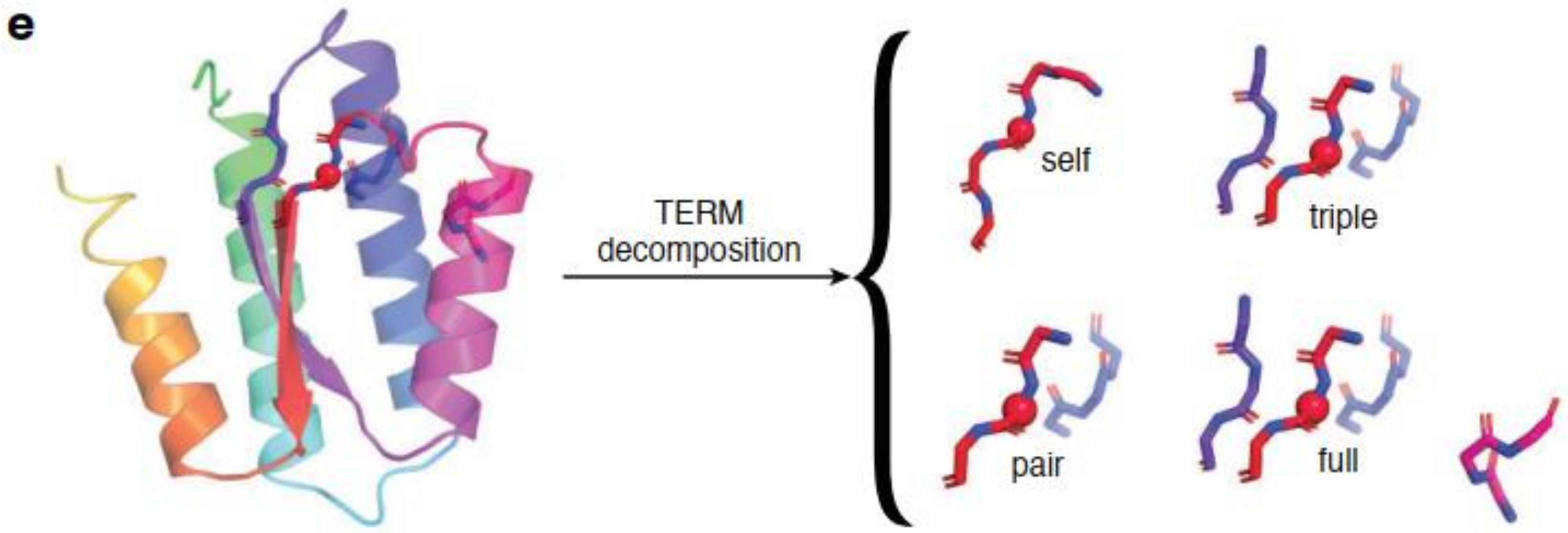
Evaluation - TM-scores



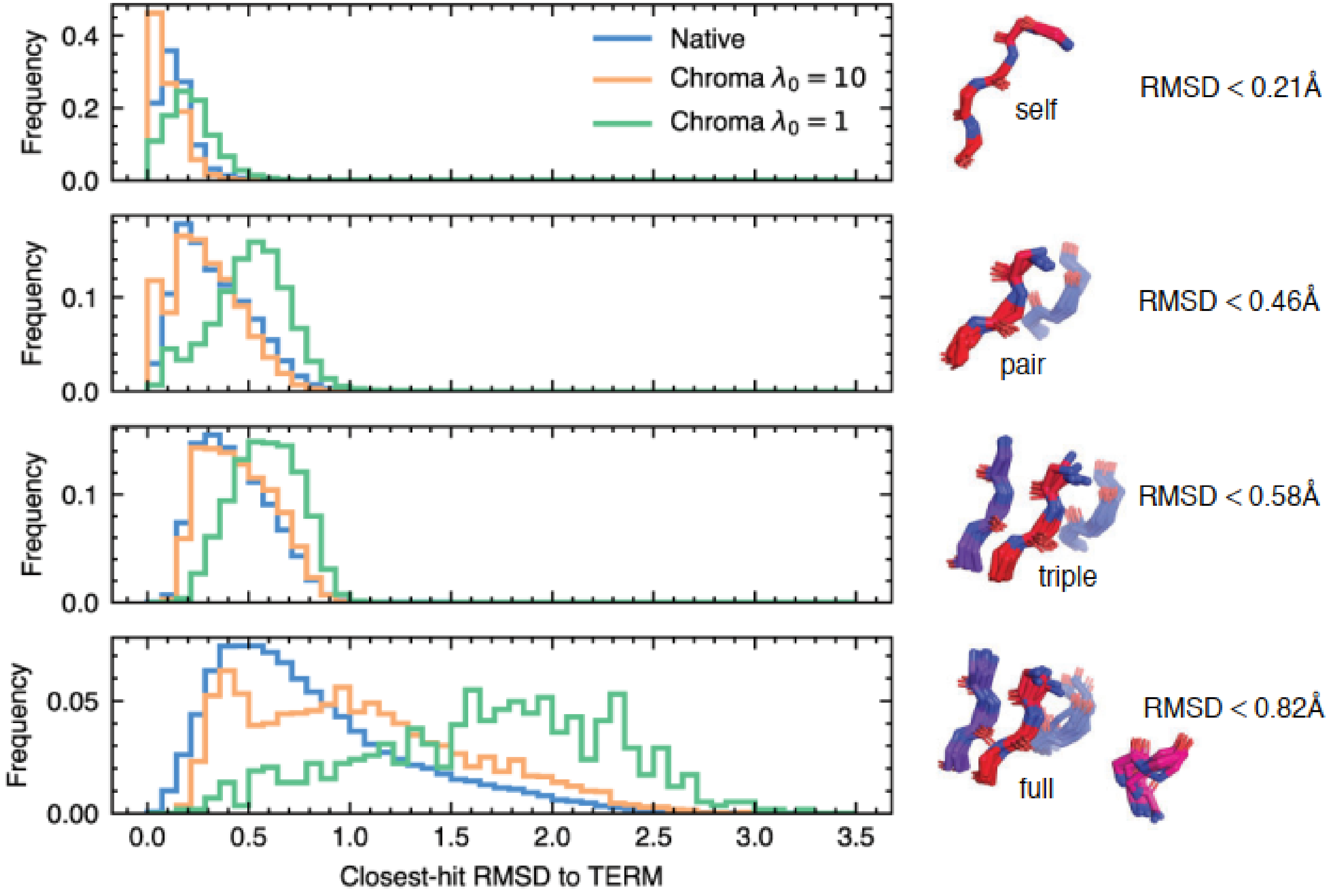
Evaluation - Novelty and structural homology



Evaluation - TERMs



Evaluation - Closest-match RMSD for TERMs



Limitations

+ combination of promising maturities in GDMs
+ elegant way they implement empirical knowledge

- missing experimental characterization
- no quantitative evaluations for many designs and design choices
- no benchmarking
- sequential generation of backbone, sequence and rotamers
- choice of model to evaluate folding

Illuminating chroma?!