Generative Models for Graph-Based Protein Design

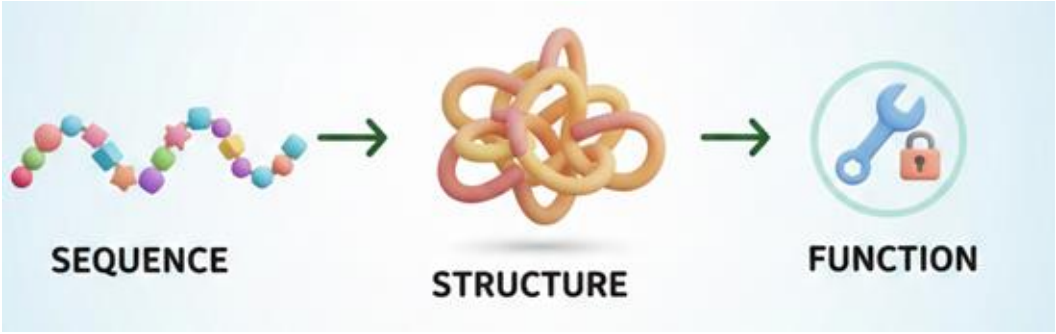
John Ingraham, Vikas K. Garg, Regina Barzilay, Tommi Jaakkola

Computer Science and Artificial Intelligence Lab, MIT

NeurIPS 2019

Presented By
Md Toki Tahmid
Department of Computer Science
Princeton University

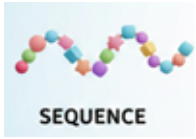
Why Protein Design Matters



A protein is:

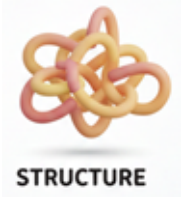
- A **sequence** of amino acids
- That **folds into a 3D structure**
- That structure determines its **function**

Why Protein Design Matters



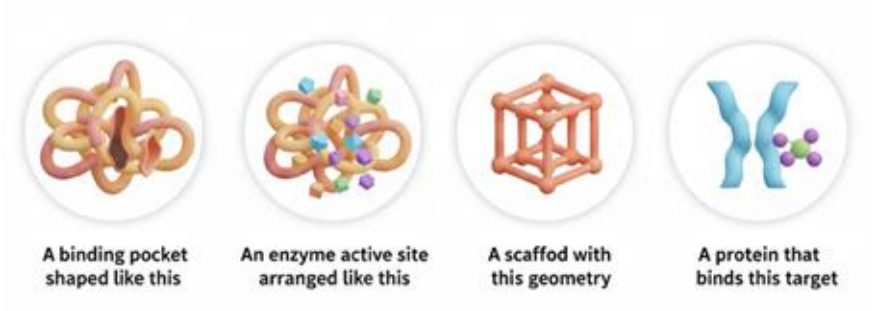
Determines

Constraints



In biology, We don't ask:

"I want this exact amino acid sequence."



What is the "Sequence" ?

The Core Problem Definition

: Given protein structure X

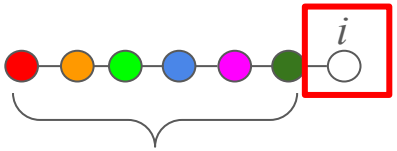
: Determine the Sequence S

The Core Problem Definition

Autoregressive Design

$$p(s|x) = \prod_i p(s_i | x, s_{<i})$$

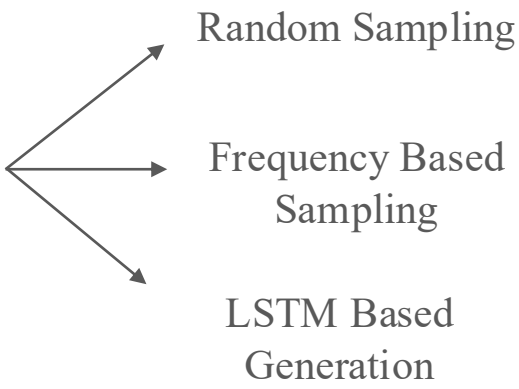
Whole 3D structure



Resolved sequence upto *i-1*

Sequence Sampling Approaches

Existing methods generate sequences without structural conditioning

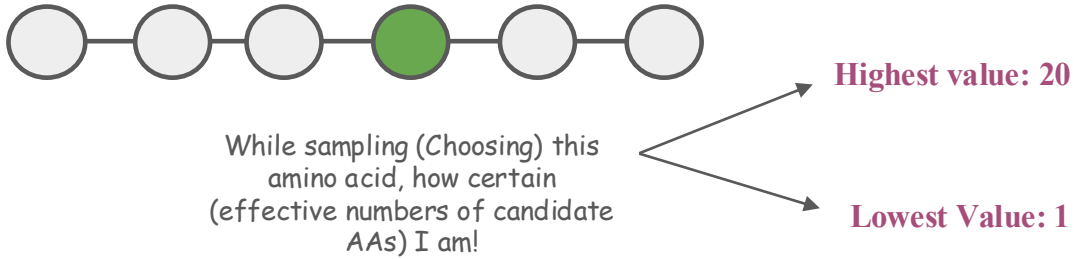


Quality of the Generated Sequences

$$\text{Perplexity} = \exp\left(-\frac{1}{N}\sum \log p(s_i)\right)$$

Perplexity measures ***how uncertain the model is*** when predicting each amino acid.

You can think of it as the effective number of amino acids the model is choosing between



Uniform Sampling

Table 1: **Null perplexities** for common statistical models of proteins.

Null model	Perplexity	Conditioned on
Uniform	20.00	-
Natural frequencies	17.83	Random position in a natural protein
Pfam HMM profiles	11.64	Specific position in a specific protein family

Choose Anything!
20 Options

Natural Frequency Bases Sampling

Table 1: **Null perplexities** for common statistical models of proteins.

Null model	Perplexity	Conditioned on
Uniform	20.00	-
Natural frequencies	17.83	Random position in a natural protein
Pfam HMM profiles	11.64	Specific position in a specific protein family

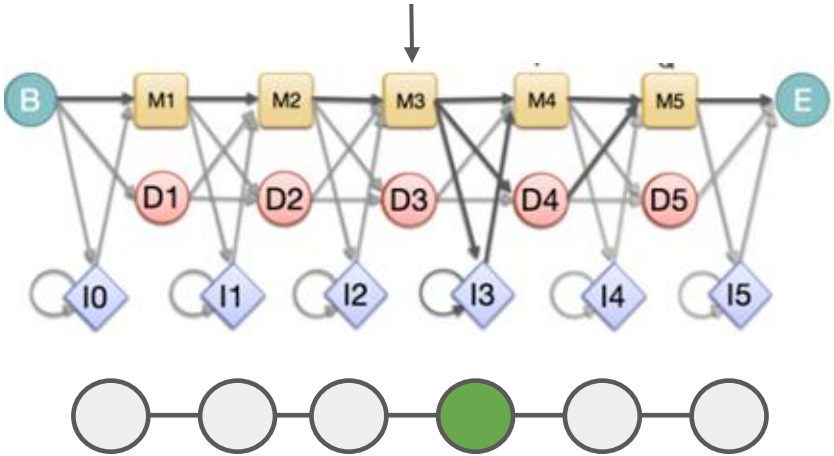
Choose following the natural AA frequency

Profile Based Sampling

Table 1: **Null perplexities** for common statistical models of proteins.

Null model	Perplexity	Conditioned on
Uniform	20.00	-
Natural frequencies	17.83	Random position in a natural protein
Pfam HMM profiles	11.64	Specific position in a specific protein family

This is expected (Natural)



Why Existing Methods Are Not Enough

Table 1: **Null perplexities** for common statistical models of proteins.

Null model	**Perplexity**	**Conditioned on**
Uniform	20.00	-
Natural frequencies	17.83	Random position in a natural protein
Pfam HMM profiles	11.64	Specific position in a specific protein family

Table 2: **Per-residue perplexities for protein language modeling** (lower is better). The protein chains have been cluster-split by CATH topology, such that test includes only unseen 3D folds. While a structure-conditioned language model can generalize in this structure-split setting, unconditional language models struggle.

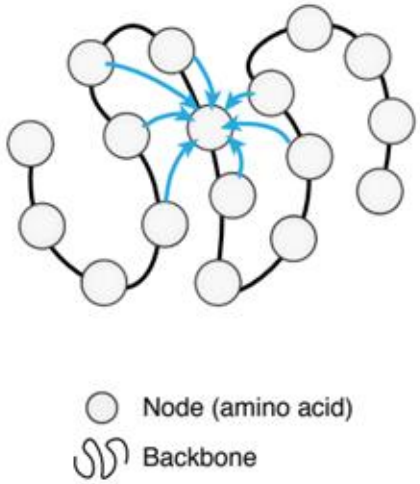
Test set	Short	Single chain	All
Structure-conditioned models			
Structured Transformer (ours)	**8.54**	**9.03**	**6.85**
SPIN2 [8]	12.11	12.61	-
Language models			
LSTM ($h = 128$)	16.06	16.38	17.13
LSTM ($h = 256$)	16.08	16.37	17.12
LSTM ($h = 512$)	15.98	16.38	17.13
Test set size	94	103	1120

Why Existing Methods Are Not Enough

Unconditional language models barely beat null models on *unseen folds*.

Hypothesis:
Structure conditioning is essential.

Key Approach (Representation)

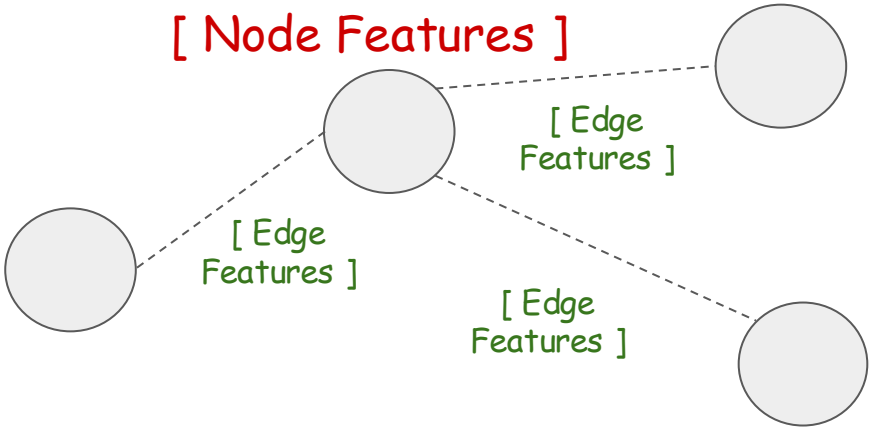


Long-range in sequence
→ Short-range in 3D

So instead of dense attention:

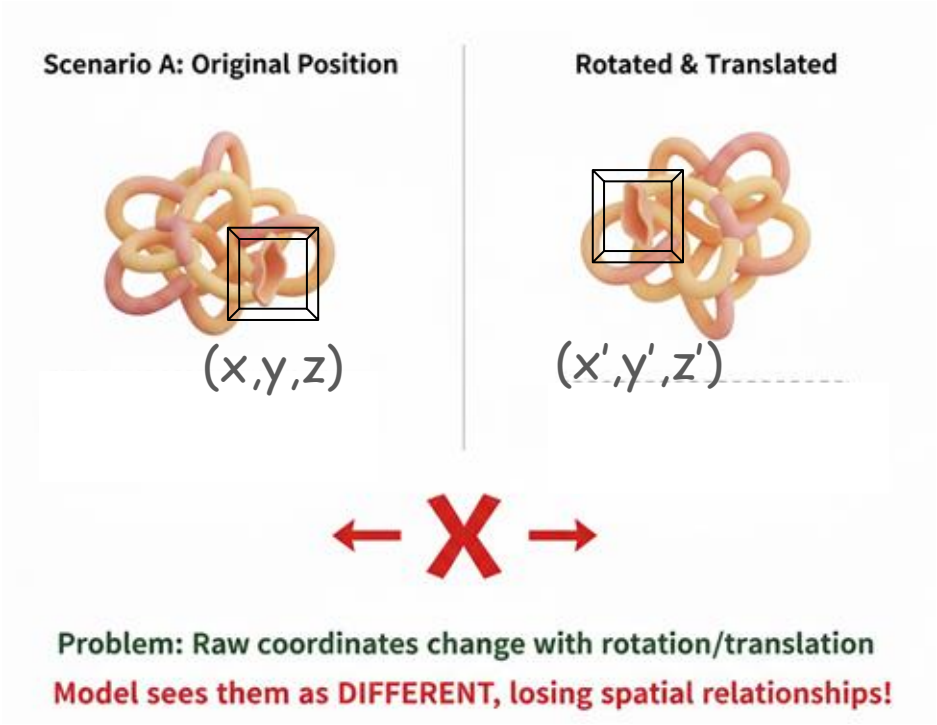
- Use **k-NN graph in 3D space**
- k = 30 neighbors
- $O(N^2) \rightarrow O(Nk)$

Protein as a Graph!

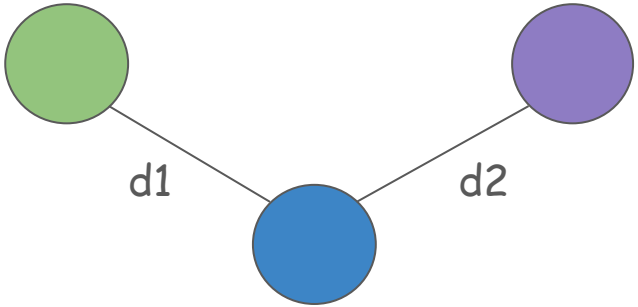


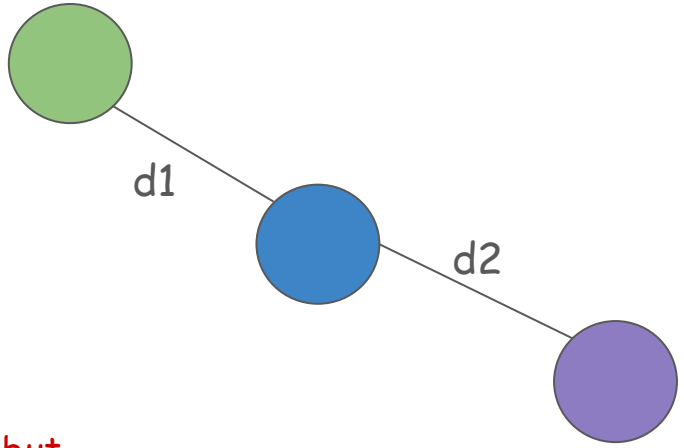
K= 30

Problem with raw node features



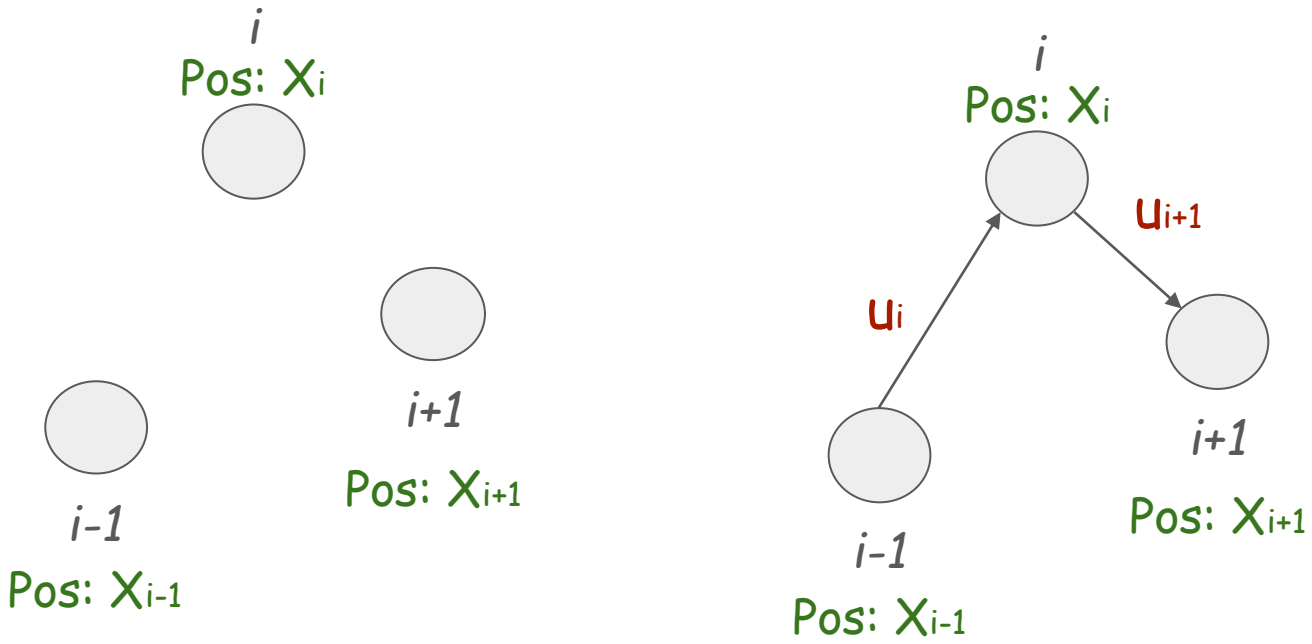
Problem with only distance as edge feature





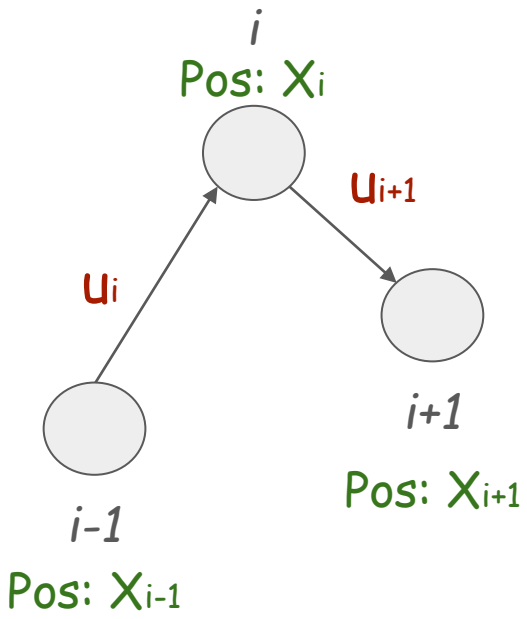
Distances are same, but geometry is completely different

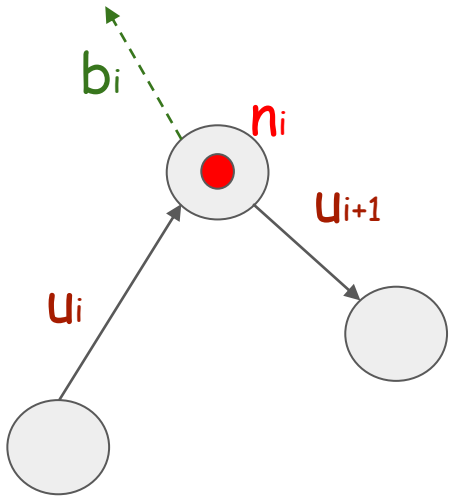
Locally aware node representation



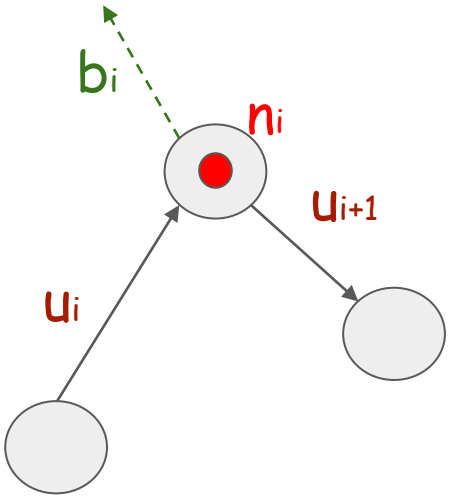
Locally aware node representation

$$u_i = \frac{x_i - x_{i-1}}{||x_i - x_{i-1}||}, \quad b_i = \frac{u_i - u_{i+1}}{||u_i - u_{i+1}||}, \quad n_i = \frac{u_i \times u_{i+1}}{||u_i \times u_{i+1}||}$$





Locally aware node representation



$$u_i = \frac{x_i - x_{i-1}}{\|x_i - x_{i-1}\|}$$

$$b_i = \frac{u_i - u_{i+1}}{\|u_i - u_{i+1}\|}$$

$$n_i = \frac{u_i \times u_{i+1}}{\|u_i \times u_{i+1}\|}$$

$$t_i = b_i \times n_i$$

$$O_i = \begin{bmatrix} | & | & | \\ b_i & n_i & t_i \\ | & | & | \end{bmatrix}$$

$$O_i = \begin{bmatrix} b_{i,x} & n_{i,x} & t_{i,x} \\ b_{i,y} & n_{i,y} & t_{i,y} \\ b_{i,z} & n_{i,z} & t_{i,z} \end{bmatrix}$$

Oi is a residue-centered orthonormal basis derived purely from backbone geometry.

What this rotation matrix does?

It transforms from:

Global coordinates → Local residue frame

When they compute:

$$O_i^T(x_j - x_i)$$

That means:

- Express the vector from i to j
- In i's local coordinate system

Node and Edge Features

✅ **Node features = backbone dihedral angles**

For each residue i:

- ϕ_i
- ψ_i
- ω_i

These are embedded as:

$$\{\sin, \cos\} \times (\phi_i, \psi_i, \omega_i)$$

So each angle becomes two values (sin and cos).

Total node feature dimension:

- 3 angles × 2 = 6 values per residue

Node and Edge Features

Edge Features

For each edge *(i, j)* :

Distance: $r(\|x_j - x_i\|)$

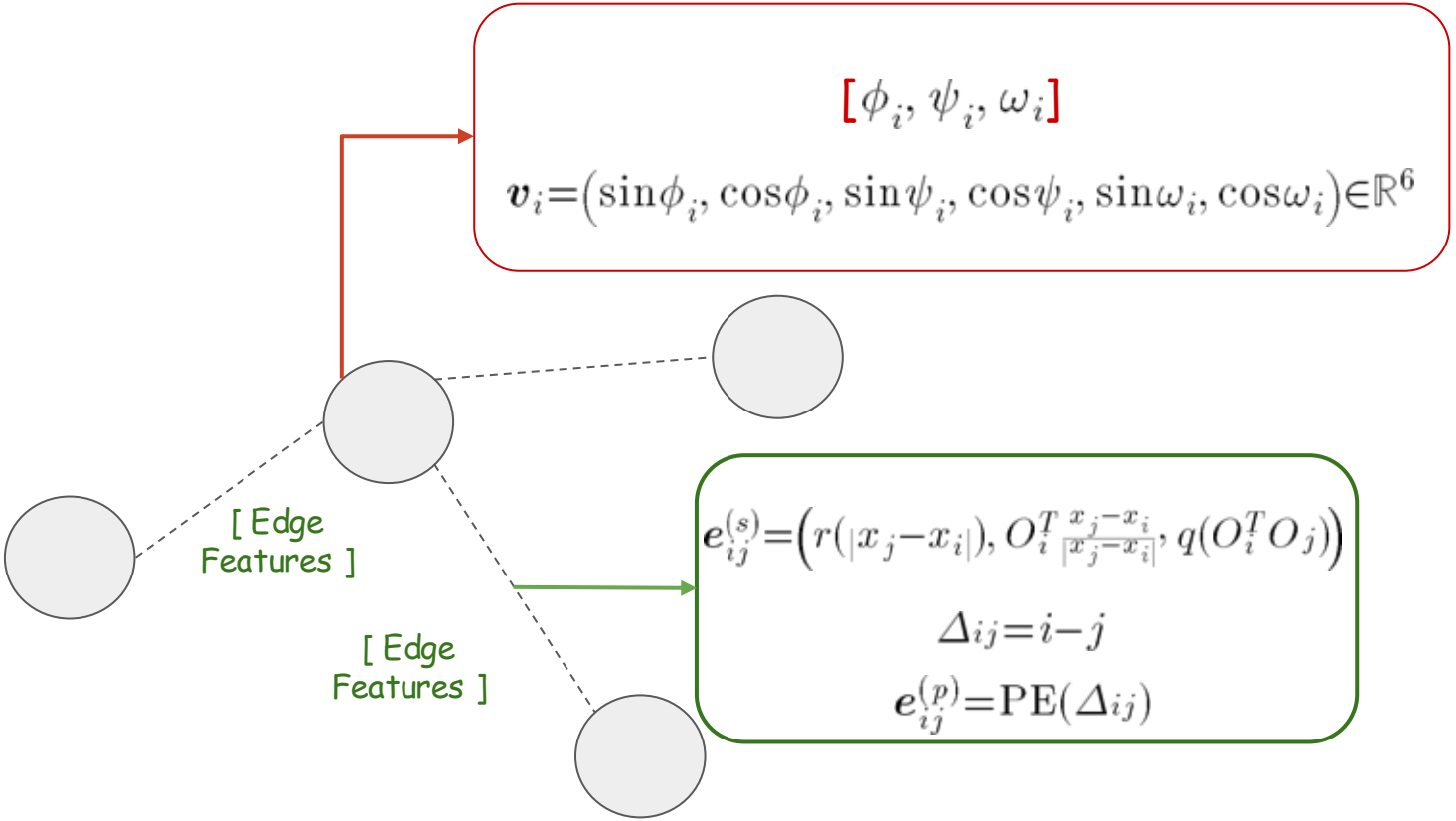
Direction: $O_i^T \frac{x_j - x_i}{\|x_j - x_i\|}$

Orientation: $q(O_i^T O_j)$

Also includes relative positional encoding. That is the index difference between node i and j:

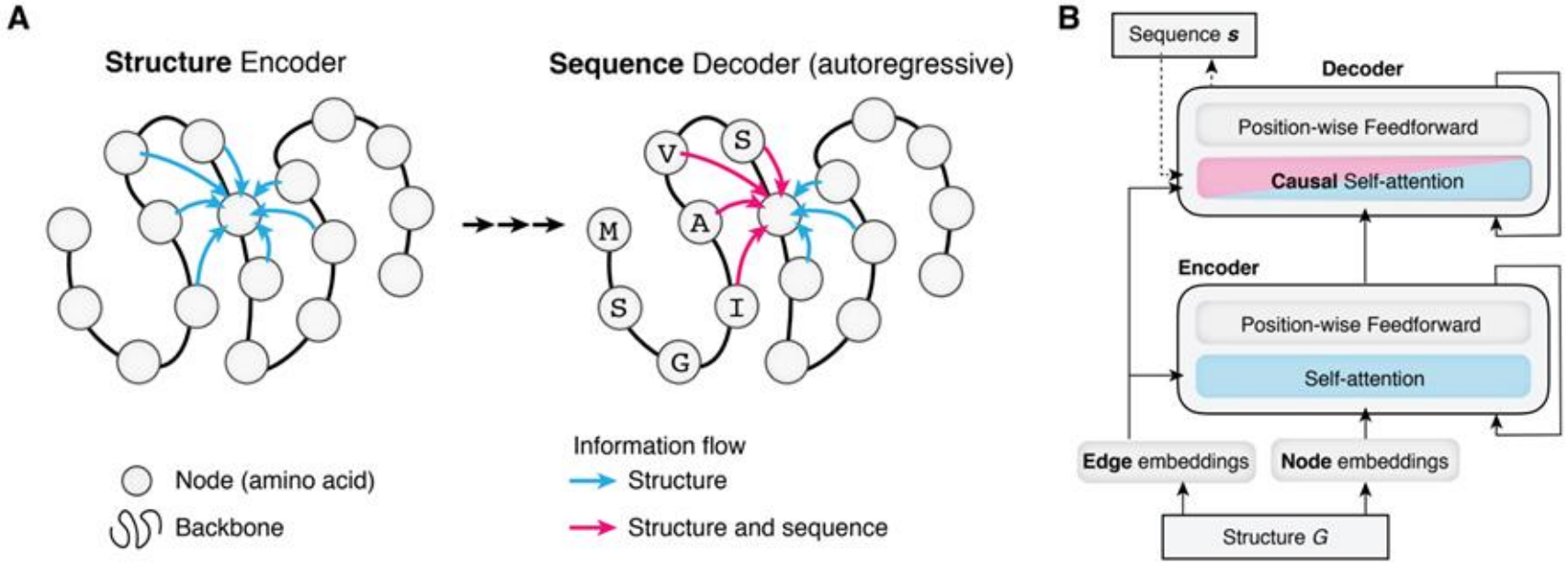
$\| i - j \|$

Final Representation

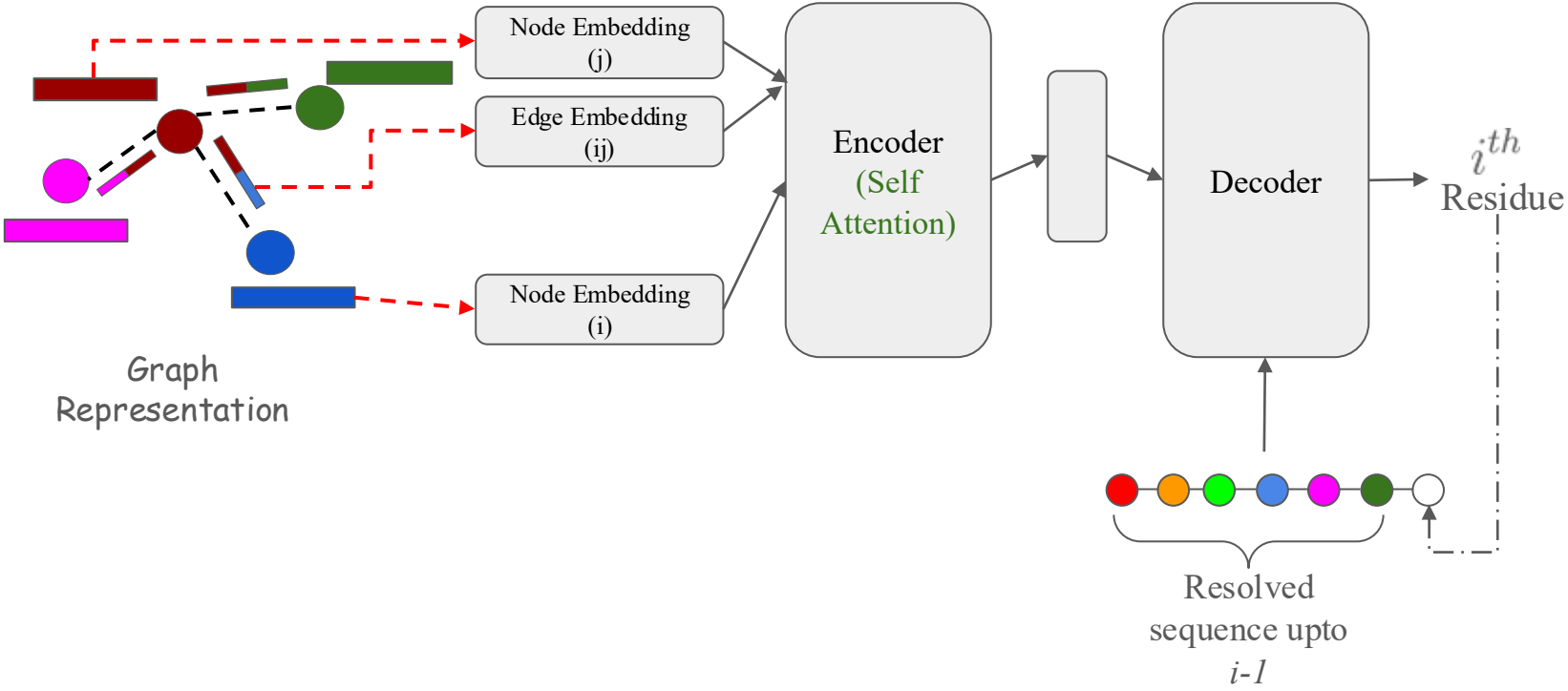


K= 30

Structured Transformer



Structure Transformer



Objective Function

- Autoregressive Factorization

$$p(s|x) = \prod_i p(s_i|x, s_{<i})$$

- Training Objective (Maximum Likelihood)

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \log p_\theta(s_i|x, s_{<i})$$

Dataset

🔷 Goal

Test whether the model generalizes to ***unseen protein folds***

🔷 Data Source

- CATH 4.2 structural classification
- 40% sequence identity non-redundant set
- Full chains up to length 500

🔷 Key Design Choice

🚨 **Zero fold overlap between splits**

- No CATH topology shared across train, val, test
- Test set contains completely unseen 3D folds

🔷 Final Dataset Size

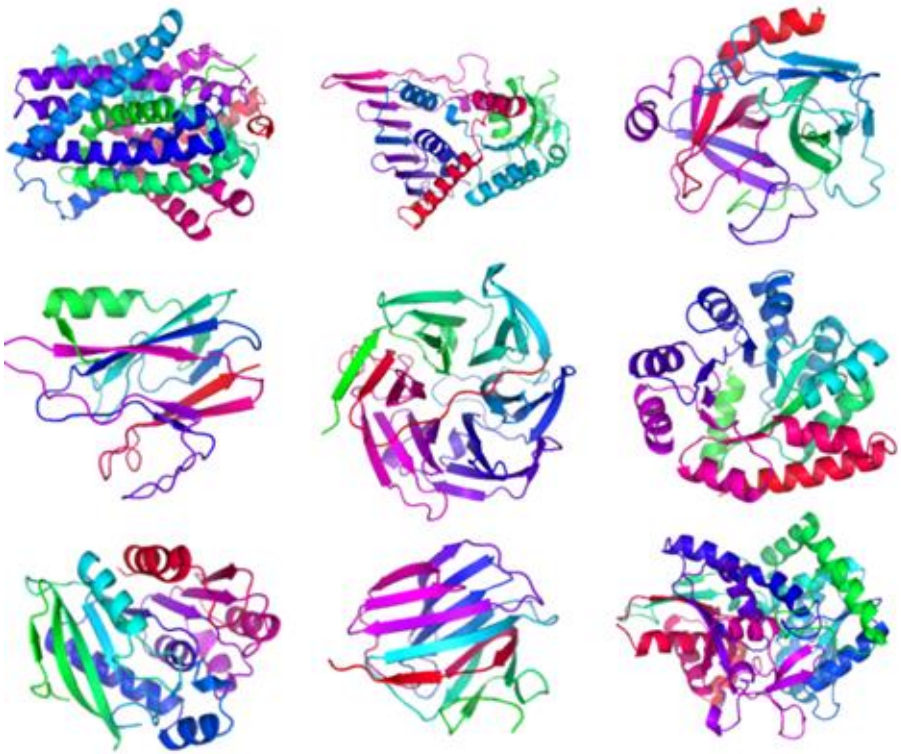
- Train: 18,024 chains
- Validation: 608 chains
- Test: 1,120 chains

Dataset

🔷 **Key Design Choice**

🚨 **Zero fold overlap between splits**

- No CATH topology shared across train, val, test
- Test set contains completely unseen 3D folds



Evaluation Metrics

1 Perplexity (Likelihood-Based Evaluation)

$$\text{Perplexity} = \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log p(s_i|x, s_{<i})\right)$$

- Measures how well the model assigns probability to true sequences
- ***Lower = better***
- Interpreted as “effective number of amino acids the model is confused between”

Tests structural generalization across unseen folds

Evaluation Metrics

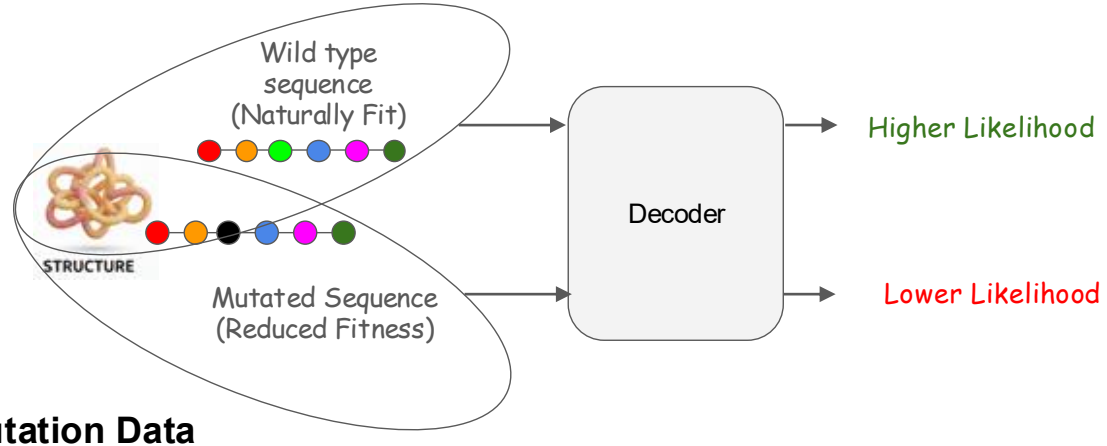
2 Native Sequence Recovery

$$\text{Recovery} = \frac{\#\text{ correctly predicted residues}}{\text{total residues}}$$

- Compare generated sequence to native sequence
- Measures how often the model recovers ground-truth residues
- Compared against Rosetta

Tests practical design accuracy

Evaluation Metrics



3 Correlation with Experimental Mutation Data

- Pearson correlation between model likelihood and mutational fitness

$$\log p(s \mid x) \quad \text{and} \quad \text{measured mutational fitness}$$

- Higher correlation = better biological relevance

→ Tests whether model likelihood reflects real stability/function

Structure Conditioning Dramatically Improves Perplexity

Test set	Short	Single chain	All
Structure-conditioned models			
Structured Transformer (ours)	**8.54**	**9.03**	**6.85**
SPIN2 [8]	12.11	12.61	-
Language models			
LSTM ($h = 128$)	16.06	16.38	17.13
LSTM ($h = 256$)	16.08	16.37	17.12
LSTM ($h = 512$)	15.98	16.38	17.13
Test set size	94	103	1120

What Makes the Model Strong?

Table 3: **Ablation of graph features and model components**. Test perplexities (lower is better).

Node features	Edge features	Aggregation	Short	Single chain	All
Rigid backbone					
Dihedrals	Distances, Orientations	Attention	8.54	9.03	6.85
Dihedrals	Distances, Orientations	PairMLP	**8.33**	**8.86**	**6.55**
C_α angles	Distances, Orientations	Attention	9.16	9.37	7.83
Dihedrals	Distances	Attention	9.11	9.63	7.87
Flexible backbone					
C_α angles	Contacts, Hydrogen bonds	Attention	11.71	11.81	11.51

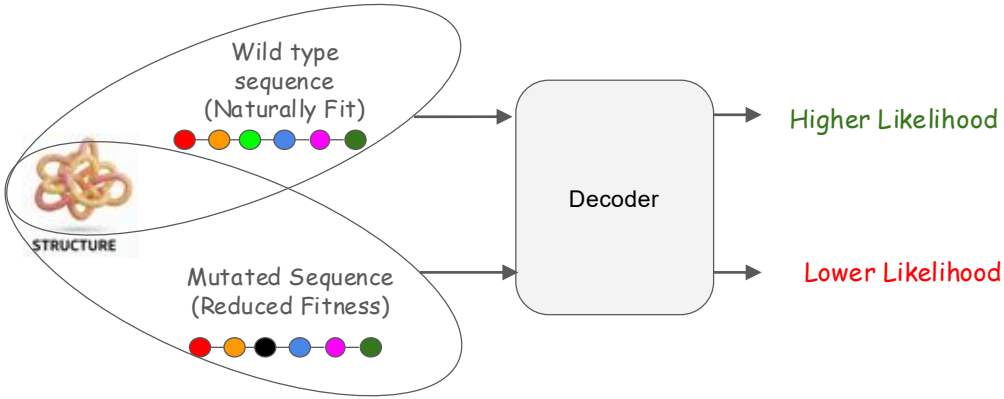
Improved Design Accuracy and Speed

Method	Recovery (%)	Speed (AA/s) CPU	Speed (AA/s) GPU
Rosetta 3.10 `fixbb`	17.9	4.88×10^{-1}	N/A
Ours ($T = 0.1$)	**27.6**	$\mathbf{2.22 \times 10^{2}}$	$\mathbf{1.04 \times 10^{4}}$

(a) Single chain test set (103 proteins)

Model Likelihood Reflects Biological Reality

Design	$\beta\beta\alpha\beta\beta_{37}$	$\beta\beta\alpha\beta\beta_{1498}$	$\beta\beta\alpha\beta\beta_{1702}$	$\beta\beta\alpha\beta\beta_{1716}$	$\alpha\beta\beta\alpha_{779}$
Rigid backbone	0.47	0.45	0.12	0.47	0.57
Flexible backbone	0.50	0.44	0.17	0.40	0.56



Conclusions & Takeaways

What the Paper Achieves

• Formulates protein design as conditional generation

$$p(s|x)$$

• Introduces a **geometry-aware Structured Transformer**

• Encodes 3D structure using invariant local coordinate frames

• Achieves:

- Dramatically lower perplexity (~7 vs ~17)
- Higher native sequence recovery than Rosetta
- Orders-of-magnitude faster inference
- Meaningful correlation with experimental mutation data

Conclusions & Takeaways

What was still missing at that time?

- Super large language models (like ProGEN-2), which can generate natural sequences without even structural data (If the goal is to only generate nature like sequences)
- Foldability is now confirmed mostly via AF3 pIDDT.
- No SE(3)-equivariant GNNs like EGNN, TFN, etc. They engineered invariance manually
- No joint structure + sequence generative models.

Thank You!