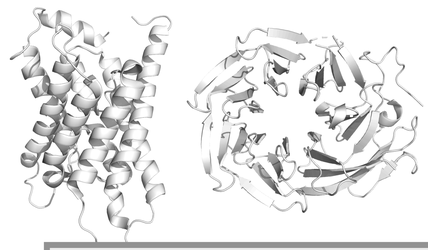
Generative Models for Graph-Based Protein Design

John Ingraham, Vikas K. Garg, Regina Barzilay, Tommi Jaakkola
{ingraham, vgarg, regina, tommi}@csail.mit.edu
MIT CSAIL, Cambridge MA, USA



Learning protein design, directly

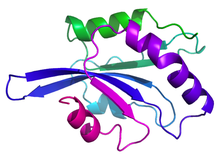
Sequence **s**

MSGIAVSDDCVQKFNELKLG

Protein folding

Protein design
"inverse folding"

Structure **x**



Opportunities: *de novo* therapeutics, catalysts, and materials

Challenges: Despite many successes from conventional methods such as Rosetta, first designs often fail (**unreliability**), outcomes are sensitive to methodology **(non-robustness)**, and design throughput is **slow**

... could we learn to generate designs directly?

The protein *sequence design* problem

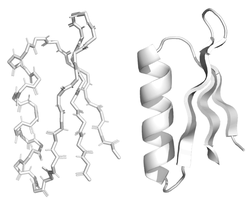
Given a 3D protein structure **x**, what sequence(s) **s** will fold into it?

Input: 3D backbone

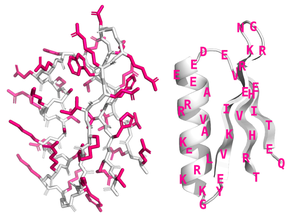
$p(\mathbf{s}|\mathbf{x})$

Output: Sequence of amino acids along backbone

QETIEVEDEEEARRVAKELRKKGYEVKIERRGNKWHVHRT



Atoms *Cartoon*

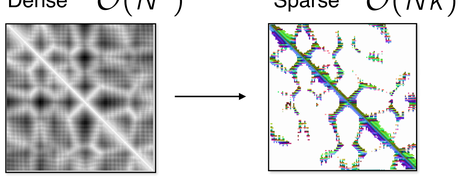


Atoms *Cartoon*

Most sequence dependencies are *spatially local*

Spatial adjacency gives the relevant context for sequences

Dense $\mathcal{O}(N^2)$ → Sparse $\mathcal{O}(Nk)$



Distances

k-NN Orientations, $k = 30$

Evolutionary evidence: sequence covariation predicts spatial contacts [Marks et al 2012]

Engineering evidence: Rosetta uses *spatially local* pairwise Markov Random Fields [Leaver-Fay et al 2011]

Implication:
Structure the computation to focus on **spatial interaction**

Structure $\mathbf{x}$ → Attributed Graph $\mathcal{V}(\mathbf{x}), \mathcal{E}(\mathbf{x})$ → Sequence $\mathbf{s}$

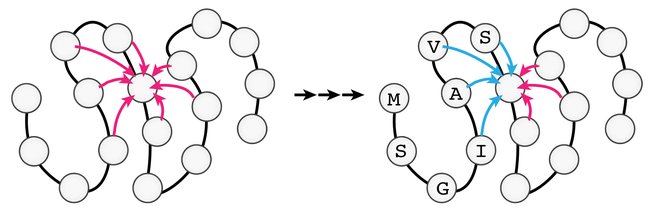
Approach: Graph-based sequence generation

1. Represent **structure as graph**
2. **Language modeling** decoder

Structure $\mathbf{x}$ → Attributed Graph $\mathcal{V}(\mathbf{x}), \mathcal{E}(\mathbf{x})$ → Sequence $\mathbf{s}$

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

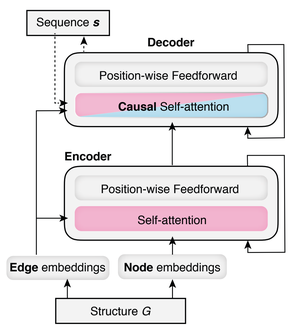
Structure Encoder → **Sequence** Decoder (autoregressive)



Node (amino acid)
Backbone

Information flow
Structure
Structure and sequence

Local attention builds up context for structure (and sequence)



Multi-head attention like Transformer [Vaswani et al 2017, Shaw et al 2018] with node embeddings $\mathbf{h}_i$ and edges $\mathbf{r}_{ij}$

$\mathbf{h}_i$ → Queries $\quad$ $\mathbf{r}_{ij}$ → Keys, Values

Encoder attends to node & edge tuples:

$$\mathbf{r}_{ij}^{(\text{enc})} = (\mathbf{h}_j^{(\text{enc})}, \mathbf{e}_{ij})$$

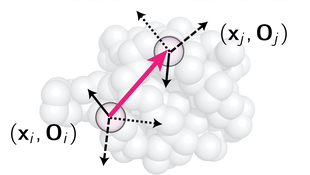
Decoder attends to node, edge, sequence triples with masking:

$$\mathbf{r}_{ij}^{(\text{dec})} = \begin{cases} (\mathbf{h}_j^{(\text{dec})}, \mathbf{e}_{ij}, \mathbf{g}(s_j)) & i > j \\ (\mathbf{h}_j^{(\text{enc})}, \mathbf{e}_{ij}, \mathbf{0}) & i \leq j \end{cases}$$

Graph features represent molecular geometry

Goal: Expressive but invariant descriptors

Point cloud with local frames



Relations

Distance Direction Rotation

Node features: Dihedral angles of backbone

Edge features are relative transformations between frames (SE(3)-invariant)

Translation $\mathbf{O}_i^T(\mathbf{x}_j - \mathbf{x}_i)$

Distance $||\mathbf{x}_j - \mathbf{x}_i||$ → Radial basis

Direction $\mathbf{O}_i^T \frac{\mathbf{x}_j - \mathbf{x}_i}{||\mathbf{x}_j - \mathbf{x}_i||}$ → Unit vector

Rotation $\mathbf{O}_j^T \mathbf{O}_i$ → Quaternion

Result: Improved speed and accuracy over conventional methods

Our single chain test set (103 structures)

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**	$\mathbf{2.22 \times 10^{2}}$	$\mathbf{1.04 \times 10^{4}}$

Ollikainen *et al* Benchmark (40 structures; re-split training for 0 topology overlap)

Method	Recovery (%)
Rosetta, `fixbb` 1	33.1
Rosetta, `fixbb` 2	38.4
Ours ($T = 0.1$)	**39.2**

- ~ **400x** speedup on one core of CPU
  ~ **20,000x** speedup GPU
- Why so low? This set contains many NMR structures (rather than X-ray) for which conventional methods are not robust

Result: Structure-conditioned language models can generalize to unseen 3D structures

Dataset creation

CATHdb 40%NR
↓
Full chains, 500 AA
↓
Split by *topology*
↓
~18,000 chains
In train

Perplexity (per amino acid)

Test set	Short	Single chain	All
Structure-conditioned models			
Structured Transformer (ours)	**8.54**	**9.03**	**6.85**
SPIN2	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

- Significant boost in statistical performance vs other neural method
- Why so low? Sequences in test are from different *fold topologies*

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

Result: Comparison of features and architecture

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

- Simpler message passing aggregation offers room for improvement (Thanks, reviewer!)

Conclusion: Deep generative models can learn to design proteins directly from structure

References

1. O'Connell, James, et al. "SPIN2: Predicting sequence profiles from protein structures using deep neural networks." 2018
2. Conchúir, Shane Ó., et al. "A web resource for standardized benchmark datasets, metrics, and Rosetta protocols for macromolecular modeling and design." 2015
3. Leaver-Fay, Andrew, et al. "ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules." 2011
4. Vaswani, Ashish, et al. "Attention is all you need." 2017
5. Marks, Debora S., Thomas A. Hopf, and Chris Sander. "Protein structure prediction from sequence variation." 2012