CarbonNovo: Joint Design of Protein Structure and Sequence Using a Unified Energy-based Model

Authors: Milong Ren[1,2], Tian Zhu[1,2], Haicang Zhang[1,2,#]

Presenter: Milong Ren

1. Institute of Computing Technology, Chinese Academy of Sciences.
2. University of Chinese Academy of Sciences.

Background

- Diffusion-based generative models like RFdiffusion show great promise in structure design, they face inherent limitations within the two-stage framework.

- First, the sequence design module risks overfitting as the accuracy of the gen- erated structures may not align with that of the crystal structures used for training.

- Second, the sequence design module lacks interaction with the structure design module to further optimize the generated structures.

Contribution

- We develop CarbonNovo, a unified framework capable of simultaneously generating sequences and structures for general protein families.

- We are the first to integrate a protein language model to enhance the generation of both protein structure and sequences.

- We explore various techniques for efficient training and inference of the joint model, such as a multi-stage training strategy and the discrete version of M-H Langevin algorithm for sequence sampling.

- CarbonNovo demonstrates superior performance compared to two-stage approaches across various metrics, including designability, novelty, Rosetta energy, and sequence plausibility.

Methods

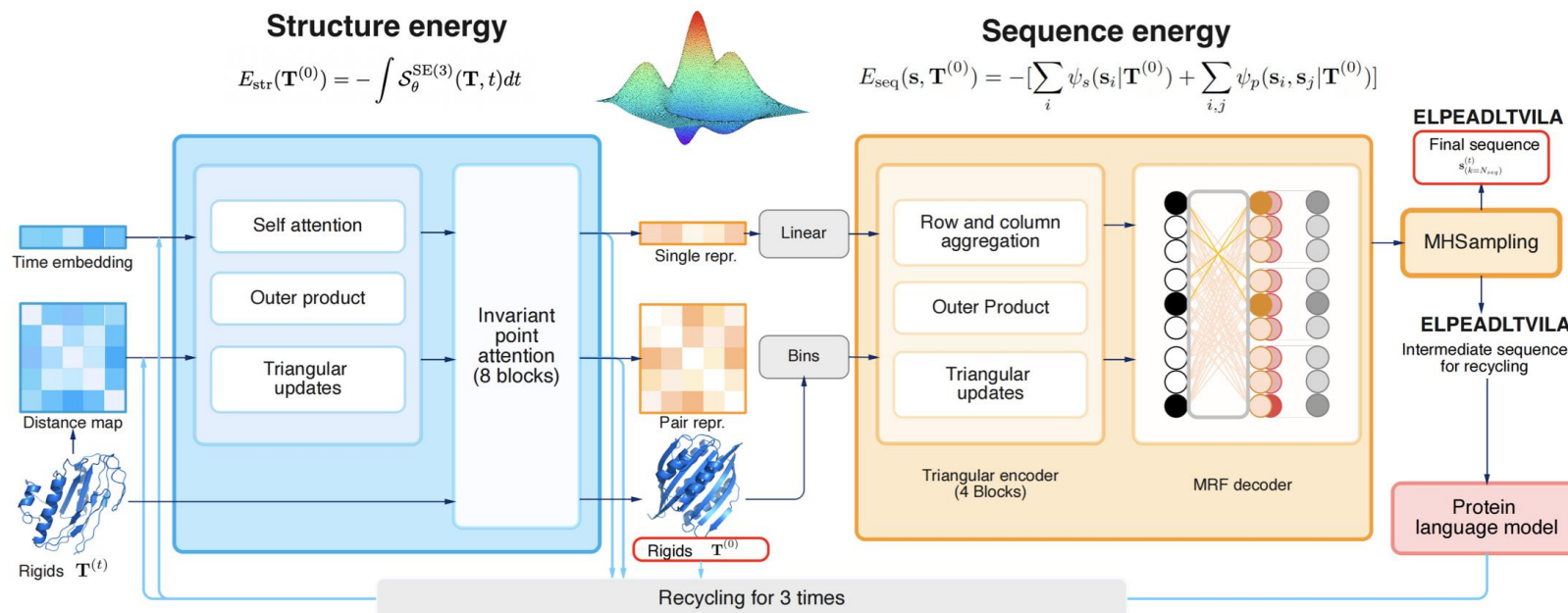- Unified Energy-based Model for Sequence and structure

Figure 1. CarbonNovo architecture which jointly generates protein backbone structure and sequence.

4

Methods

- Structure sampling

$$q_{\text{str}}(\mathbf{T}^{(t-\Delta t)}|\mathbf{T}^{(t)}) = q_{\text{str}}(\mathbf{R}^{(t-\Delta t)}|\mathbf{R}^{(t)})q_{\text{str}}(\mathbf{t}^{(t-\Delta t)}|\mathbf{t}^{(t)}),$$

$$q_{\text{str}}(\mathbf{R}^{(t-\Delta t)}|\mathbf{R}^{(t)}) \sim \mathcal{IG}_{SO(3)}(\Delta t \mathcal{S}_\theta^{\mathbf{R}}(\mathbf{R}^{(t)}), \Delta t \text{Id})^{\otimes N},$$

$$q_{\text{str}}(\mathbf{t}^{(t-\Delta t)}|\mathbf{t}^{(t)}) \sim P\mathcal{N}(\mu_\theta, \Delta t \text{Id}_3)^{\otimes N}, \qquad (8)$$

$$\mu_\theta = \frac{1}{2}\Delta t \cdot \mathbf{t}^{(t)} + \Delta t \cdot \mathcal{S}_\theta^{\mathbf{t}}(\mathbf{t}^{(t)}).$$

- Sequence Sampling:

We obtain the initial sequence $\mathbf{s}_{(0)}^{(t)}$ only from the single representation $\mathbf{r}^s$. The sequence proposal distribution $q_{\text{seq}}(\mathbf{s}_{(k+1)}^{(t)}|\mathbf{s}_{(k)}^{(t)}, \mathbf{T}_\theta^{(0)})$ is as follows:
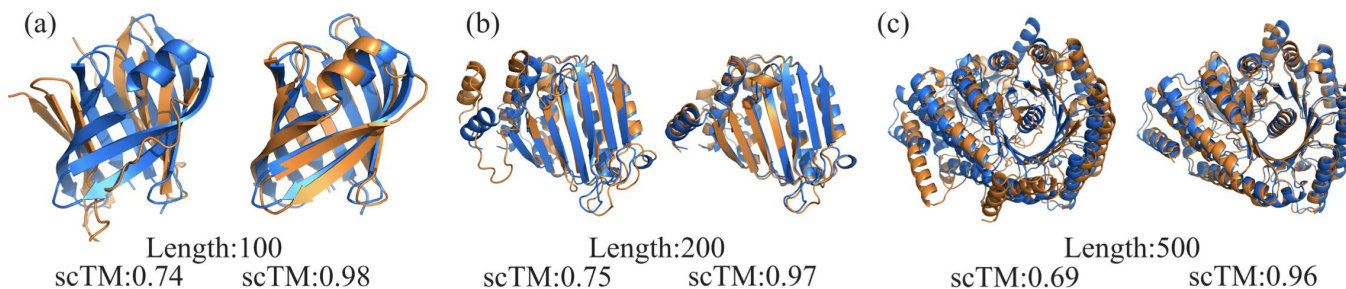
$$q_{\text{seq}}(\mathbf{s}_{(k+1)}^{(t)}|\mathbf{s}_{(k)}^{(t)}, \mathbf{T}_\theta^{(0)}) \sim \text{Categorical}\left(\mathbf{M}^{\text{seq}}\right),$$

$$\mathbf{M}^{\text{seq}} = \text{Softmax}\left(\frac{1}{2}\nabla E_{\text{seq}}(\mathbf{s}_{(k)}^{(t)}, \mathbf{T}_\theta^{(0)})\Delta\mathbf{s} - \frac{\Delta\mathbf{s}^2}{2\gamma}\right), \quad (9)$$

$$\Delta\mathbf{s} = \mathbf{s}_{(k+1)}^{(t)} - \mathbf{s}_{(k)}^{(t)}.$$

Results

- ## Experiment 1: Sequence and Structure Co-design

Table 1. Evaluation of Designability, Diversity, and Novelty. As for the designability and novelty metrics, the results are presented with structures predicted by ESMFold on the left and OmegaFold on the right of the slash line.
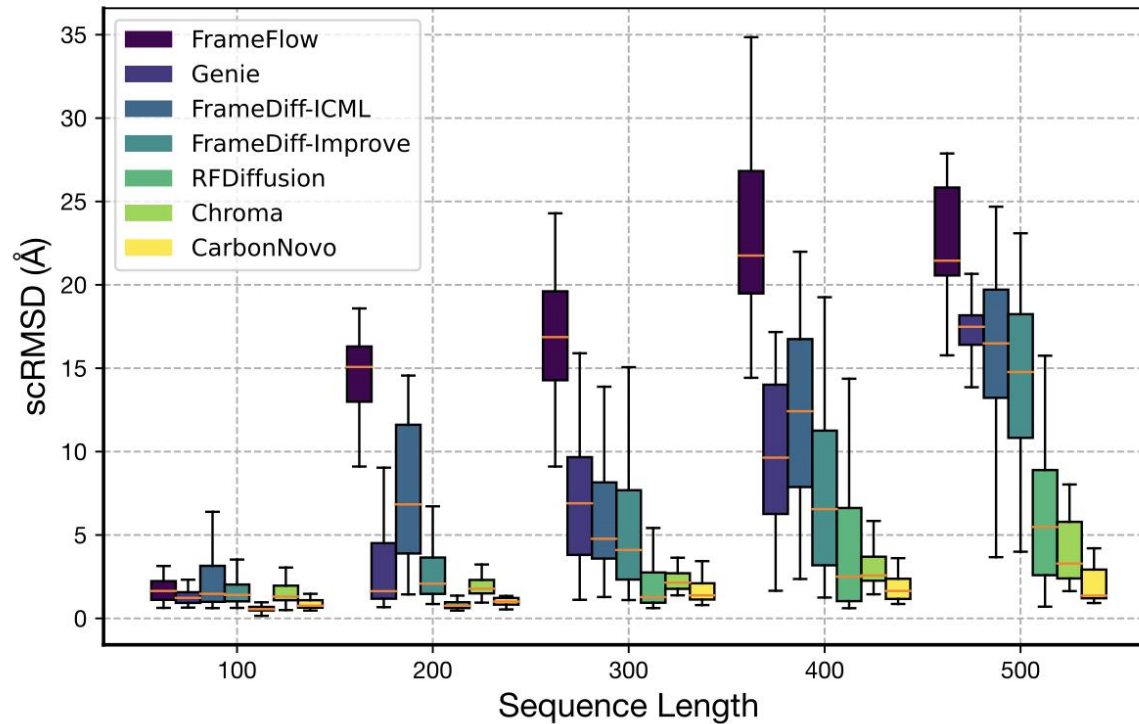
	Designability			Diversity ($\downarrow$)	Novelty ($\uparrow$)
	scRMSD ($\downarrow$)	scTMscore ($\uparrow$)	Fraction ($\uparrow$)		
RFdiffusion	3.494 / 3.891	0.897 / 0.753	69.81% / 60.00%	0.221	36.50% / 32.65%
Genie	7.581 / 8.929	0.672 / 0.589	31.43% / 28.75%	0.229	16.50% / 13.35%
FrameDiff-ICML	8.197 / 9.768	0.656 / 0.498	19.96% / 15.63%	0.239	5.31% / 4.65%
FrameDiff-Improve	6.524 / 6.793	0.755 / 0.629	27.81% / 28.75%	0.279	6.31% / 5.60%
FrameFlow	18.827 / 26.02	0.320 / 0.285	11.88% / 11.25%	0.275	5.60% / 4.65%
Chroma v1 (GitHub)	3.209 / 3.620	0.868 / 0.742	45.70% / 40.10%	**0.204**	25.62% / 23.92%
CarbonNovo+MPNN	2.431 / 2.541	0.917 / 0.834	73.16% / 70.15%	0.217	39.75% / 36.94%
CarbonNovo (default)	**1.943 / 1.990**	**0.924 / 0.859**	**81.38% / 77.38**%	0.217	**43.15% / 40.92%**

(a) Length:100 scTM:0.74 scTM:0.98

(b) Length:200 scTM:0.75 scTM:0.97

(c) Length:500 scTM:0.69 scTM:0.96
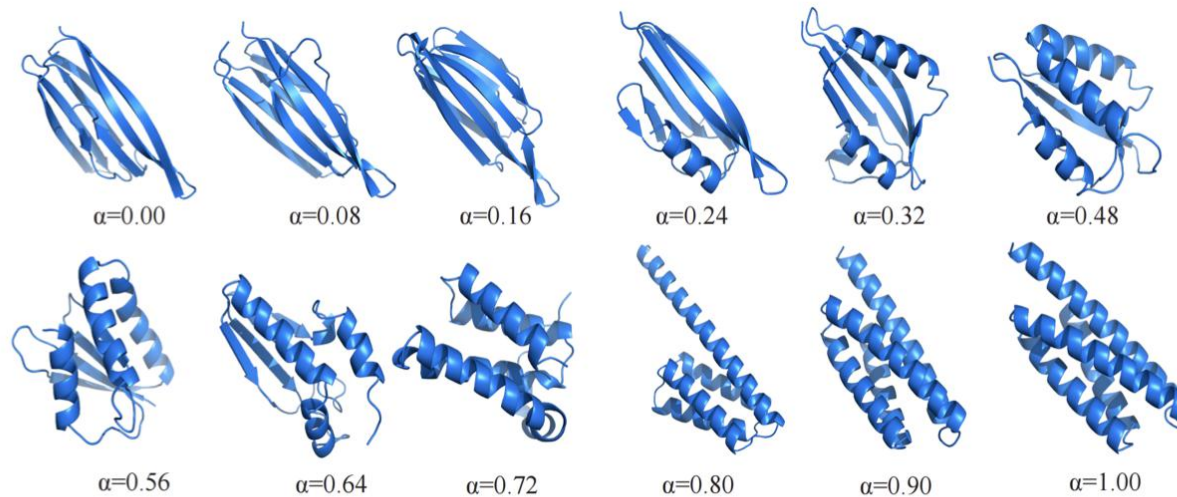
Results

- Experiment 1:

Figure 3. scRMSD of designed proteins vs. predicted proteins under various length.

Results

- Experiment 2:

Figure 4. An example of structure morphing: starting from the top left, a protein consisting solely of beta sheet secondary structures gradually transitions to a protein with only alpha-helices in the bottom right.

Thanks