中國人民大學

RENMIN UNIVERSITY OF CHINA

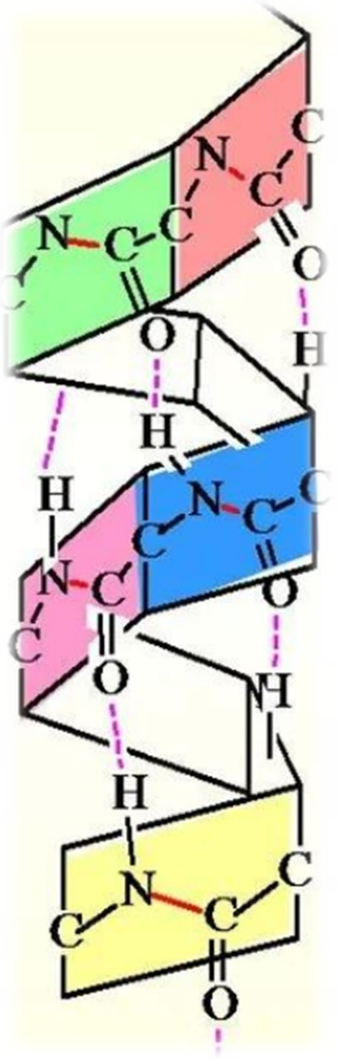
ProteinWeaver: A Divide and Assembly Approch for Protein Backbone Design

Paper Review: Yitian Wang

Research Background

- Current Challenges in Protein Backbone Design
  - Advances in protein backbone design have enabled the generation of novel and diverse structures, but designability decreases significantly as the protein backbone length increases
  - Existing methods such as RFdiffusion fail to effectively capture inter-domain interactions, particularly for long-chain proteins. This results in lower structural quality and less functional diversity for complex multi-domain backbones

- Nature-Inspired Strategy: Nature uses a “divide-and-assembly” approach to construct diverse and complex protein structures by recombining a limited number of building blocks (protein domains)



Research Background

- Objective of ProteinWeaver
  - Develop a two-stage framework to enable flexible assembly of protein domains into high-quality, novel backbones
  - Address the limitations of current methods, particularly in long-chain protein backbone design

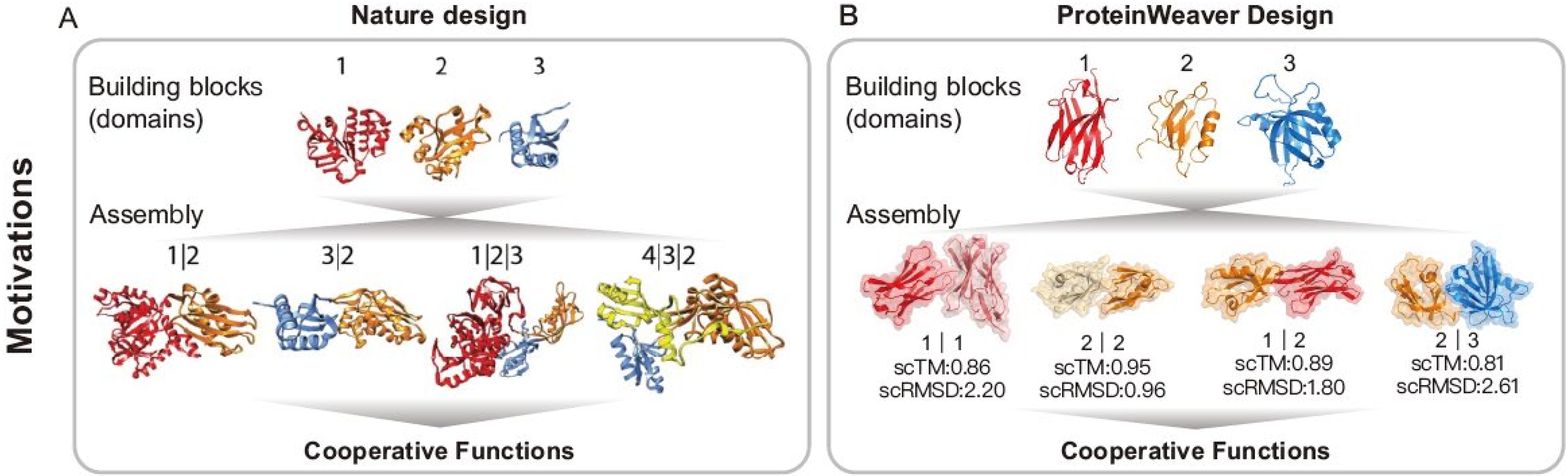
Method Overview

Framework Architecture

- Domain Generation (Divide): $\bar{S}$
  - Protein backbones are divided into multiple domains, with each domain being independently generated
  - FrameDiff, Chroma, and RFdiffusion can be applied here to generate individual domains

- Domain Assembly (Assembly): S
  - The generated domains are flexibly assembled using a SE(3) diffusion model, which learns inter-domain spatial relationships and interactions
  - Introduce Preference Alignment to optimize the interaction between domains by conducting pairwise comparisons of generated structures

Method Overview



Divide and Assembly Diffusion Framework

Protein Backbone Representation

- Following AlphaFold2 (Varadi et al., 2022):

$$\mathbf{T} = [T_1, T_2, ..., T_L] \in \mathrm{SE}(3)^L \qquad T_i = (r_i, x_i) \qquad r_i \in \mathrm{SO}(3) \qquad x_i \in \mathbb{R}^3$$

$$\mathrm{N}^*, \mathrm{C}_\alpha^*, \mathrm{C}^*, \mathrm{O}^* \in \mathbb{R}^3, \text{ with } \mathrm{C}_\alpha^* = (0, 0, 0)$$

$$\mathbf{S}_i = [\mathrm{N}, \mathrm{C}_\alpha, \mathrm{C}, \mathrm{O}]^i = T_i \circ [\mathrm{N}^*, \mathrm{C}_\alpha^*, \mathrm{C}^*, \mathrm{O}^*] \in \mathbb{R}^{4 \times 3}$$

$$\mathbf{S} \in \mathbb{R}^{L \times 4 \times 3}$$

$$\text{TM-score}(\mathbf{T}_{\text{predicted}}, \mathbf{T}_{\text{target}}) = \max \left(\frac{1}{L_{\text{target}}} \sum_{i=1}^{L_{\text{aligned}}} \frac{1}{1 + \left(\frac{d_i}{d_0}(L_{\text{target}}) \right)^2} \right),$$

Divide and Assembly Diffusion Framework

- Divided Domain Generation

$$D = \{D_1, D_2, \cdots, D_m\} \quad D_i \cap D_j = \varnothing \quad \bigcup_{j=1}^{m} D_j = D$$

Given L_i, $f_\theta : \mathbb{N}^+ \rightarrow \mathbb{R}^{L_i \times 3 \times 4}$, generates individual domain $\overline{S}_{D_i}$

  - FrameDiff, Chroma, and RFdiffusion can be applied here to generate individual domains

Divide and Assembly Diffusion Framework

- Domain Assembly Generation

$\{\bar{\mathbf{S}}_{D_1}, \bar{\mathbf{S}}_{D_2}, ..., \bar{\mathbf{S}}_{D_m}\}$ extracting C_α distance maps $\{\bar{\mathbf{M}}_{D_1}, \bar{\mathbf{M}}_{D_2}, ..., \bar{\mathbf{M}}_{D_m}\}$

$$\bar{\mathbf{M}} = \mathrm{SDM}(\bar{\mathbf{M}}_{D_1}, \bar{\mathbf{M}}_{D_2}, ..., \bar{\mathbf{M}}_{D_m}) = \begin{bmatrix} \bar{\mathbf{M}}_{D_1} & -\mathbf{1} & \cdots & -\mathbf{1} \\ -\mathbf{1} & \bar{\mathbf{M}}_{D_2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & -\mathbf{1} \\ -\mathbf{1} & \cdots & -\mathbf{1} & \bar{\mathbf{M}}_{D_m} \end{bmatrix}$$

- SE(3) Diffusion Model: $(\hat{\mathbf{T}}^{(0)}, \hat{\psi}) = g_\phi(\mathbf{T}^{(t)}, t, \bar{\mathbf{M}}),$

- Finally obtain the backbone coordinates $\mathbf{S}$ based on $[\mathrm{N}^*, \mathrm{C}_\alpha^*, \mathrm{C}^*, \mathrm{O}^*]$ by applying $\hat{\mathbf{T}}^{(0)}$ and rotation angle $\hat{\psi}$

Method Overview

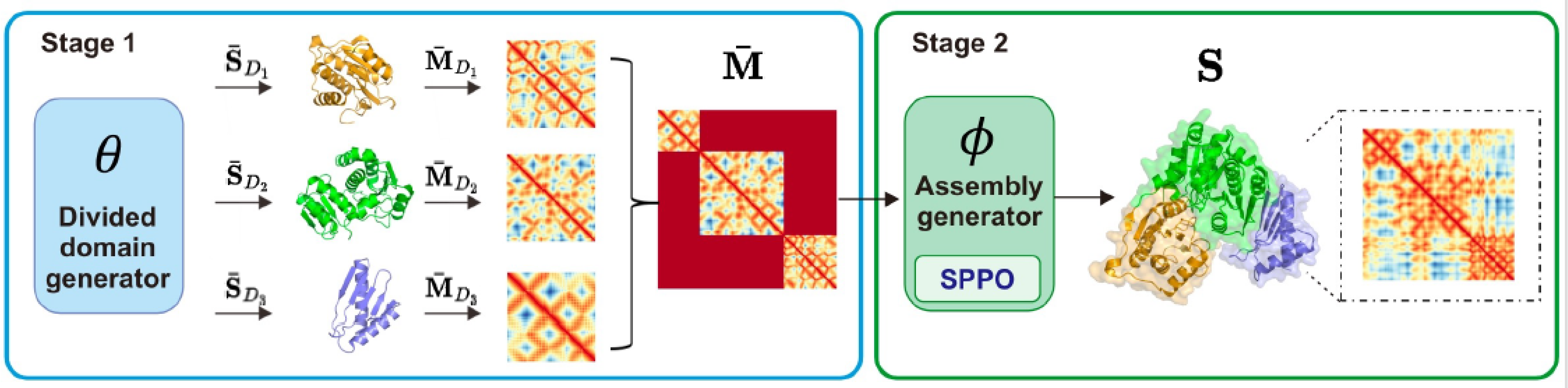


Figure 2: ProteinWeaver employs a two-staged 'divide-and-Assembly' framework, first generating individual protein domains and then using an SE(3) diffusion model to flexibly assemble these domains. $\bar{\mathrm{S}}$ represents isolated domains undergoing internal structural modifications for assembly into integrated backbones.

Training

- Dataset: Protein Data Bank
  - single-chain monomers between length 60 and 512 with resolution <5Å
- Pretraining:
  - refolded each domain by ESMFold to mimic their unassembled states for training
  - adopted the training loss from FrameDiff:
    - diffusion score-matching loss for translation and rotation
    - auxiliary losses related to the coordinate and pairwise distance loss on backbone atoms (t<0.25)

$$\mathcal{L} = \mathcal{L}_{\text{trans}} + 0.5 \cdot \mathcal{L}_{\text{rot}} + 0.25 \cdot \mathcal{L}_{\text{atom}}^{t<0.25} + 0.25 \cdot \mathcal{L}_{\text{pairwise}}^{t<0.25}.$$

Training

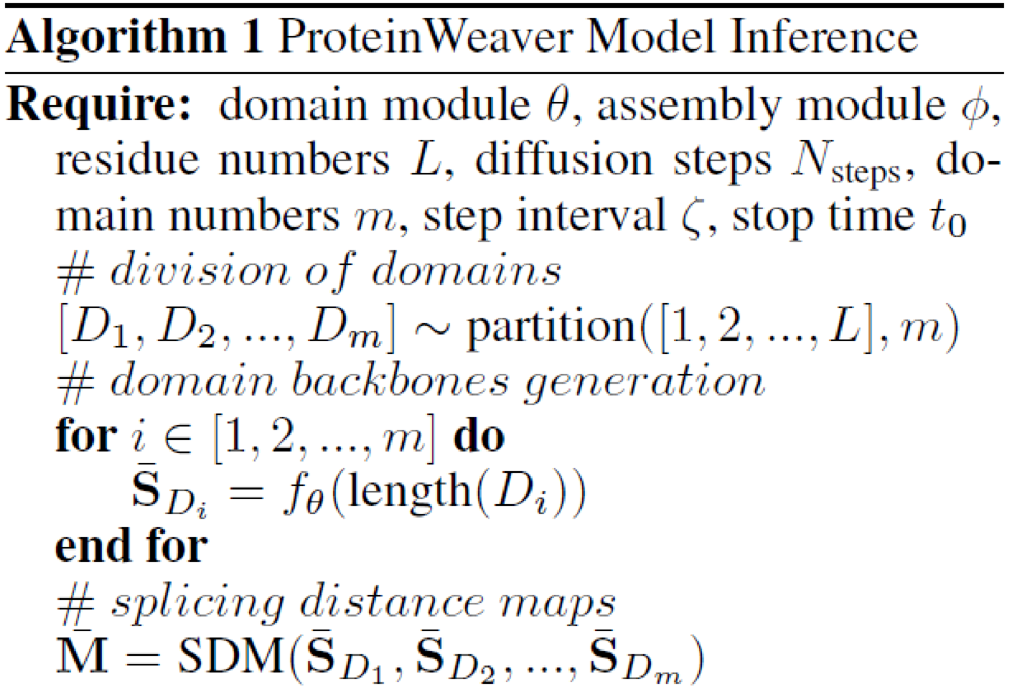
- Preference Alignment (Wu et al., 2024b)

$$\max_{\pi_\phi} \mathbb{E}_{\mathbf{S}_{\text{ref}}\sim\Omega, \bar{\mathbf{M}}=\text{SDM}(\mathbf{S}_{\text{ref}}), \mathbf{S}\sim\pi_\phi(\mathbf{S}|\bar{\mathbf{M}})}[r(\mathbf{S}, \mathbf{S}_{\text{ref}})] - \beta\mathbb{D}_{\text{KL}}[\pi_\phi(\mathbf{S}|\bar{\mathbf{M}})||\pi_{\text{ref}}(\mathbf{S}|\bar{\mathbf{M}})]. \quad (4)$$

- Dataset preparation:
  - For each spliced distance map, ProteinWeaver generates 3 structures
  - Use scTM scores to rank the generated structures and identify the "winner" structure (Sw) and the "loser" structure (Sl).
  - Constructe 10,000 data pairs of winner and loser structures for training the SPPO alignment model

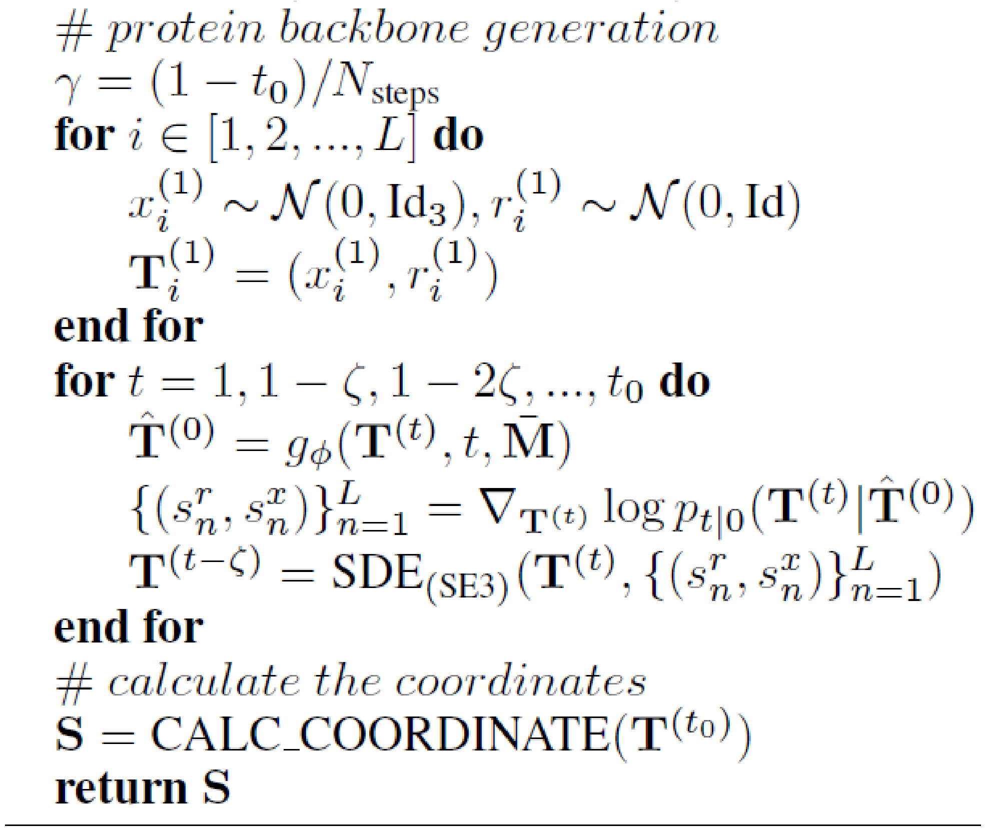
$$\mathcal{L}_{\text{SPPO}}(\bar{\mathbf{M}}, \mathbf{S}_w, \mathbf{S}_l; \pi_\phi, \pi_{\text{ref}}, \beta) := \left(\beta \log \frac{\pi_\phi(\mathbf{S}_w|\bar{\mathbf{M}})}{\pi_{\text{ref}}(\mathbf{S}_w|\bar{\mathbf{M}})} - \frac{1}{2}\right)^2 + \left(\beta \log \frac{\pi_\phi(\mathbf{S}_l|\bar{\mathbf{M}})}{\pi_{\text{ref}}(\mathbf{S}_l|\bar{\mathbf{M}})} + \frac{1}{2}\right)^2. \quad (5)$$

Sampling



Algorithm 1 ProteinWeaver Model Inference

```
Require: domain module $\theta$, assembly module $\phi$, residue numbers $L$, diffusion steps $N_{\text{steps}}$, domain numbers $m$, step interval $\zeta$, stop time $t_0$
    # division of domains
    $[D_1, D_2, ..., D_m] \sim \text{partition}([1, 2, ..., L], m)$
    # domain backbones generation
    for $i \in [1, 2, ..., m]$ do
        $\bar{\mathbf{S}}_{D_i} = f_\theta(\text{length}(D_i))$
    end for
    # splicing distance maps
    $\bar{\mathbf{M}} = \text{SDM}(\bar{\mathbf{S}}_{D_1}, \bar{\mathbf{S}}_{D_2}, ..., \bar{\mathbf{S}}_{D_m})$
```



```
    # protein backbone generation
    $\gamma = (1 - t_0)/N_{\text{steps}}$
    for $i \in [1, 2, ..., L]$ do
        $x_i^{(1)} \sim \mathcal{N}(0, \text{Id}_3), r_i^{(1)} \sim \mathcal{N}(0, \text{Id})$
        $\mathbf{T}_i^{(1)} = (x_i^{(1)}, r_i^{(1)})$
    end for
    for $t = 1, 1 - \zeta, 1 - 2\zeta, ..., t_0$ do
        $\hat{\mathbf{T}}^{(0)} = g_\phi(\mathbf{T}^{(t)}, t, \bar{\mathbf{M}})$
        $\{(s_n^r, s_n^x)\}_{n=1}^L = \nabla_{\mathbf{T}^{(t)}} \log p_{t|0}(\mathbf{T}^{(t)} | \hat{\mathbf{T}}^{(0)})$
        $\mathbf{T}^{(t-\zeta)} = \text{SDE}_{(\text{SE3})}(\mathbf{T}^{(t)}, \{(s_n^r, s_n^x)\}_{n=1}^L)$
    end for
    # calculate the coordinates
    $\mathbf{S} = \text{CALC\_COORDINATE}(\mathbf{T}^{(t_0)})$
    return $\mathbf{S}$
```

Experiments

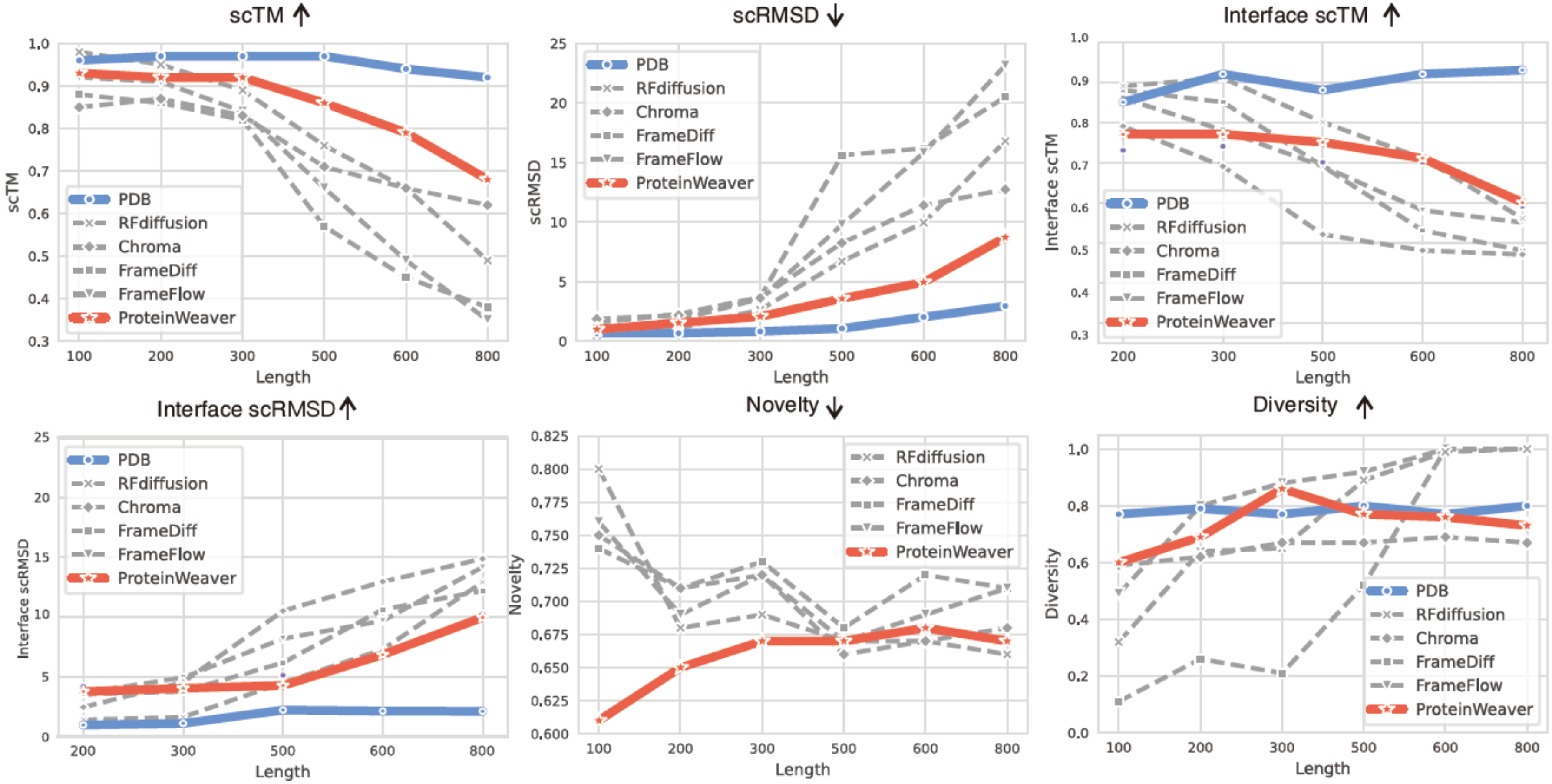


Figure 4: ProteinWeaver shows strong capacity in designing novel and high-quality backbones with significant improvement, particularly in long-chain structures.

Ablation Study

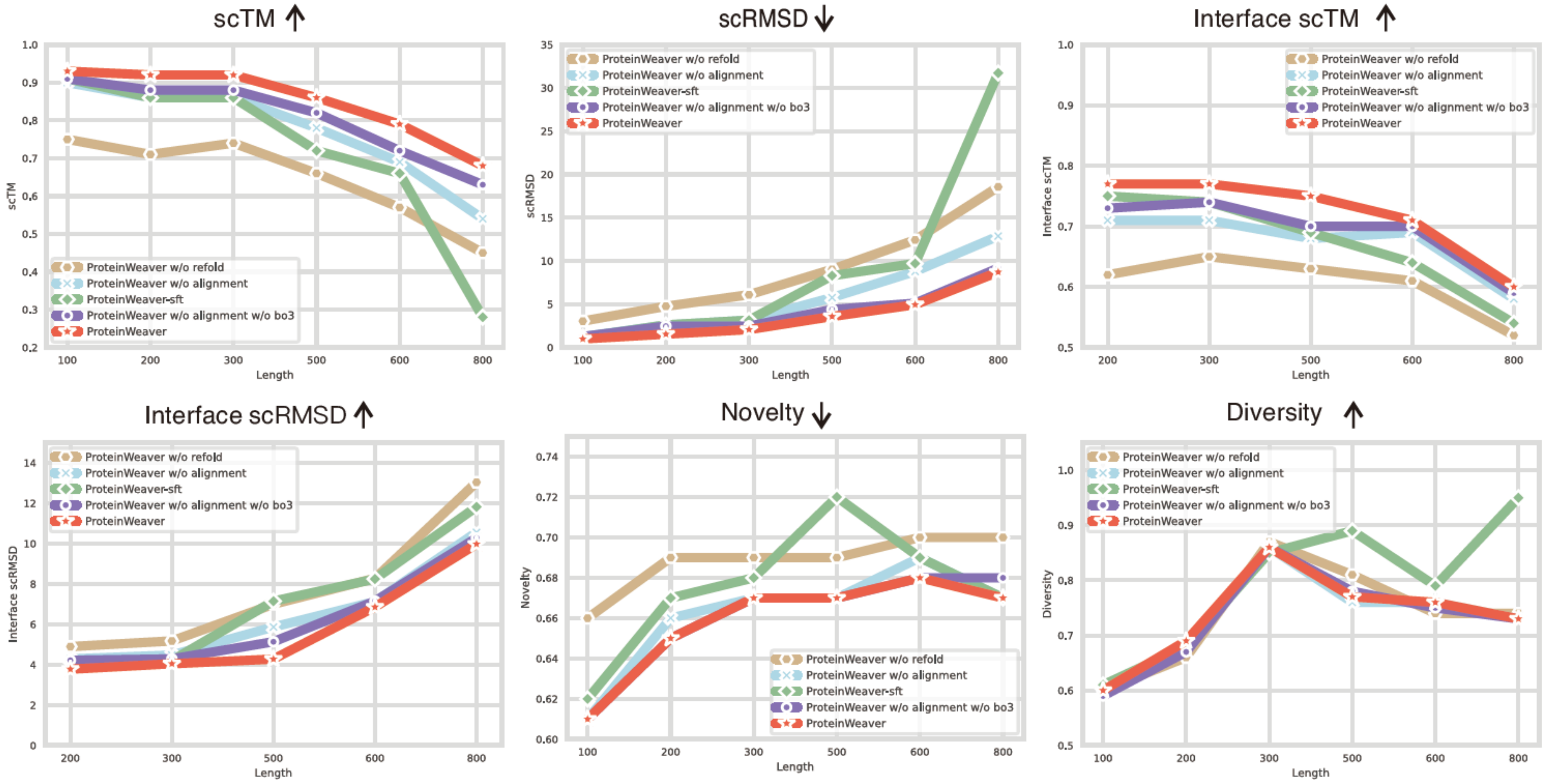


Figure 7: Ablation study on backbone design. “bo3” is abbreviation for best of 3.

Thank you!
I'd appreciate any criticisms and corrections