Protein Structure Prediction using Particle Belief Propagation

Roshan Rao, Jason Pacheco, and Erik Sudderth

Brown University, Department of Computer Science

Goal

Given an amino acid sequence and an electron density map, want to predict the location of every atom in a protein. We frame this as a problem of MAP inference for a likelihood function f and maximize f with Diverse Particle Max Product, an algorithm developed by Pacheco et al. [1].

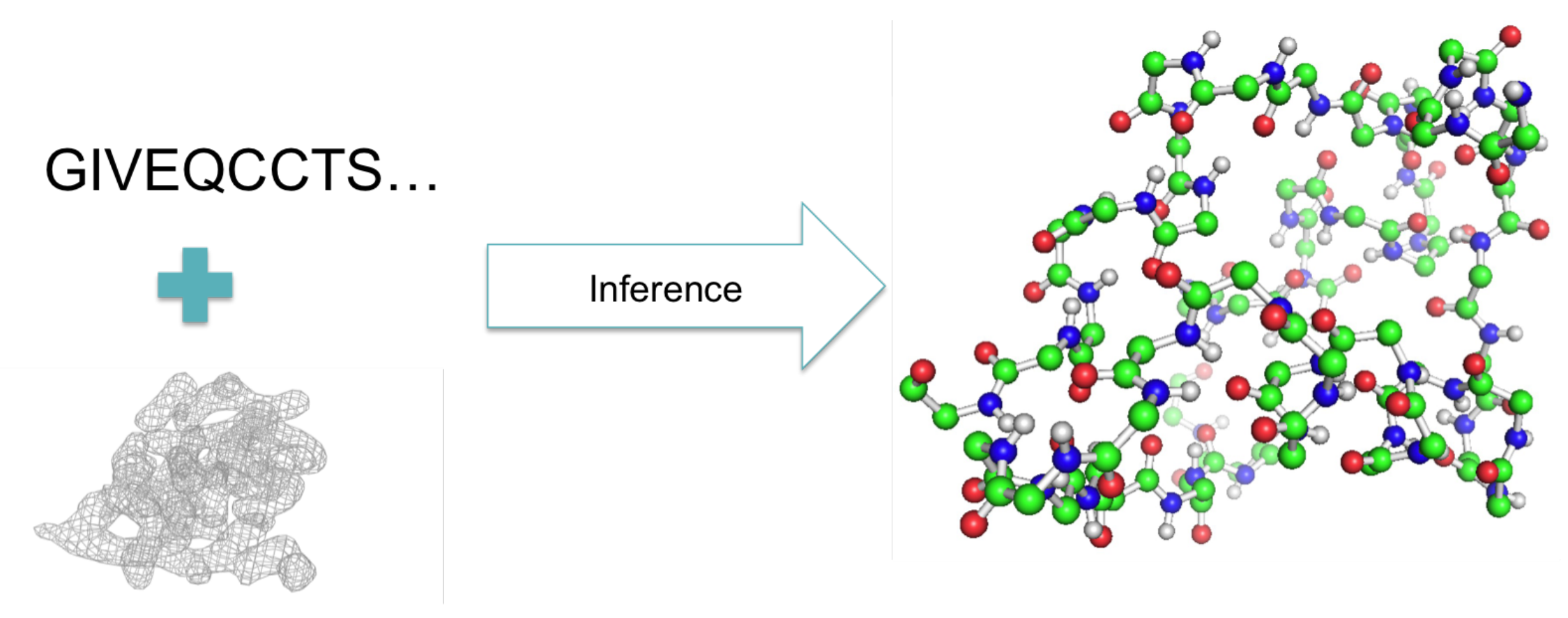


Figure: Density guided protein structure inference problem.

Background

A Markov Random Field (MRF) is a method of representing a function. Suppose f is a function of four discrete random variables that factors like so,

$$f(x_1, x_2, x_3, x_4) = \psi_{12}(x_1, x_2)\psi_{23}(x_2, x_3)\psi_{24}(x_2, x_4)$$

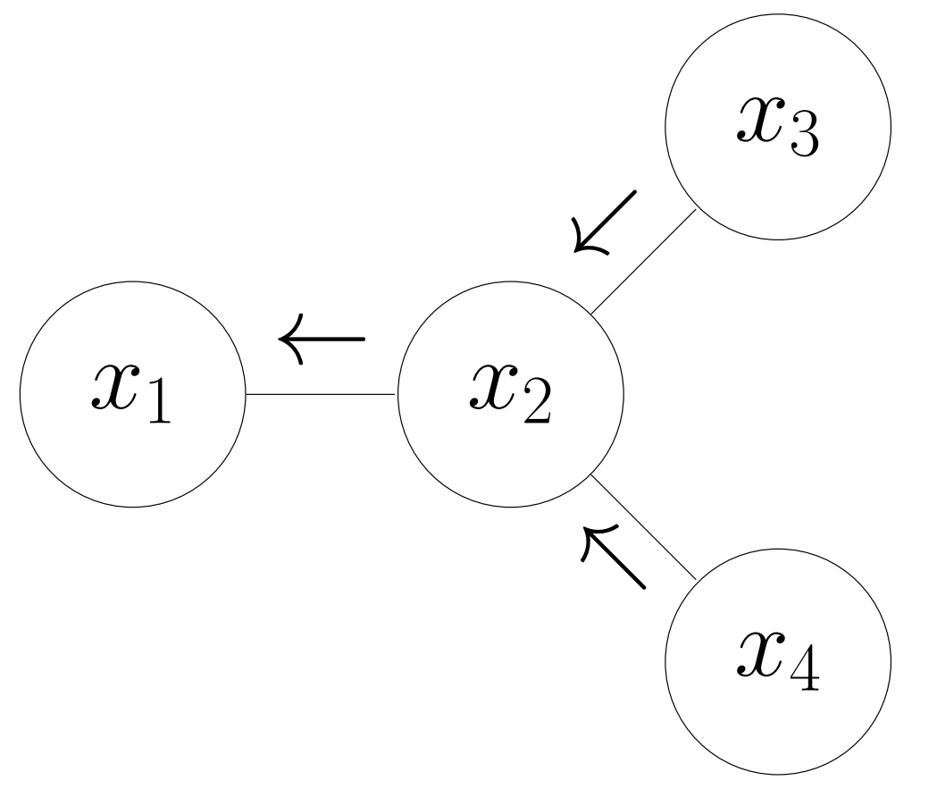


Figure: Graphical representation of f. There is one node for every variable and one edge for every function of two variables. Arrows represent 'direction' of maximization.

Max-product BP allows $O(M^2)$ maximization, where M is number of states.

- Exact if graph is a tree
- Good approximation if graph is cyclic but sparse
- Only works for discrete RVs

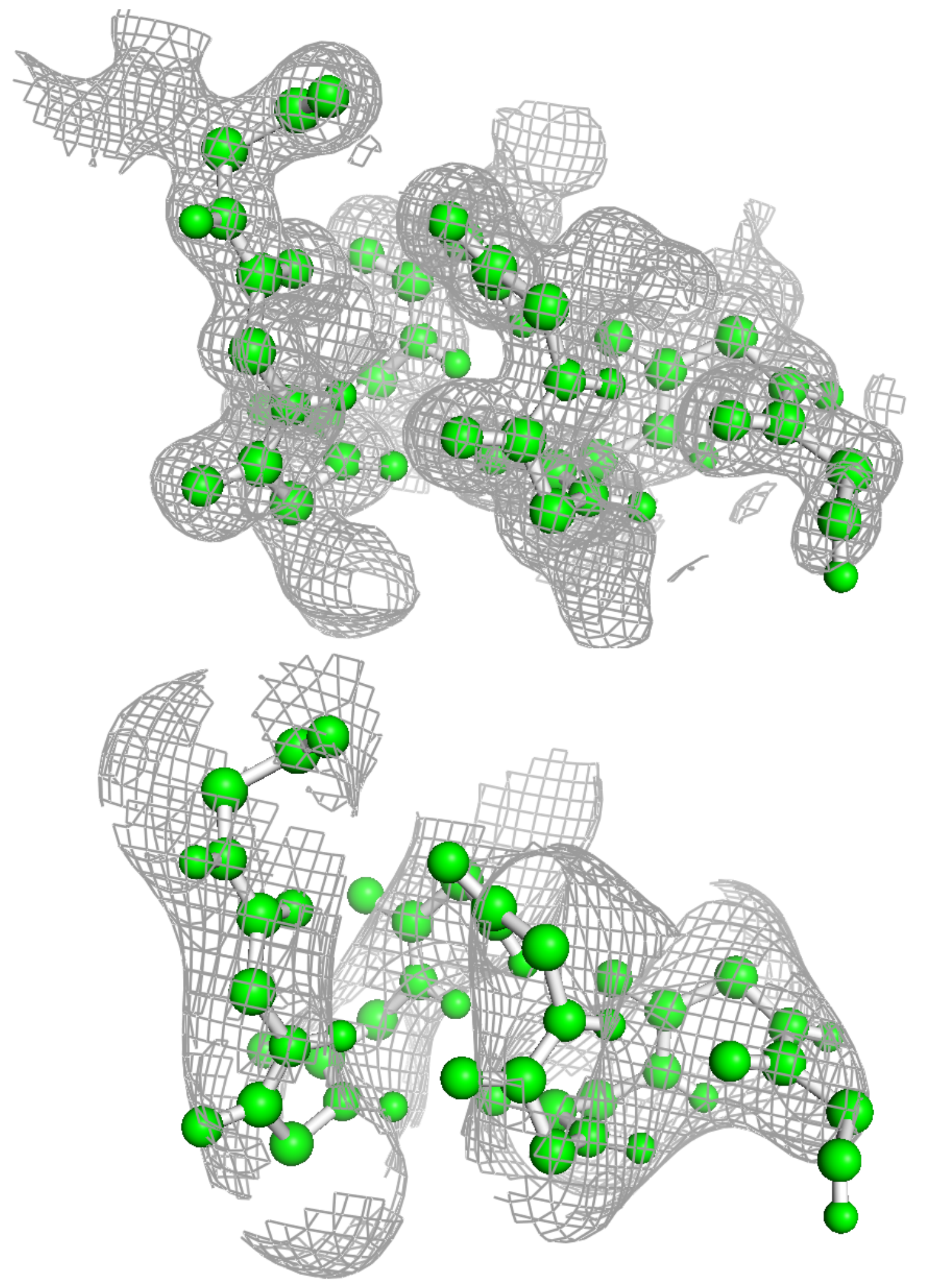
Particle Belief Propagation is the extension into the continuous domain. Iteratively optimize a finite discrete subset $\mathbb{X}$ of continuous space $\mathcal{X}$:

$$\max_{x \in \mathbb{X}} f(x) \le \max_{x \in \mathcal{X}} f(x)$$

Later, we exploit the fact that we are not optimizing over the full domain to drastically improve inference.

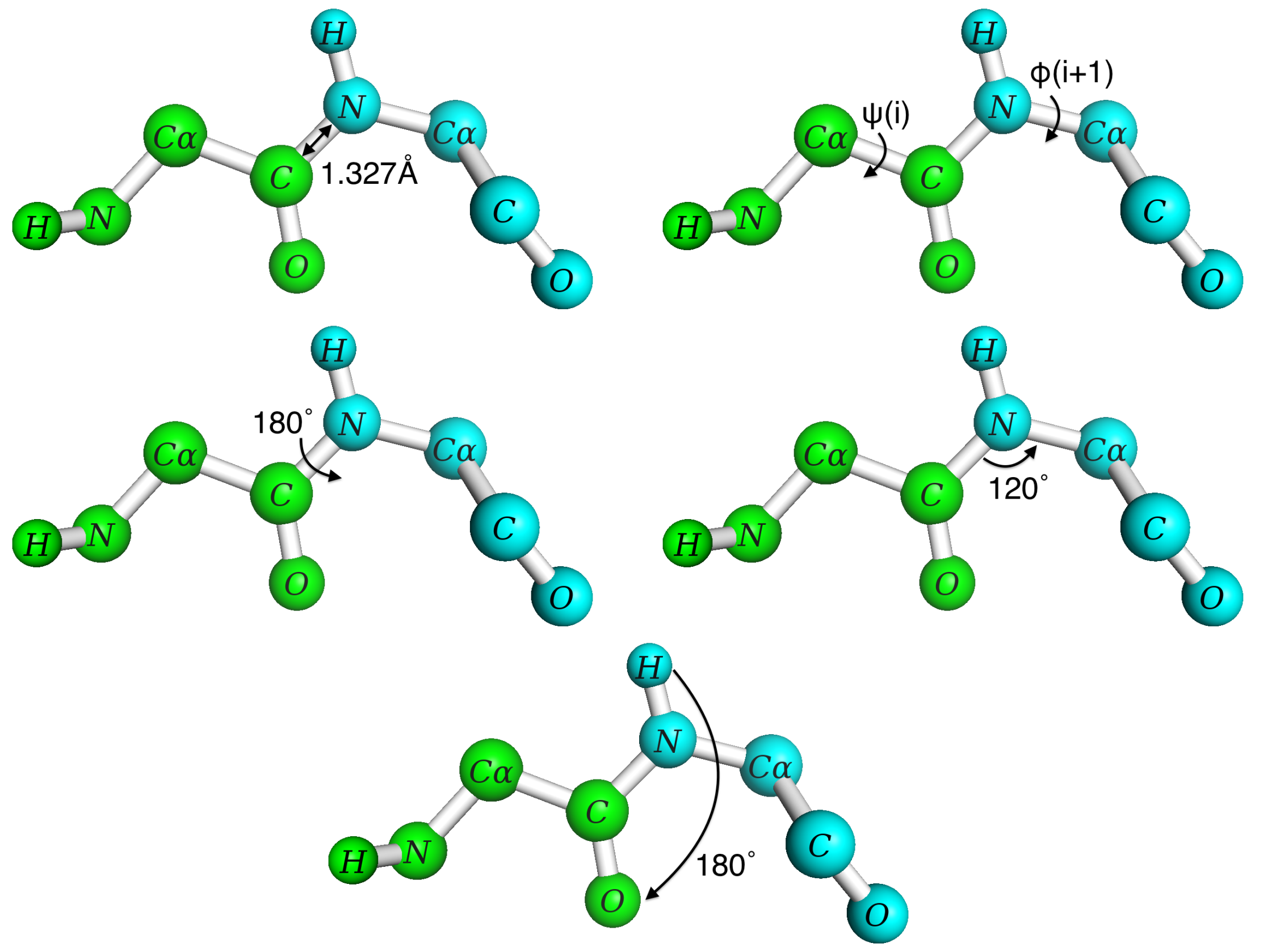
Electron Density Potential

Unary potential. Based on input density map.



Covalent Bonding Potentials

Constrain geometry between adjacent amino acids.



Van der Waals Potential

Repulsive force. Discourages atoms from clashing (occupying the same space or being too near each other) [2].

Likelihood Function

$$f(x) = \sum_{i=1}^{N} \psi_{\text{ElecDens}}(x_i) + \sum_{i=1}^{N-1} \psi_{\text{Covalent}}(x_i, x_{i+1}) + \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \psi_{\text{VDW}}(x_i, x_j)$$

Problem: Complete Graph

Max-Product BP requires a sparse underlying graph. The Van der Waals potential can apply between any two amino acids, so the underlying graph structure is complete. Inference will be slow and inaccurate [3]. Solution: Take advantage of optimization over $\mathbb{X}$ instead of $\mathcal{X}$. Estimate a function $\hat{f}$ over a sparse graph, such that

$$\operatorname*{argmax}_{x \in \mathbb{X}} \hat{f}(x) = \operatorname*{argmax}_{x \in \mathbb{X}} f(x)$$

Intuitively, treat VDW potential as a constraint and iteratively generate constraints.

Graph Structure Estimation

```
function GETESTIMATEDEDGESET($\mathbb{X}$)
    Initialize $\hat{\mathcal{E}} = \{e_{ij} \text{ s.t. } |i-j| = 1\}$, $\hat{G} = (\mathcal{V}, \hat{\mathcal{E}})$
    Let $x^{\text{MAP}} = \operatorname{argmax}_{x \subset \mathbb{X}} \hat{f}(x)$
    Let $\mathcal{E}_{\text{clashing}} = \{e_{ij} \mid \psi_{\text{VDW}}(x_i^{\text{MAP}}, x_j^{\text{MAP}}) < 0\}$
    while $|\mathcal{E}_{\text{clashing}} \cap \hat{\mathcal{E}}^c| > 0$ do
        $\hat{\mathcal{E}} = \hat{\mathcal{E}} \cup \mathcal{E}_{\text{clashing}}$
        $x^{\text{MAP}} = \operatorname{argmax}_{x \subset \mathbb{X}} \hat{f}(x)$
        $\mathcal{E}_{\text{clashing}} = \{e_{ij} \mid \psi_{\text{VDW}}(x_i^{\text{MAP}}, x_j^{\text{MAP}}) < 0\}$
    return $\hat{\mathcal{E}}$
```

Function is upper bound, equal at x^{MAP}, so

$$f(x^{\text{MAP}}) = \hat{f}(x^{\text{MAP}}) \ge \hat{f}(x) \ge f(x) \; \forall x \in \mathbb{X}$$

Results

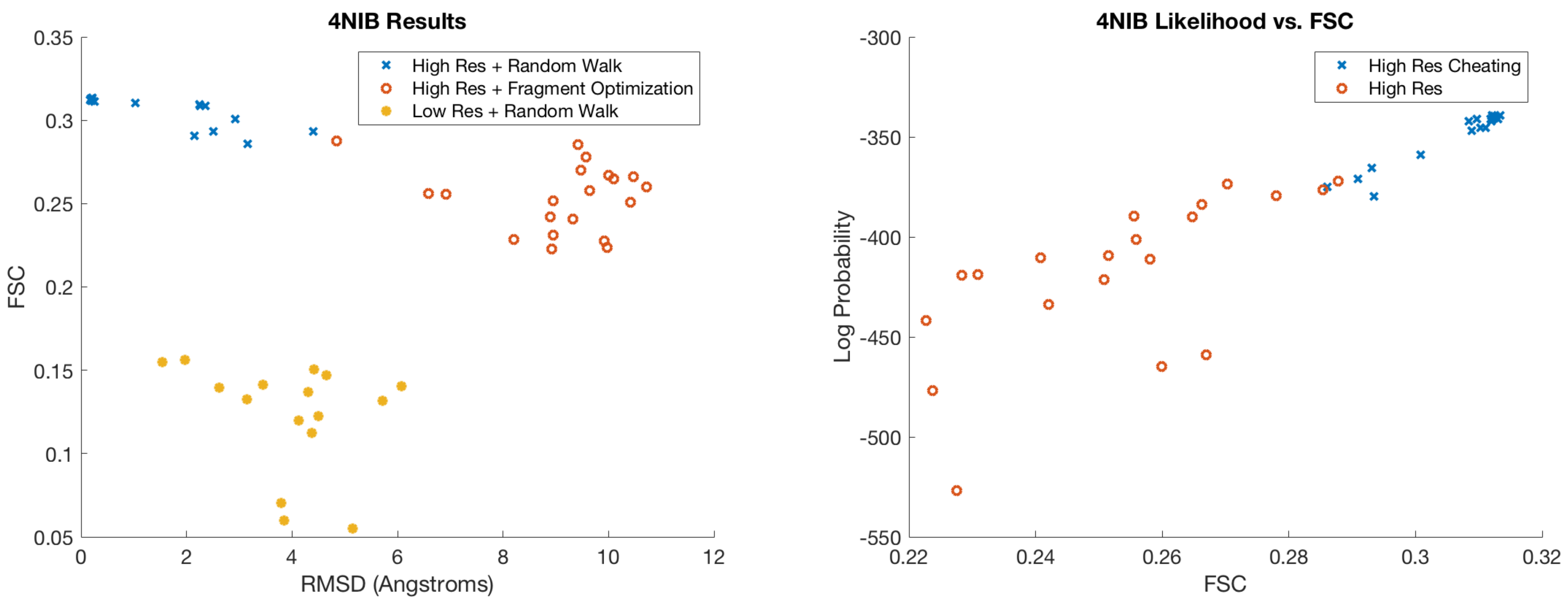


Figure: Experimental setup: 1 protein (4NIB), two resolutions (high and low), two initializations (strong and weak).

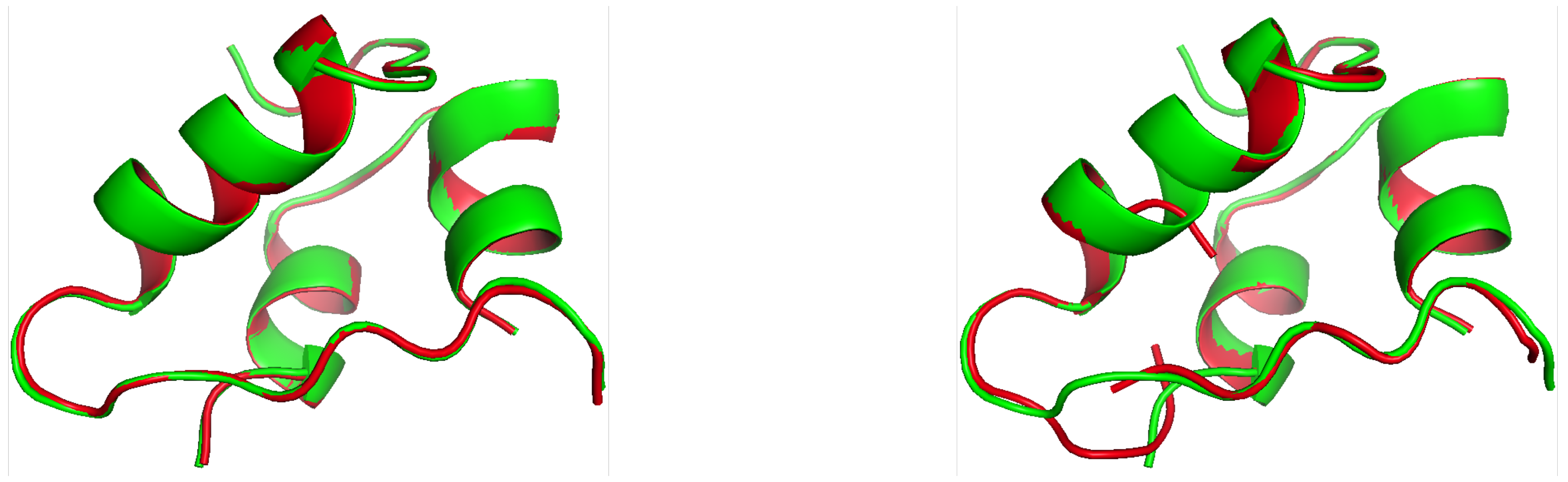


Figure: Visualization of estimated structure. Strong initialization (left) versus weak initialization (right).

References: [1] Pacheco, Sudderth, *ICML*, 2015. [2] Rohl, et al. *Methods in enzymology*, 2004. [3] Ihler, et al. *JMLR*, 2005.