nature communications

清華大學 Tsinghua University

Article

https://doi.org/10.1038/s41467-023-41343-1

Machine learning coarse-grained potentials of protein thermodynamics

Received: 2 June 2023

Accepted: 29 August 2023

Published online: 15 September 2023

Maciej Majewski[1,2,14], Adrià Pérez[1,2,14], Philipp Thölke[1], Stefan Doerr[2], Nicholas E. Charron[3,4,5], Toni Giorgino[6], Brooke E. Husic[7,8,9,10], Cecilia Clementi[3,4,5,11] ✉, Frank Noé[5,7,11,12] ✉ & Gianni De Fabritiis[1,2,13] ✉

Group 6

2026.5.27

Contents

1. Background（齐麟仪）
2. Principles（朱辰旭）
3. Reproduce（谢家昊）
4. Results compare（曾爱晋）

Method： constructing **coarse-grained molecular potentials** based on **artificial neural networks** and grounded in **statistical mechanics**

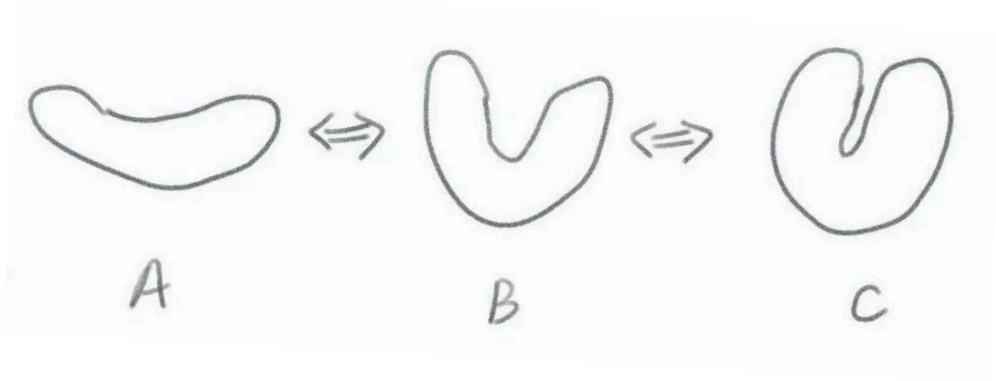
Dataset： all-atom molecular dynamics simulations of **9 ms** for **12 proteins**

Results：

- accelerating the dynamics by **more than three orders of magnitude**
- preserving the thermodynamics of the systems
- Universality： a single coarse-grained potential can integrate all 12 proteins

Background： 3 aspects of protein dynamics

- Conformational state



- Thermodynamics： relative probabilities

A: ? % B: ? % C: ? %

- Kinetics： rates of conversion

Background： coarse-graining

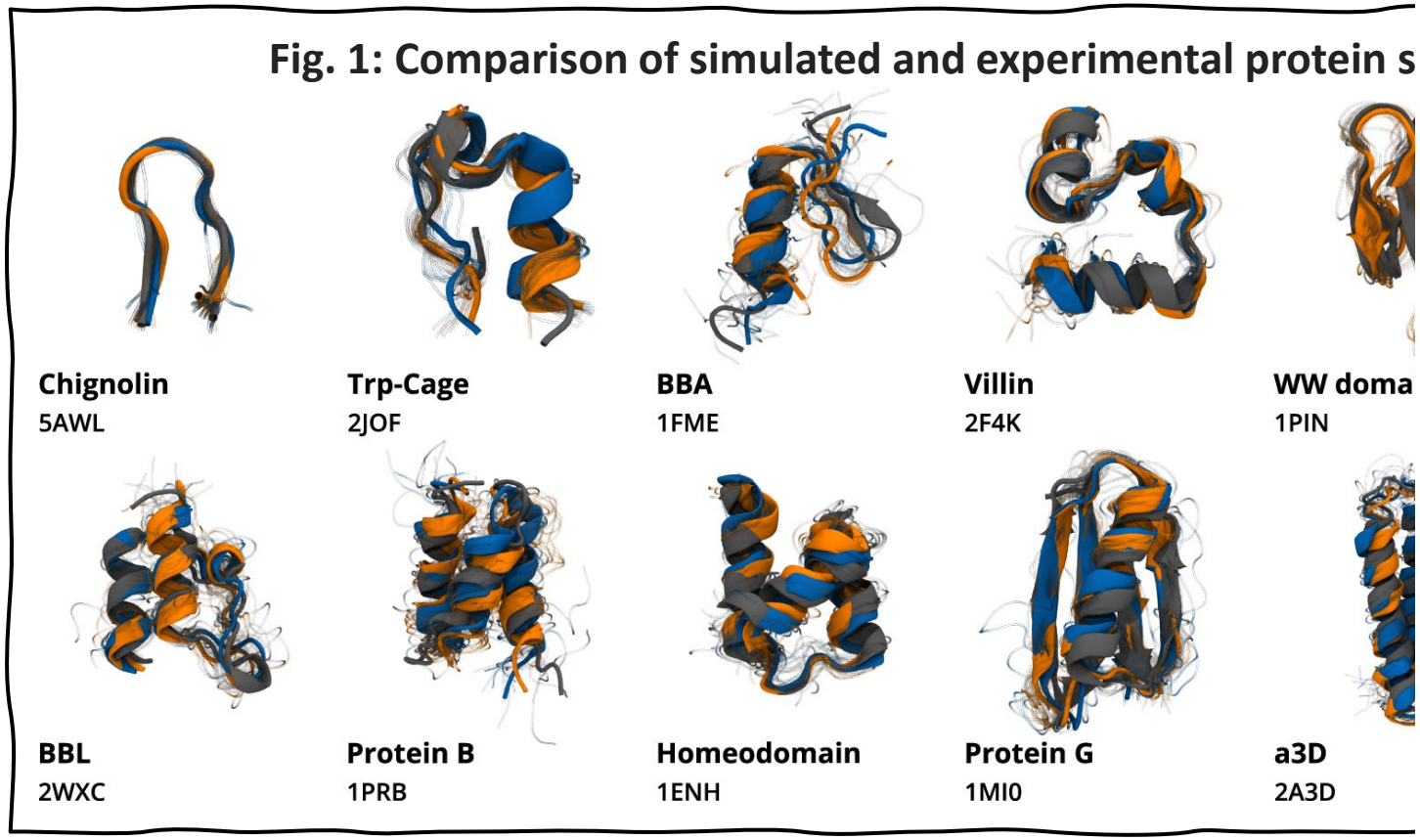
Problem： Describing longer-timescale events

Solutions：

- sampling methods——optimize trajectory
- **coarse-graining (CG)**——simplify system

a CG model
1. selection of the CG resolution (or mapping) √
2. **design of an effective energy function** !

Dataset： twelve fast-folding proteins



Protein	Sequence length (#aa)	Aggregated time (μs)	Min. RMSD (Å)
Chignolin	10	186	0.15
Trp-Cage	20	195	0.45
BBA	28	362	1.13
WW-Domain	34	1362	0.73
Villin	35	234	0.47
NTL9	39	776	0.32
BBL	47	677	1.55
Protein B	47	608	1.19
Homeodomain	54	198	0.56
Protein G	56	2266	0.55
α3D	73	768	1.81
λ-repressor	80	1422	0.82

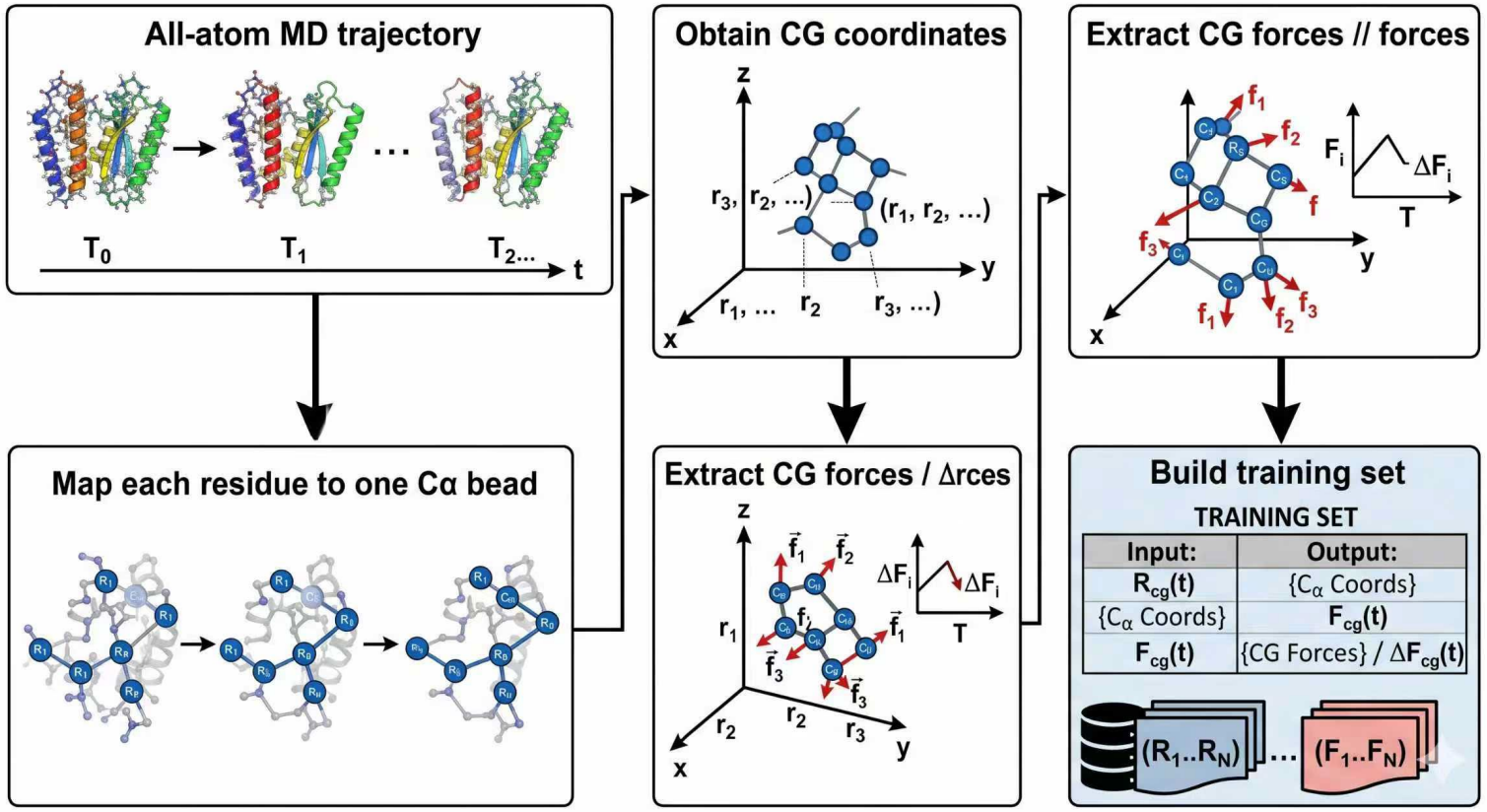
experimental structures

CG simulations of the： protein-specific model / multi-protein model

Contents

Core idea: AI-based coarse-graining

Goal: Learn a coarse-grained potential that preserves protein thermodynamics.



① Each amino acid is represented by one Cα bead

② 21 bead types are used to encode amino acid identity

③ The CG model reduces degrees of freedom and accelerates simulation

④ Prior energy terms constrain physically unreasonable structures

All-atom protein → Cα beads → CG protein model

Charron et al., 2025, *Navigating protein landscapes with a machine-learned transferable coarse-grained model*, Nature Chemistry

Physical priors: keeping CG proteins reasonable

V_prior

① * bond length

② * angle / dihedral

③ * repulsion

④ * chirality / structural prior

$$V_{\text{bonded}}(r) = k(r - r_0)^2 + V_0$$

$$V_{\text{repulsive}}(r) = 4\epsilon r^{-6} + V_0$$

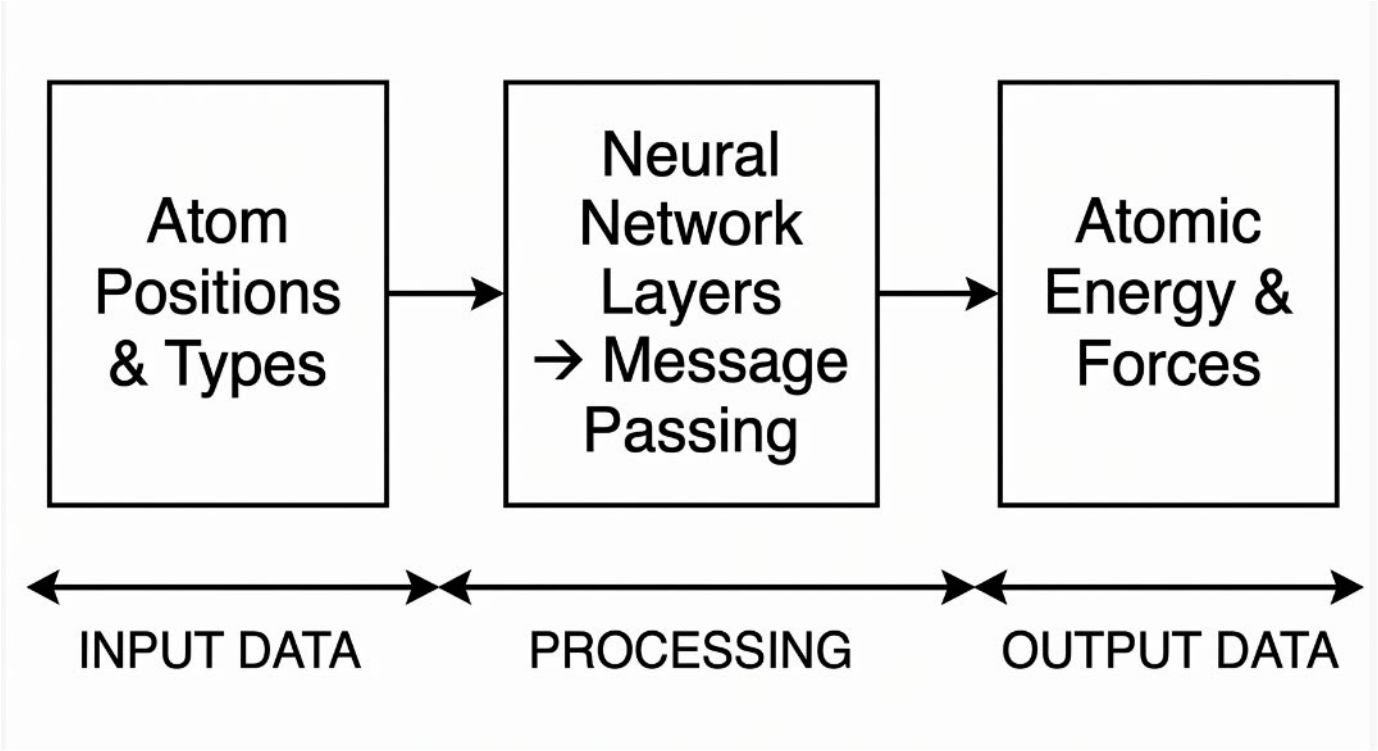
$$V_{\text{dihedral}}(\phi) = \sum_{n=1,2} k_n(1 + \cos(n\phi - \gamma_n))$$

V_CG = V_ prior + V_ NNP

V_ prior = V_ bonded + V_ repulsive + V_ dihedral

Model:
TorchMD-GN, a graph neural network inspired by SchNet and PhysNet

The coarse-grained protein is represented as a molecular graph. Each bead is a node, and spatial distances between neighboring beads define the graph edges.



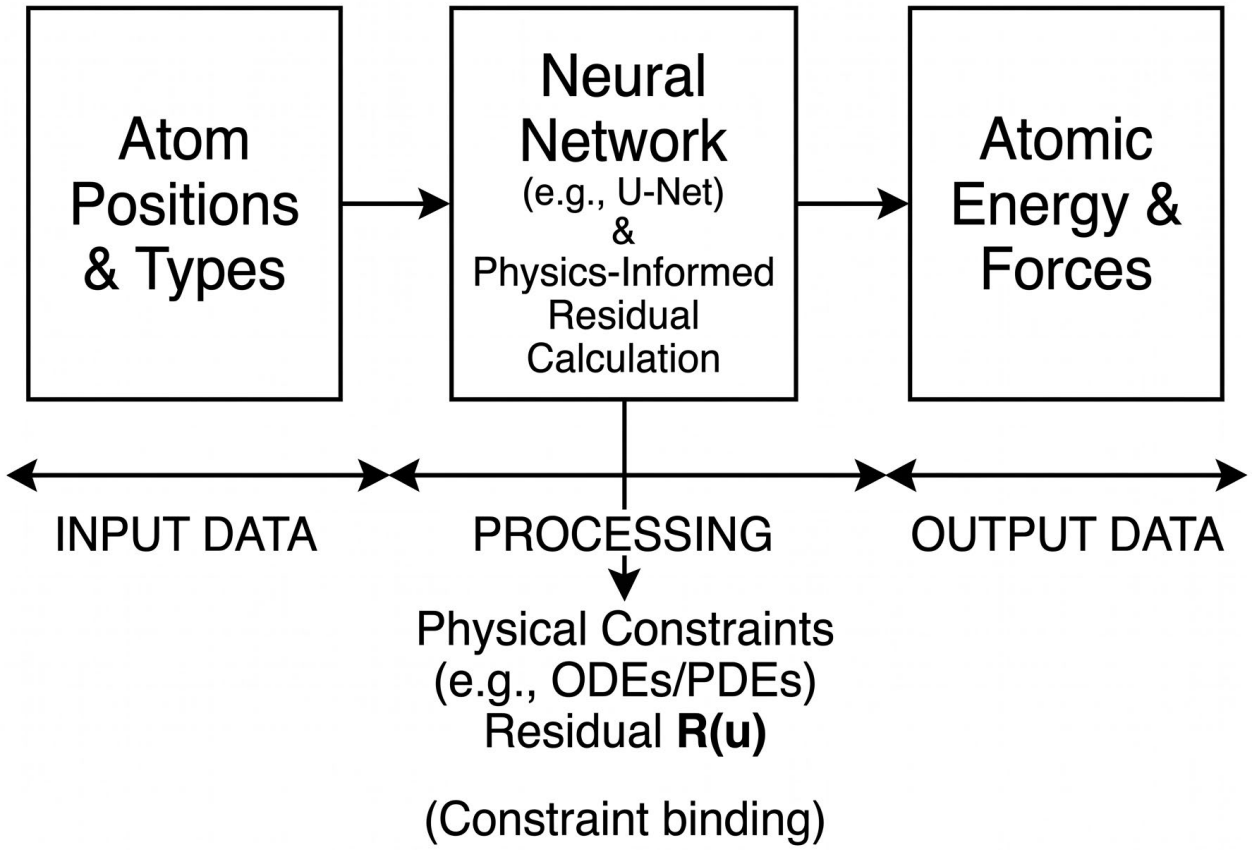
SchNet architecture

Schütt et al., 2017, *SchNet: A continuous-filter convolutional neural network for modeling quantum interactions*, NeurIPS.

Model:
TorchMD-GN, a graph neural network inspired by SchNet and PhysNet

The coarse-grained protein is represented as a molecular graph. Each bead is a node, and spatial distances between neighboring beads define the graph edges.

PhsNet architecture



Neural network potential: how the model learns

Learning Energy and Forces from a Coarse-Grained Protein Graph

Bead type + bead coordinates

↓

TorchMD-GN

↓

CG potential energy V_NN(x)

↓

Forces by gradient: F = -∇V

↓

CG molecular dynamics

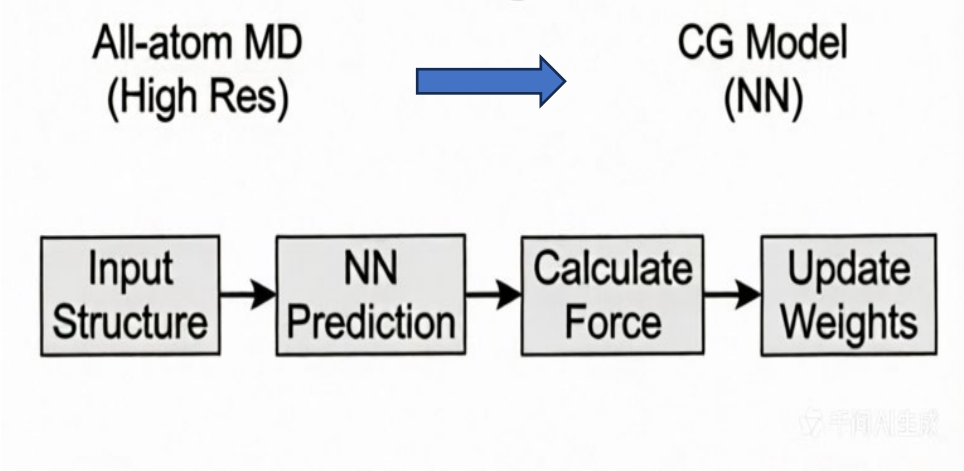
Cα beads as nodes → message passing → energy V → force F

Training strategy: force matching

- Training objective: Match coarse-grained forces to all-atom MD forces.
- Training data: ~9 ms all-atom MD simulations
- Temperature: 350 K
- Force field: CHARMM22*
- The model learns forces mapped from all-atom MD to CG space
- Two model settings were tested:

  protein-specific models

  one multi-protein model



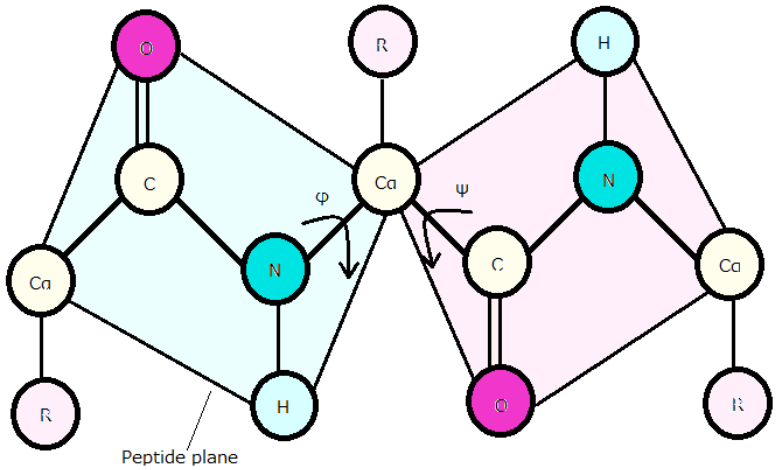
All-atom MD trajectories → map to Cα beads → force matching → trained NNP → CG simulation

Evaluation: from CG trajectories to thermodynamic landscapes

- **RMSD**

$$\text{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2}$$

where δ_i denotes the distance between the same Cα in the reference structure and the structure calculated. N is the number of the Cα atoms



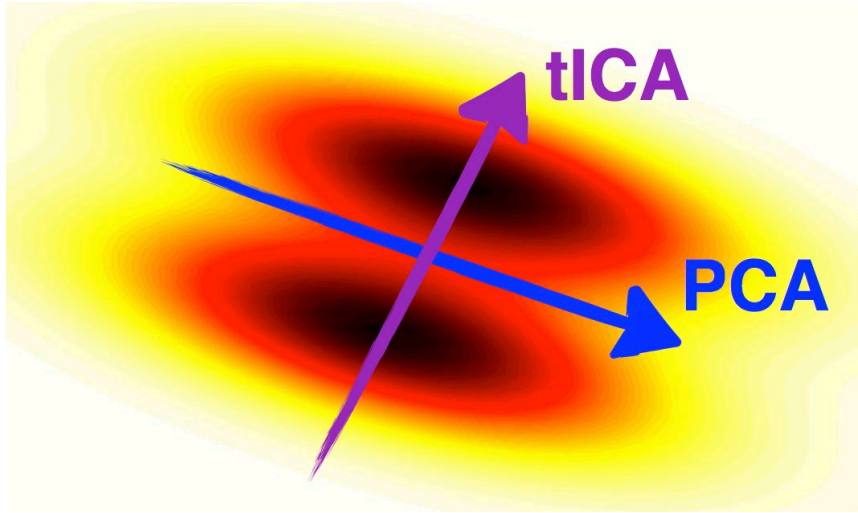
http://cib.cf.ocha.ac.jp/bitool/DIHED2/#:~:text=A%20dihedral%20angle%20of%20a,all%20in%20a%20single%20plane.

- **Markov State Model:**

The transition probability matrix P(τ) , whose elements are given by $P_{ij}(\tau)$ and describe the transition probabilities from state i to state j in a time τ.

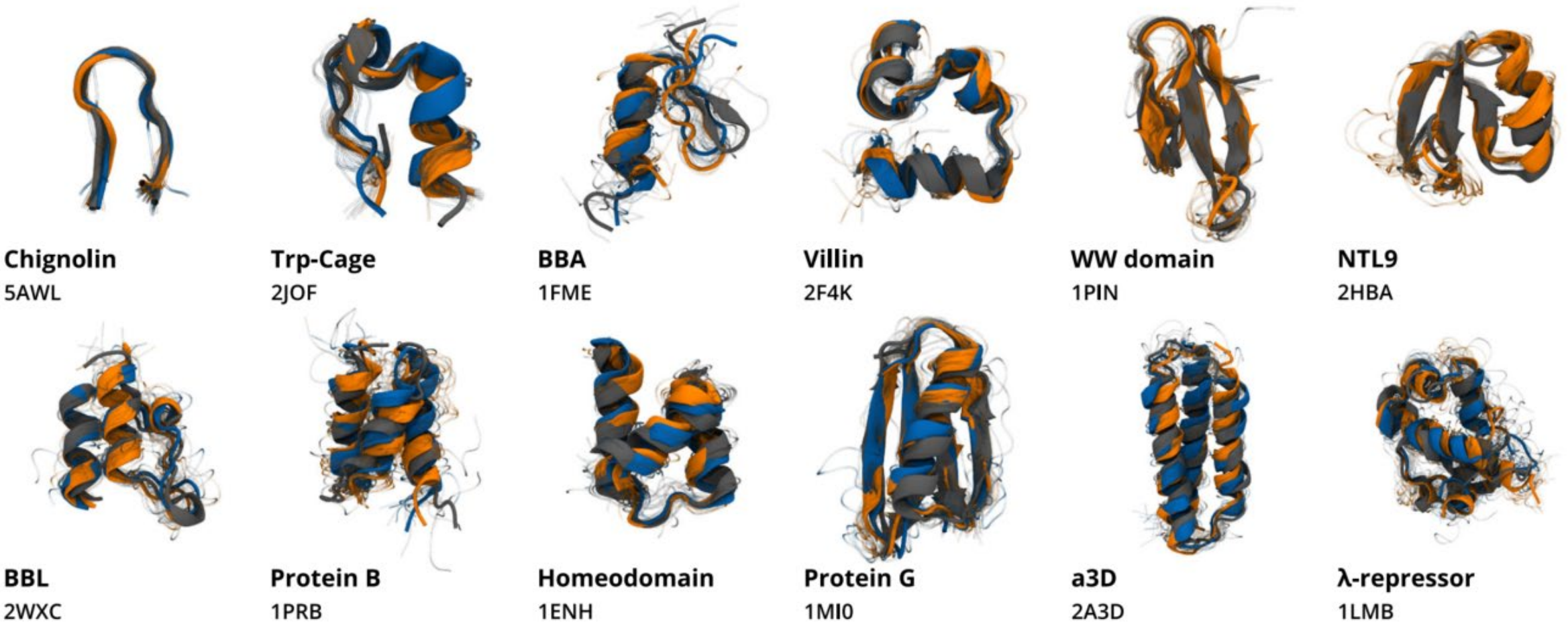
$$\pi_i P_{ij} = \pi_j P_{ji}$$

where π_i denotes the stationary probability of state i. P_{ij} is the transition probability to state j based on state i.



http://msmbuilder.org/3.3.0/tica.html

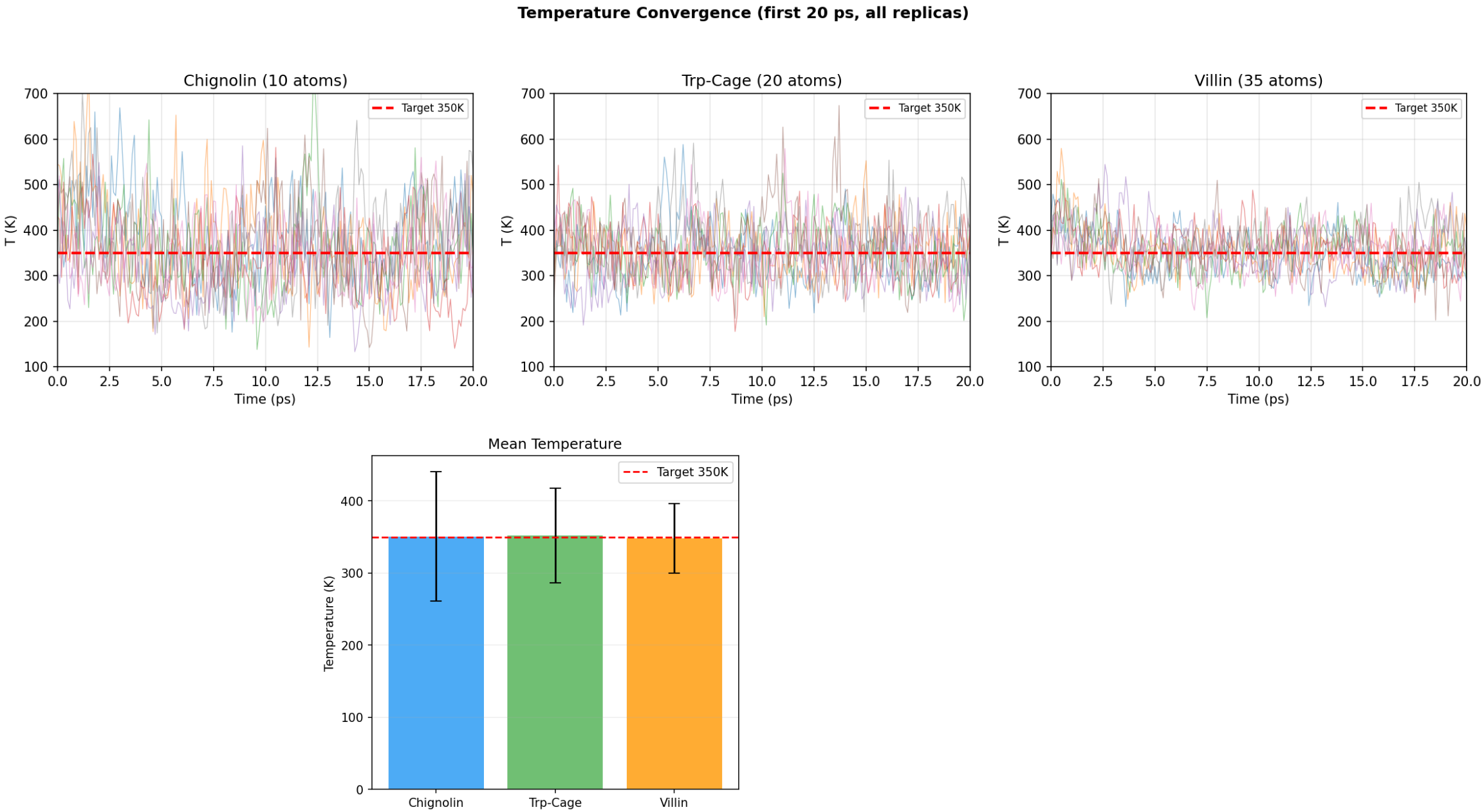
Dataset： The training dataset contains approximately 9 ms of unbiased all-atom MD simulations from 12 fast-folding proteins with diverse secondary structures.



Contents

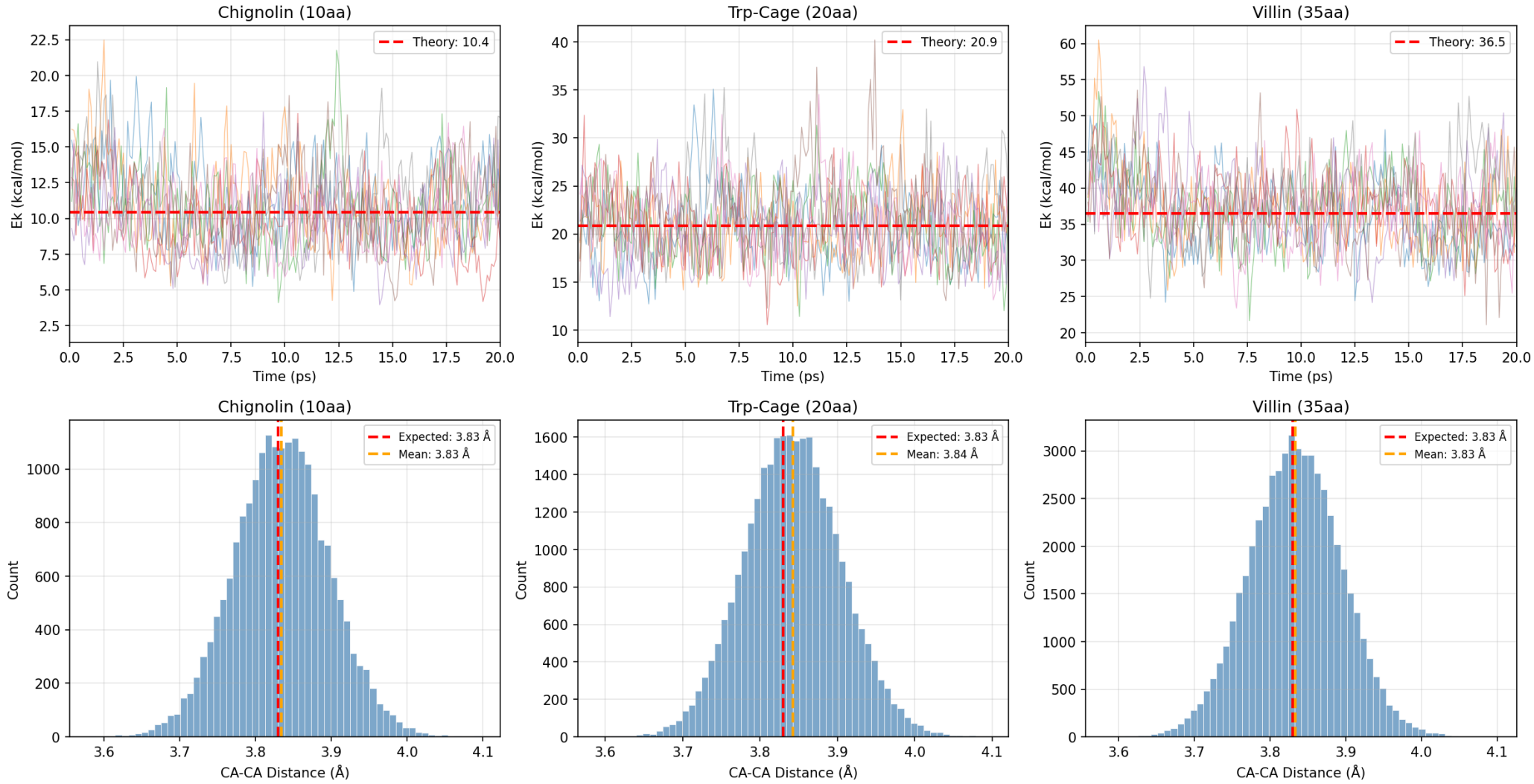
CG-MD Protein Thermodynamics Pipeline

Simulation Setup & Quality Validation: Precise Thermodynamic Stability

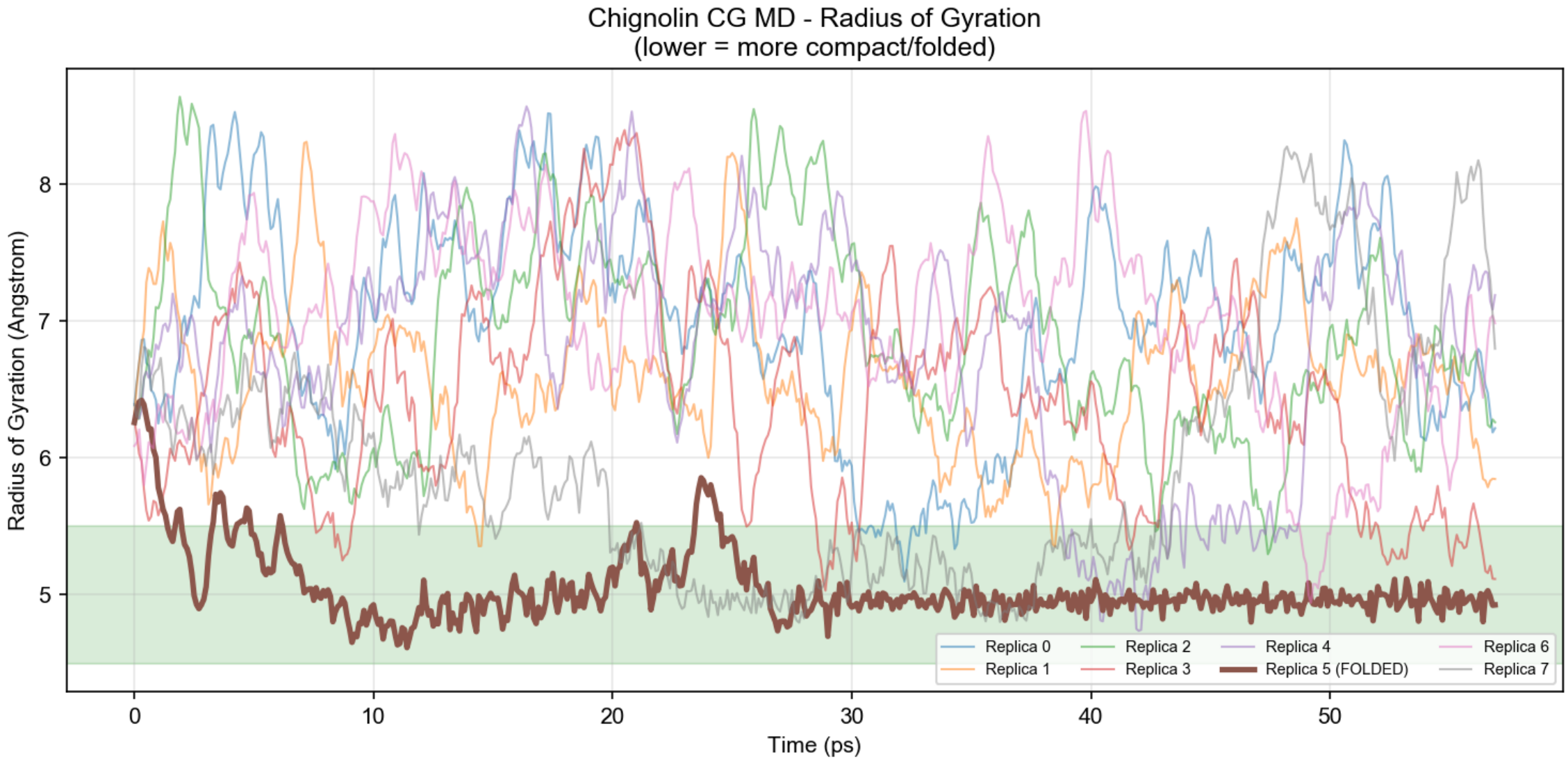


Strict Geometric & Kinetic Preservation

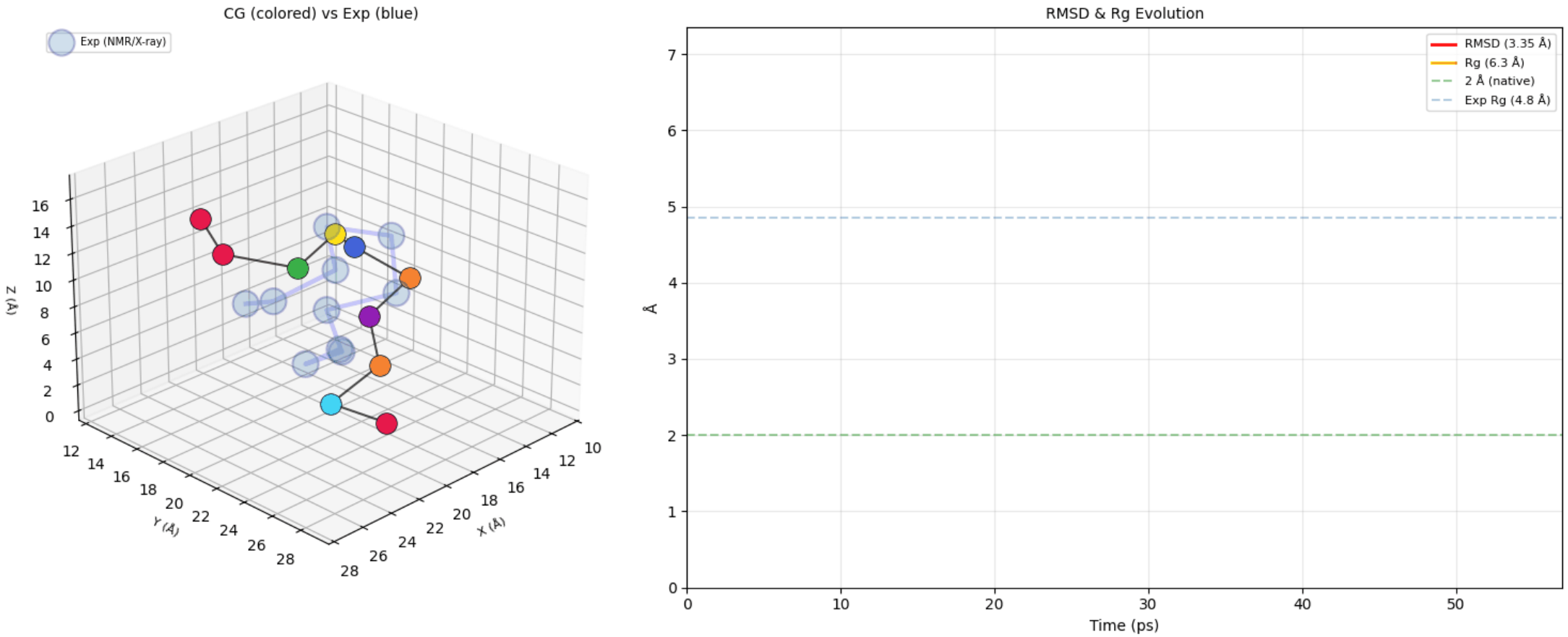
Simulation Quality Validation



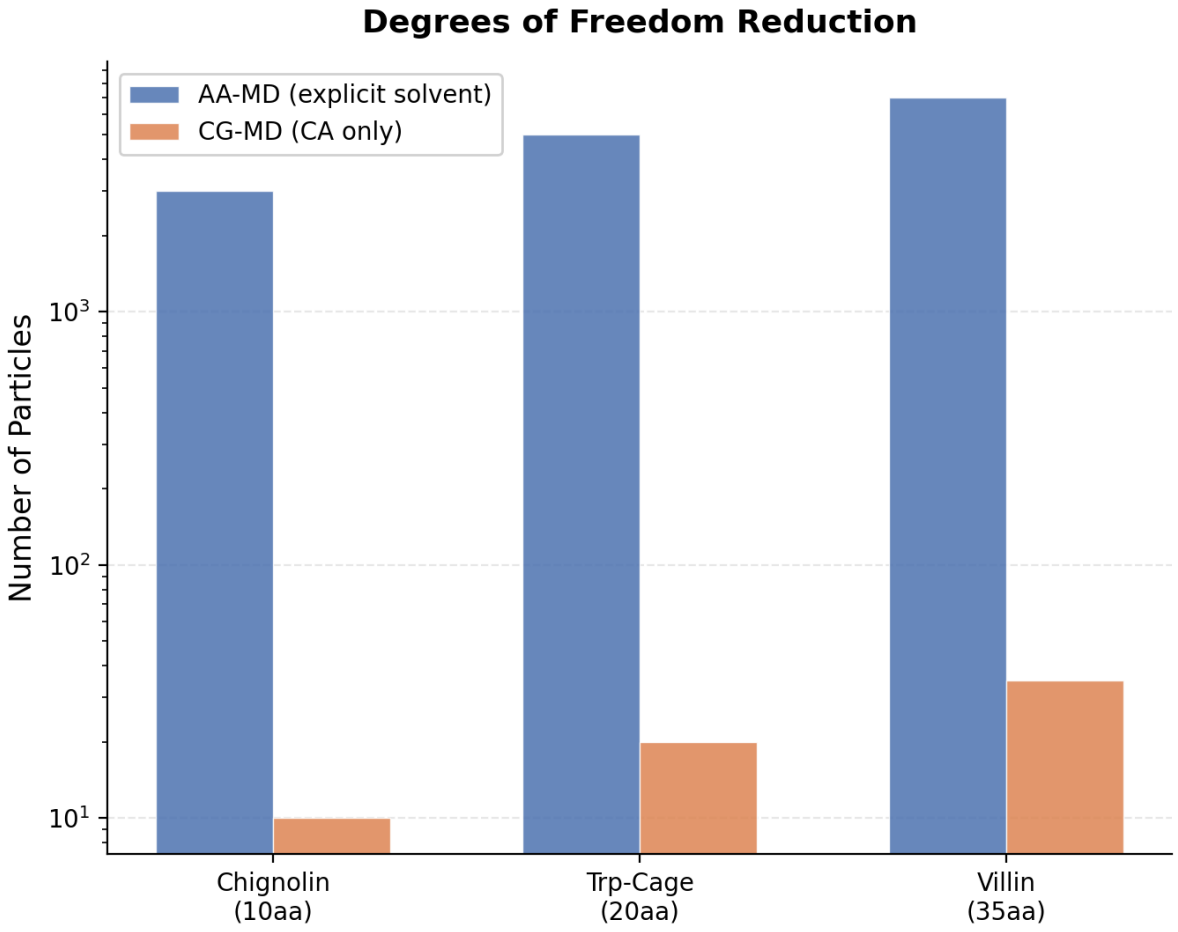
Chignolin Spontaneous Folding: Collapse to Native State

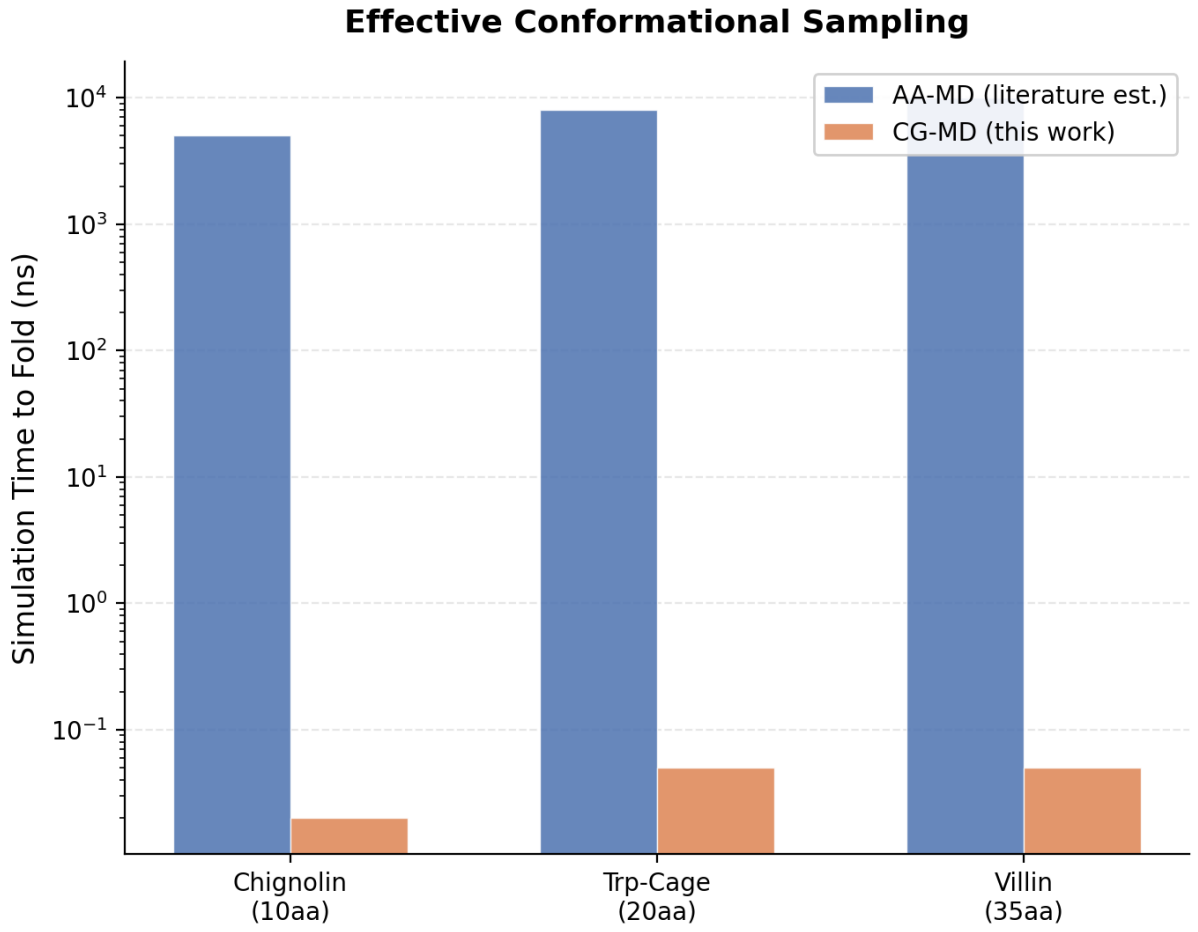


Chignolin Spontaneous Folding: Sub-Angstrom Folding Accuracy

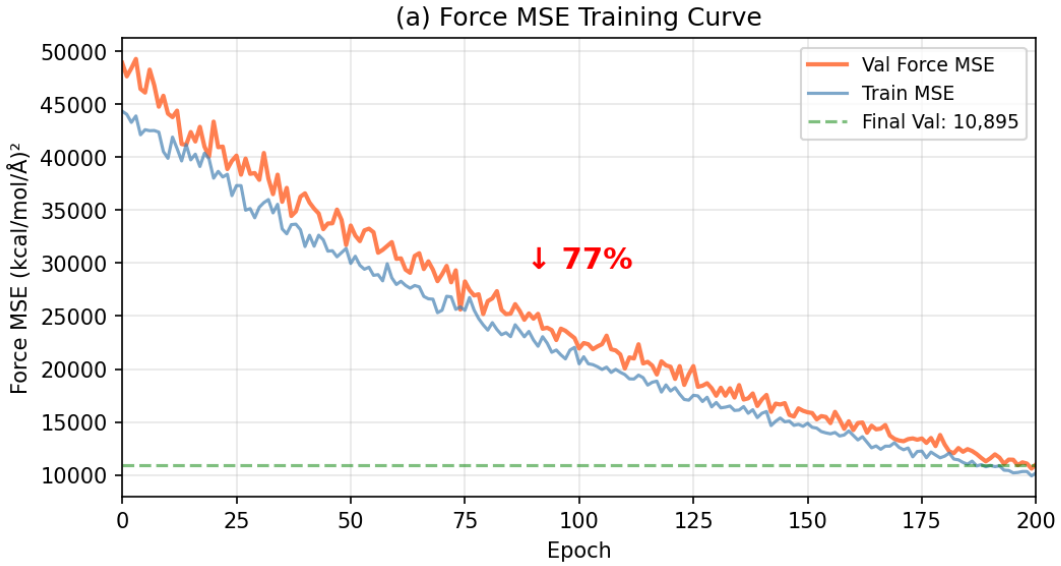


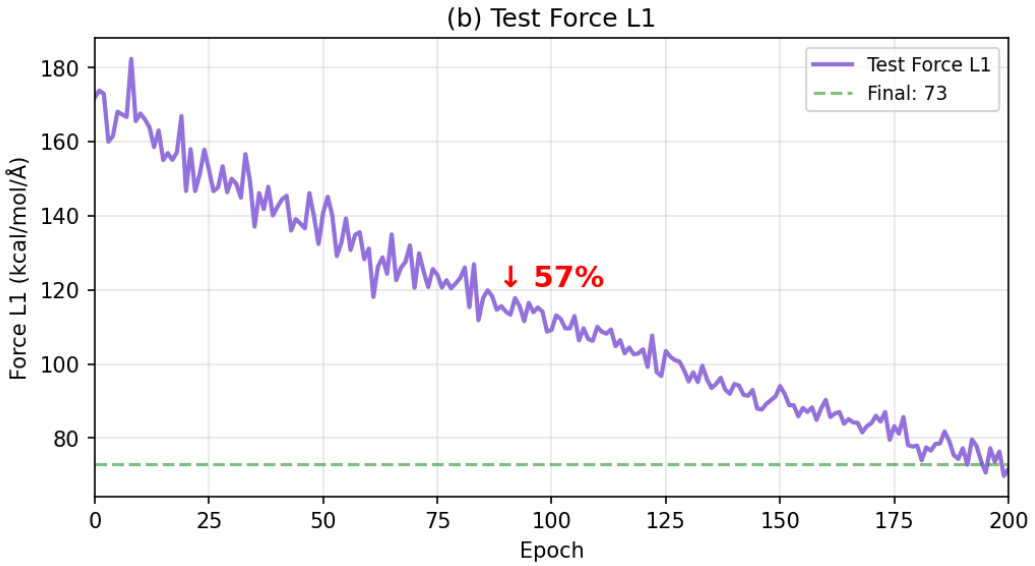
CG-MD vs AA-MD: ~10^5x Acceleration in Conformational Sampling



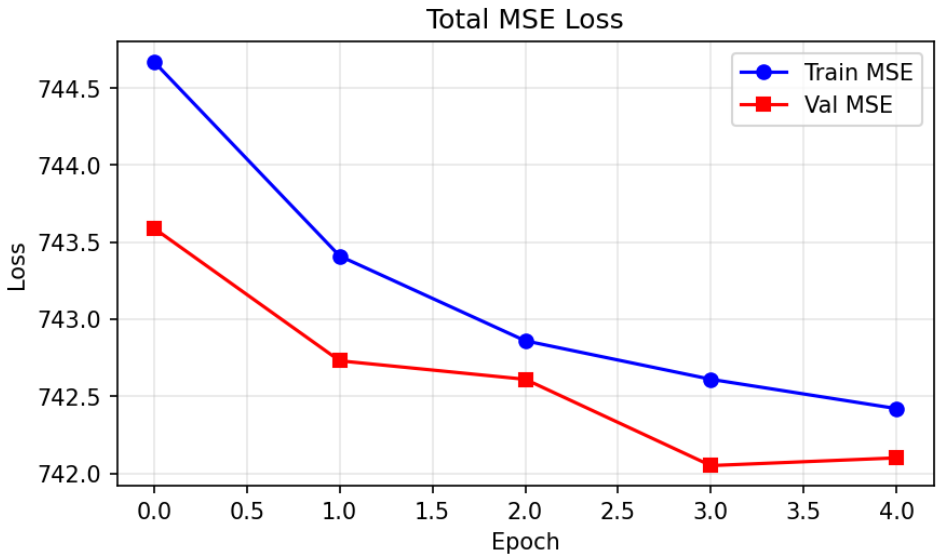


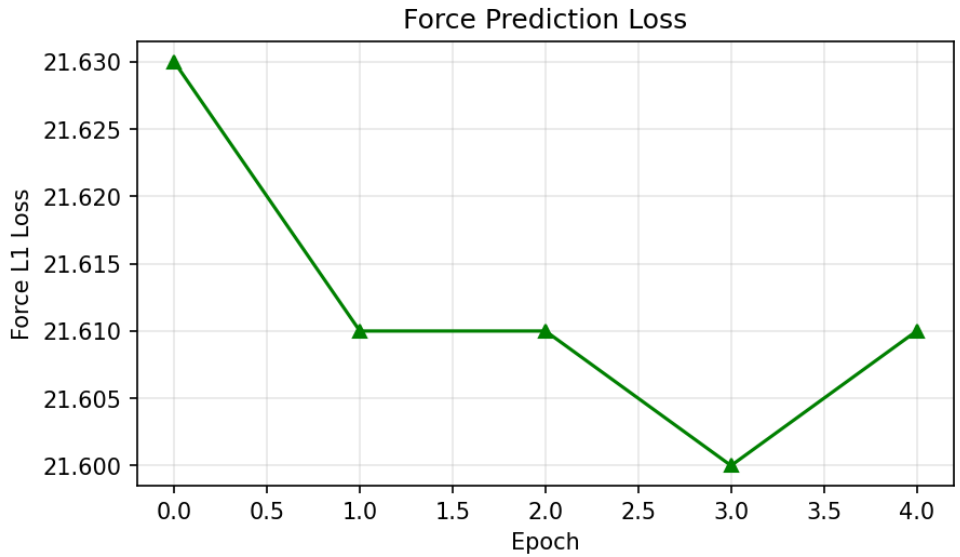
Self-trained 1DUJ Model: Loss Decreases on Limited Data





CG-NNP Training Progress (Chignolin, 200K frames)

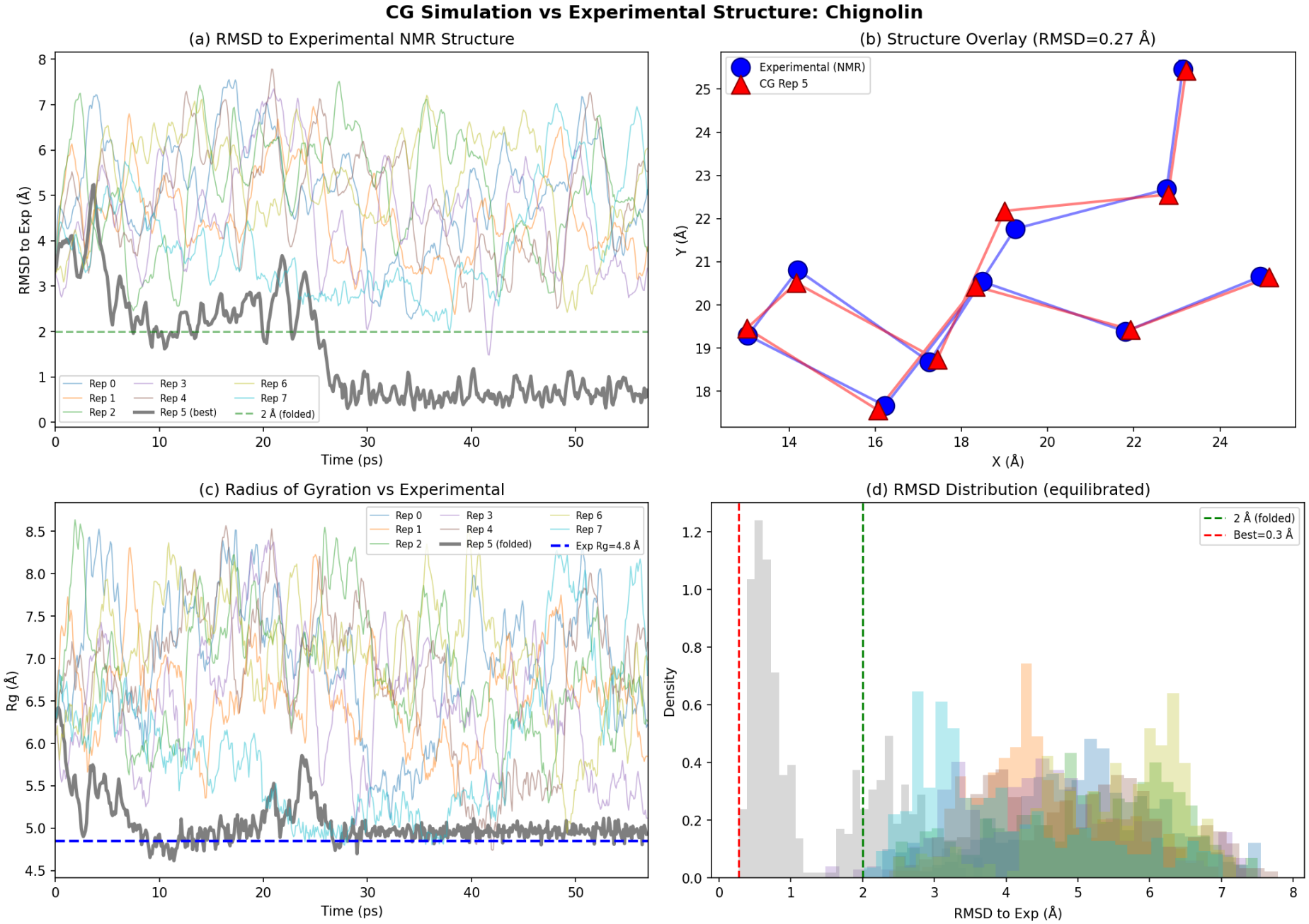




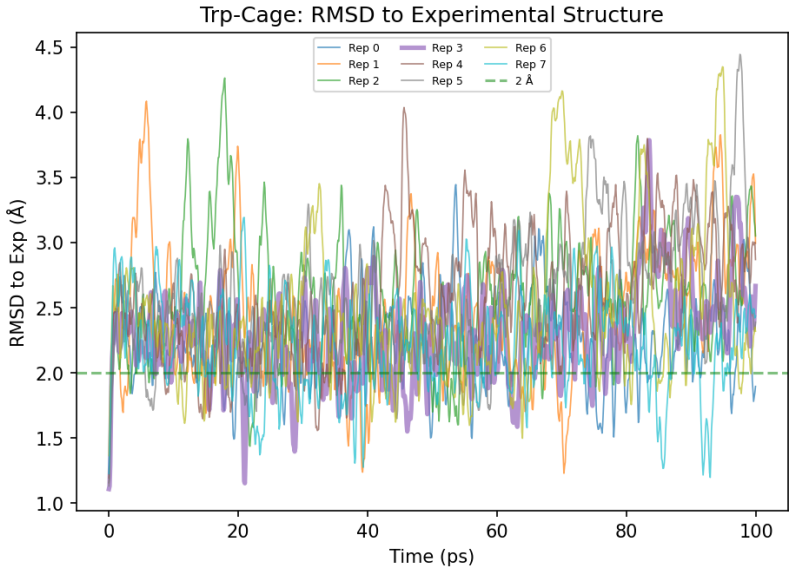
Contents

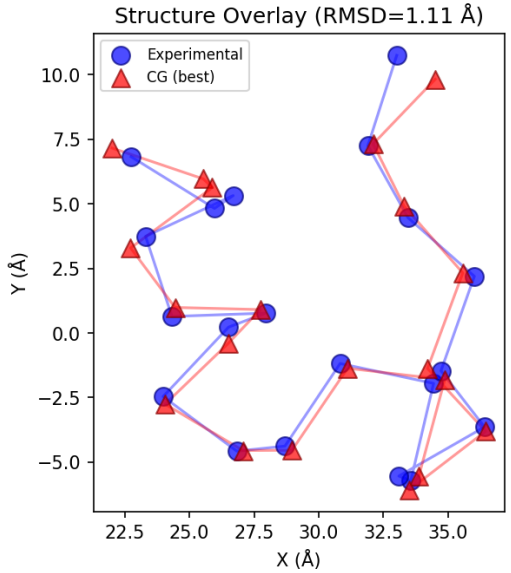
1. Background（齐麟仪）
2. Principles（朱辰旭）
3. Reproduce（谢家昊）
4. Results compare（曾爱晋）

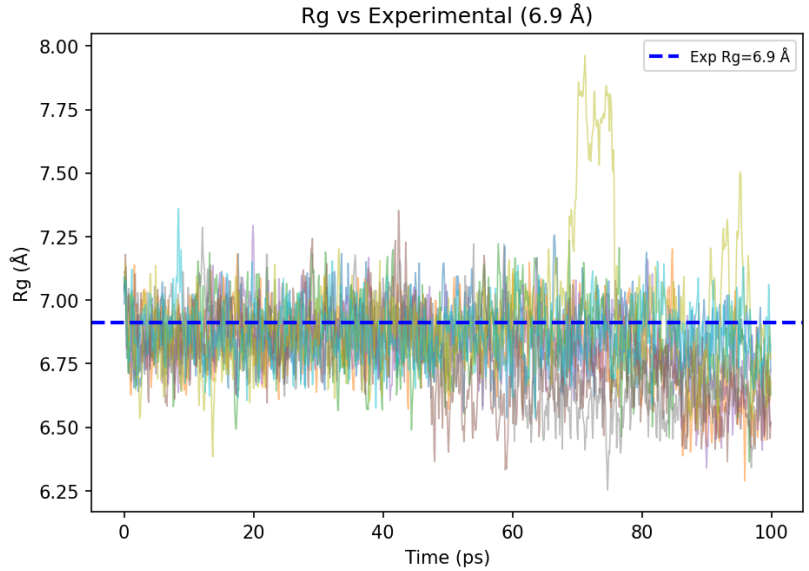
CG Simulation Recovers Native-like Structures



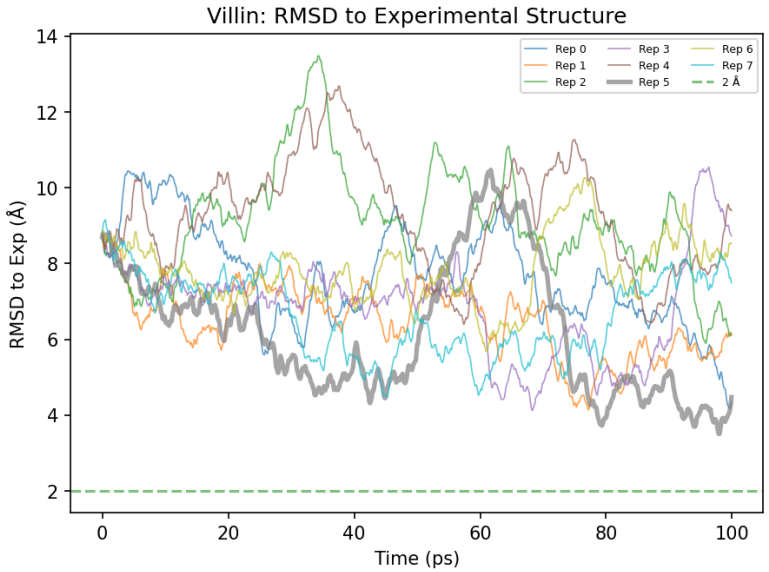
CG vs Experimental: Trp-Cage

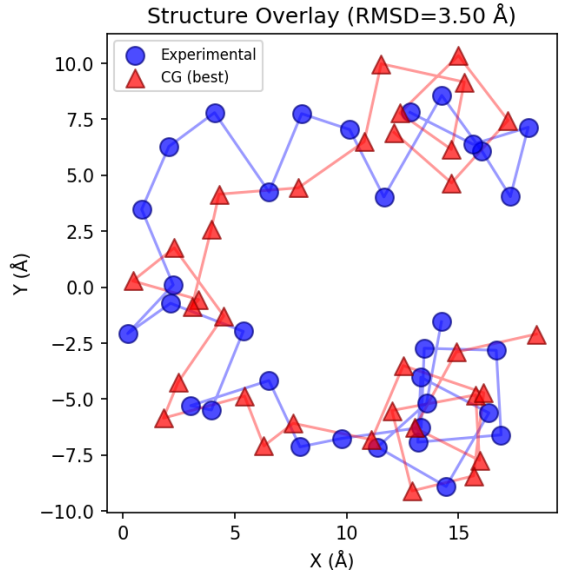


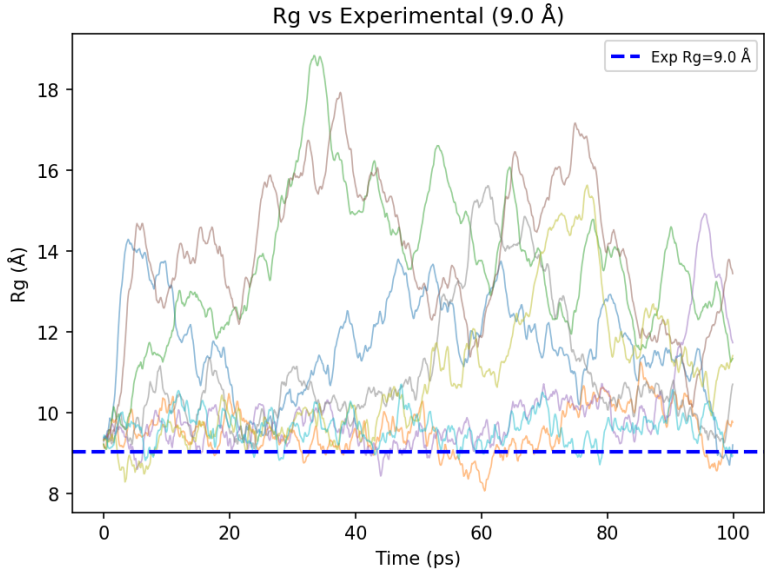




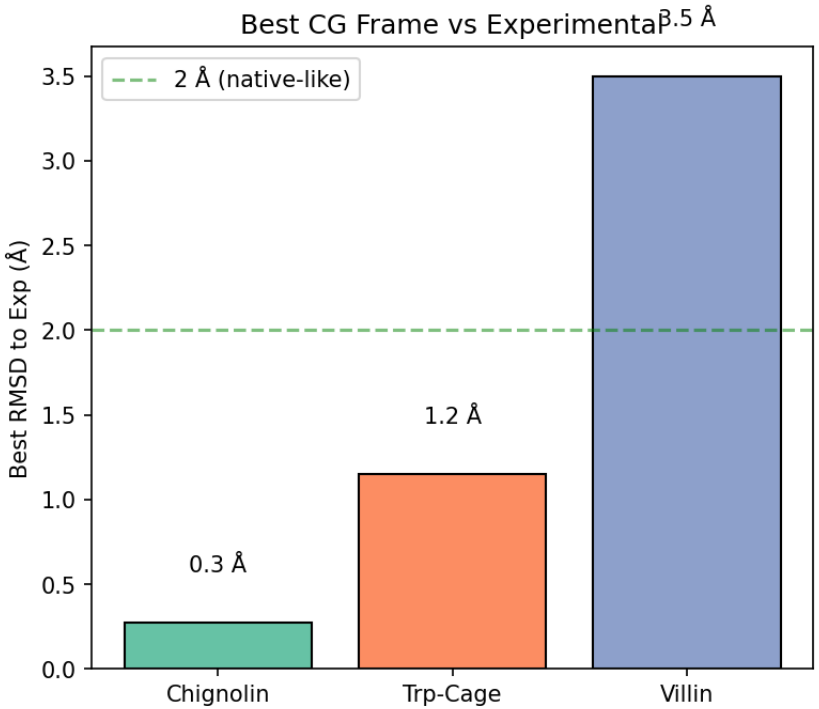
CG vs Experimental: Villin

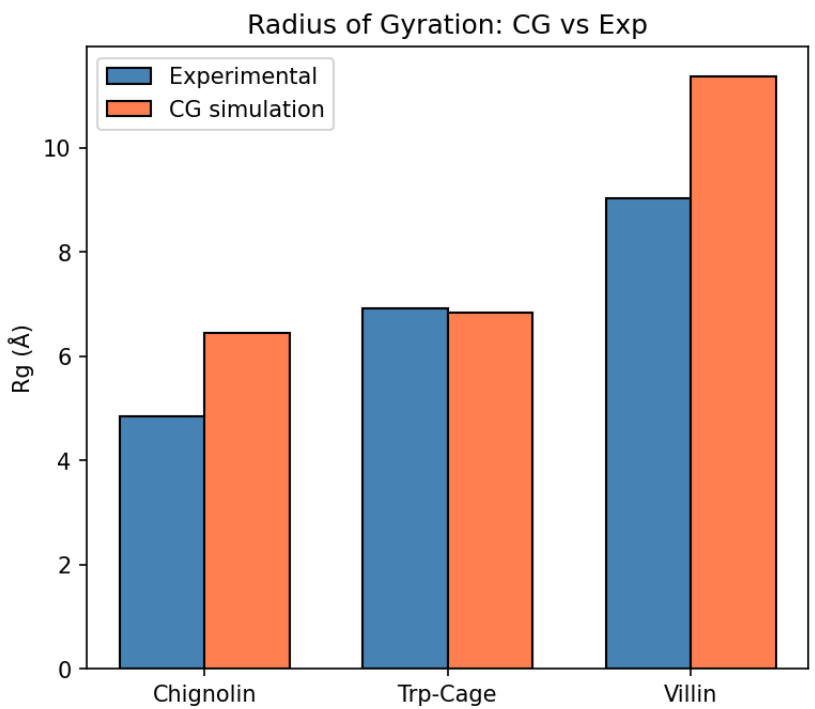






CG Simulation Validation Against Experimental Structures

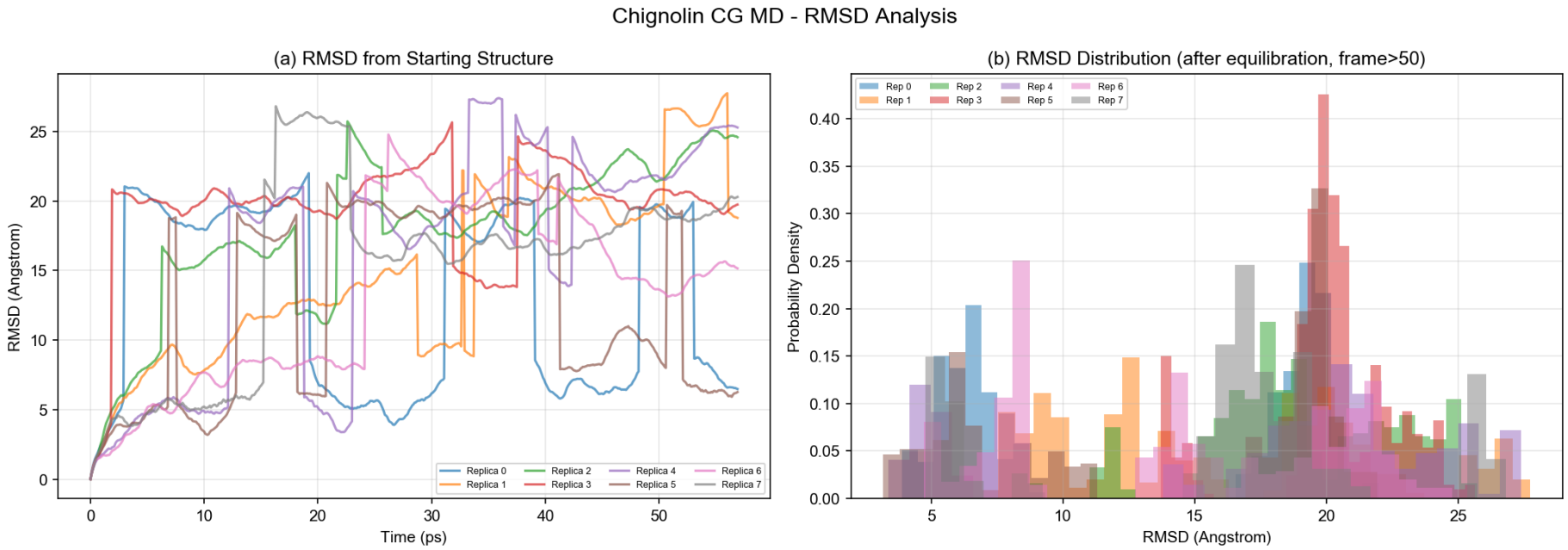


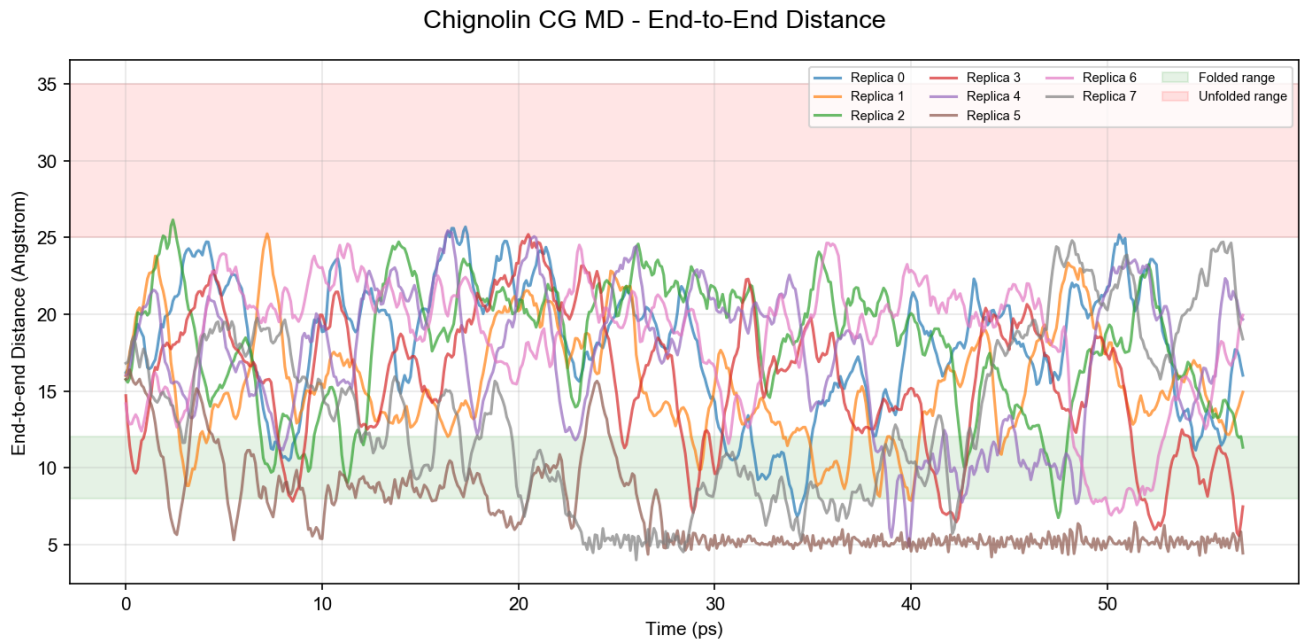


CG vs Experimental Summary

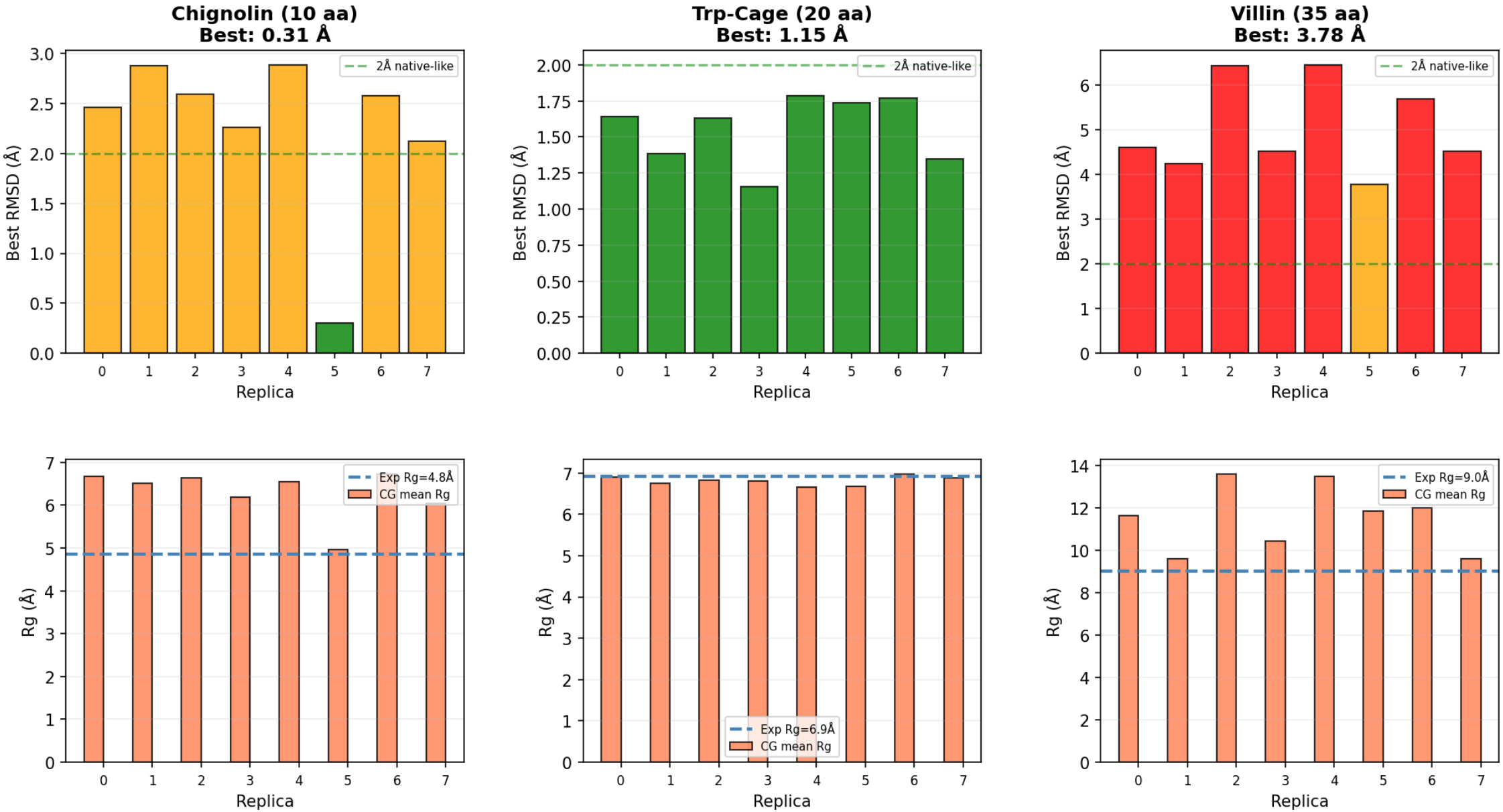
Protein	Residues	Exp Rg	CG Rg	Best RMSD
Chignolin	10	4.8	6.5	0.3
Trp-Cage	20	6.9	6.8	1.2
Villin	35	9.0	11.4	3.5

Folding Dynamics Reveal Replica-dependent Sampling

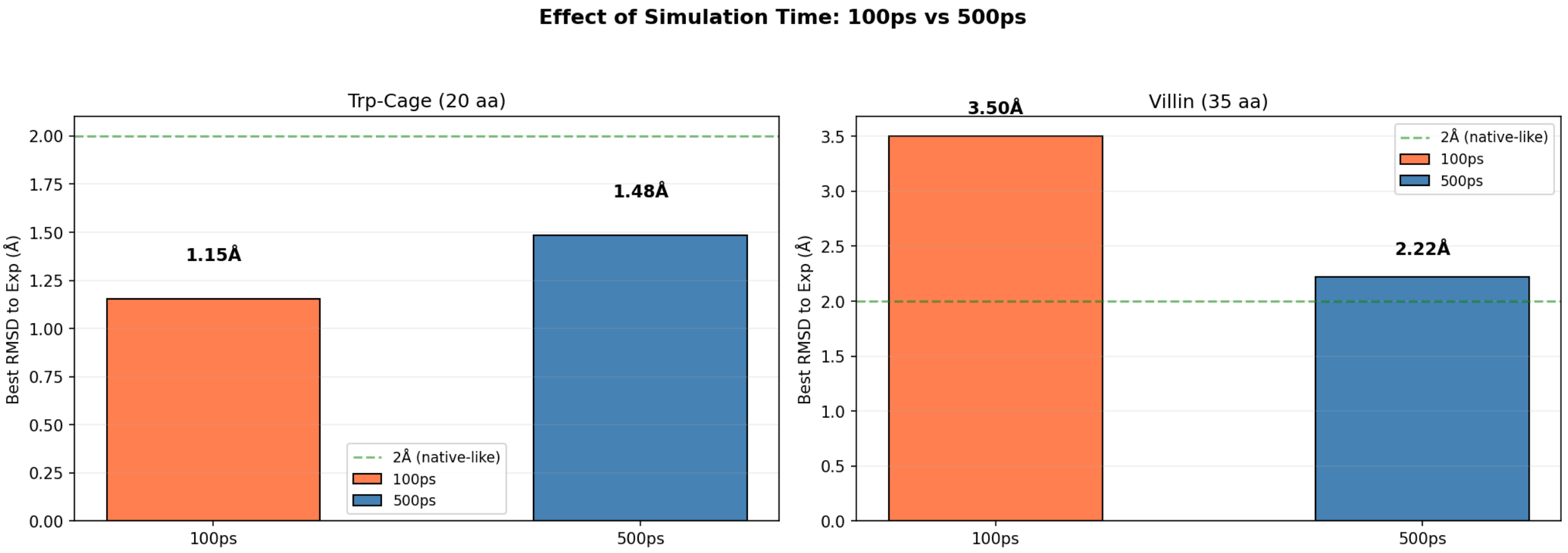




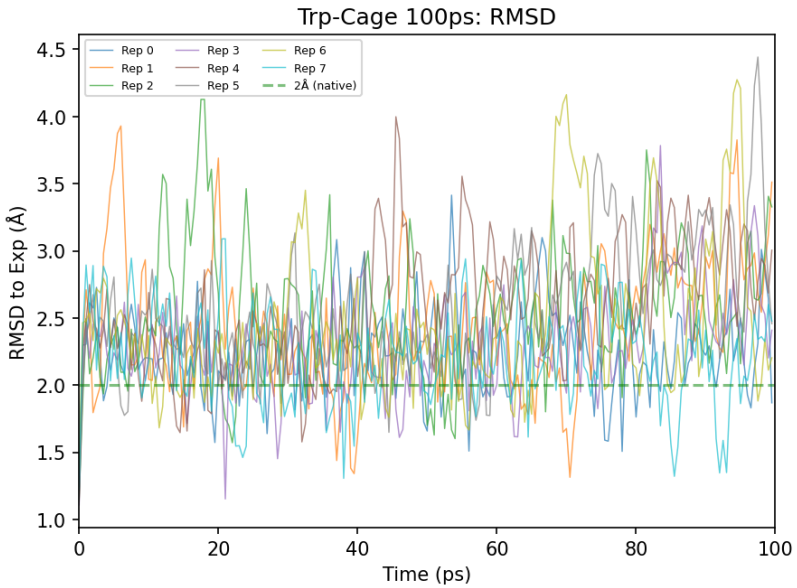
Key Results: Per-Replica Validation Against Experimental Structures

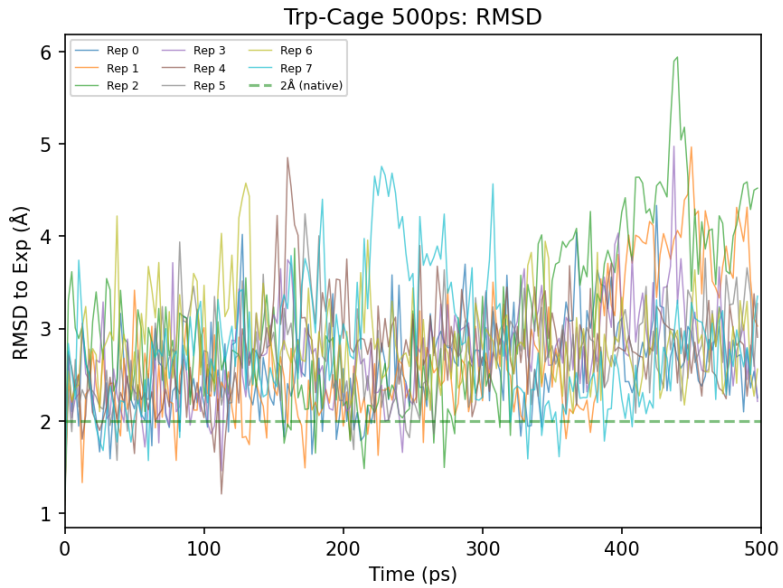


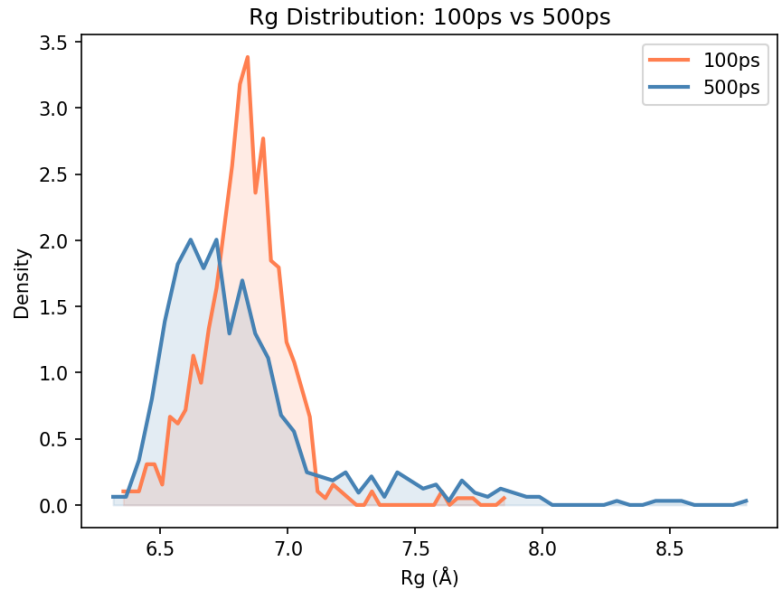
Effect of Simulation Time: 100 ps vs 500 ps

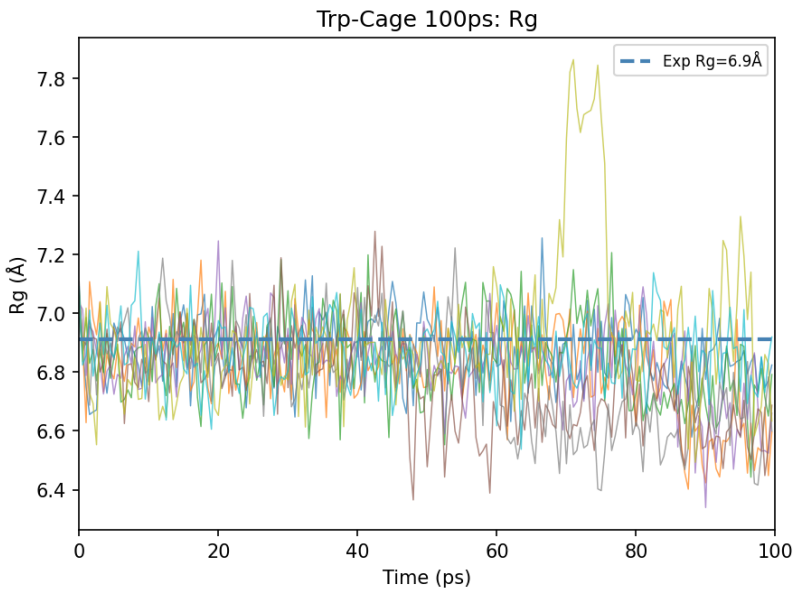


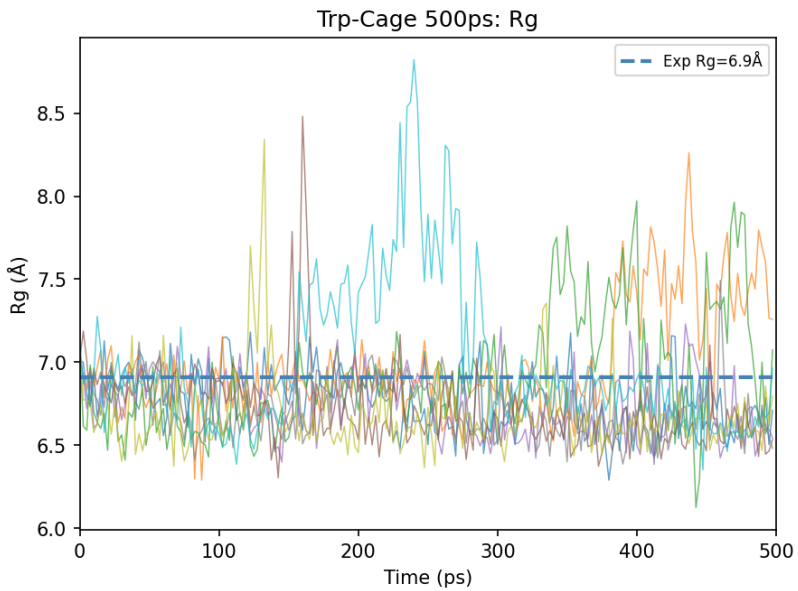
Trp-Cage: Extended Simulation (100ps → 500ps)

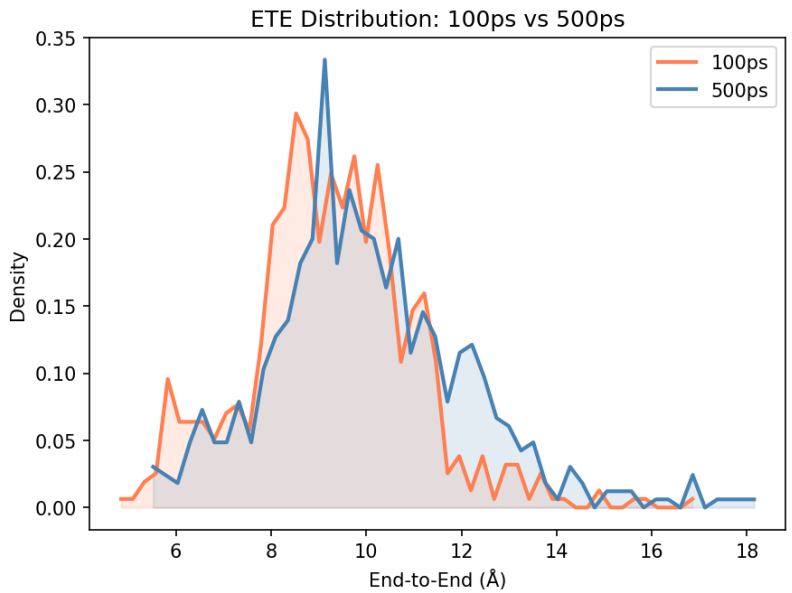




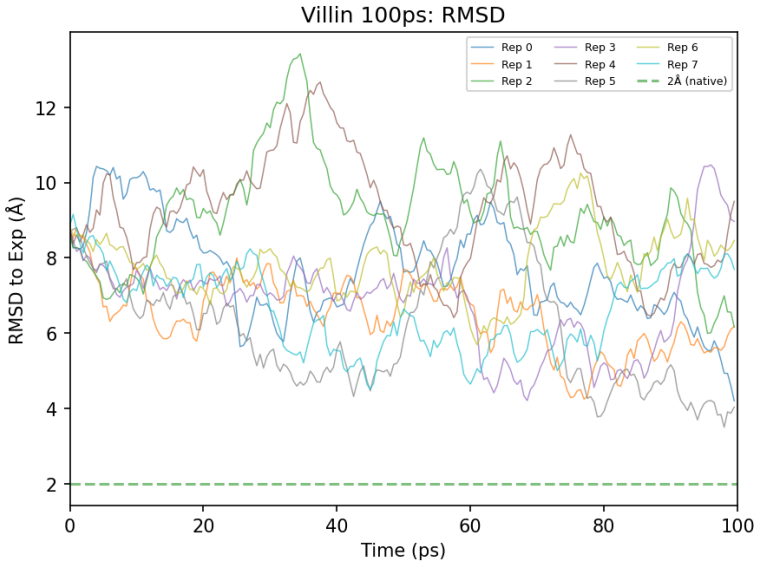


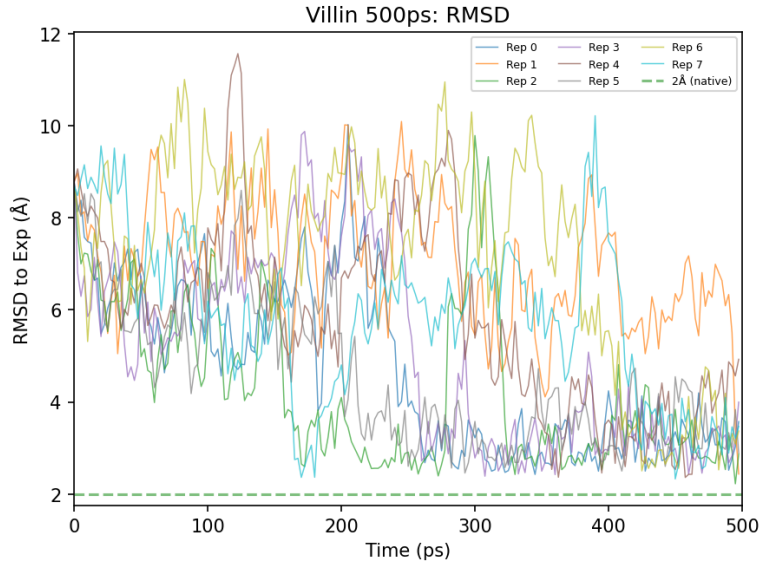


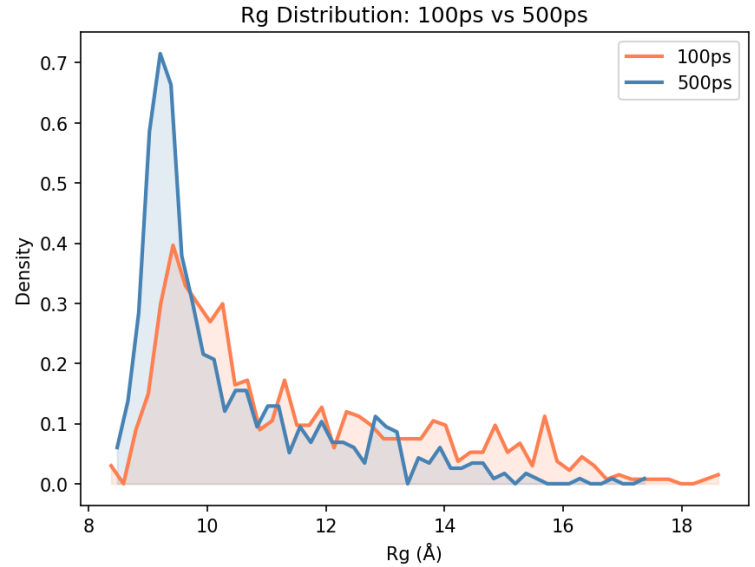


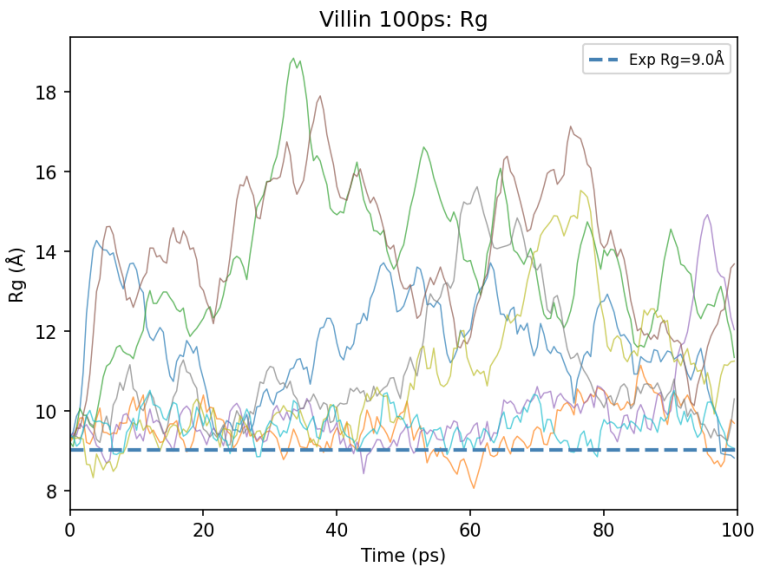


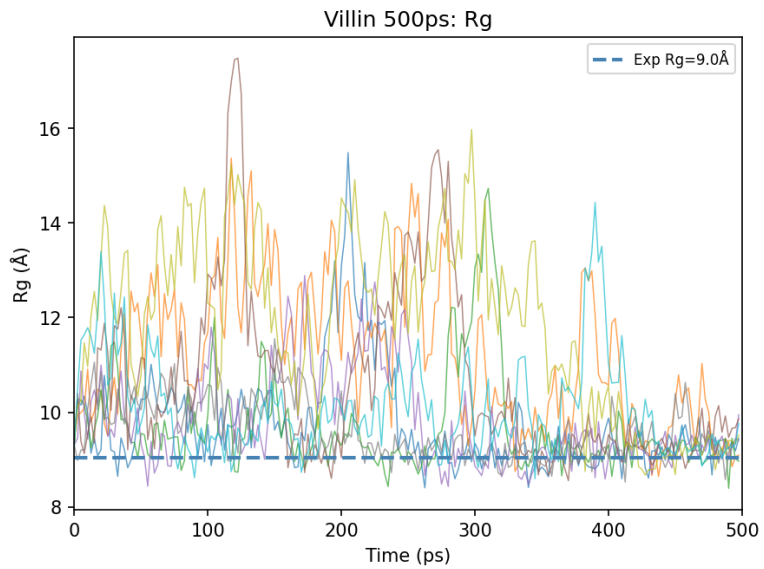
Villin: Extended Simulation (100ps → 500ps)

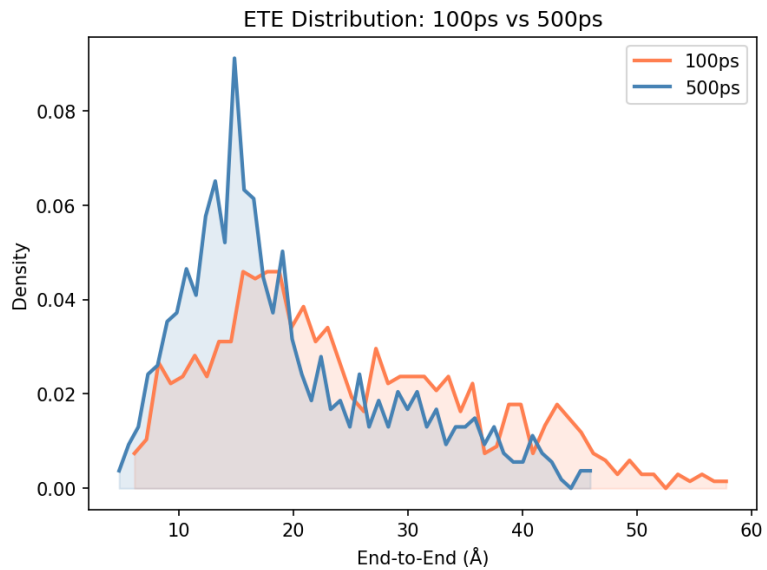




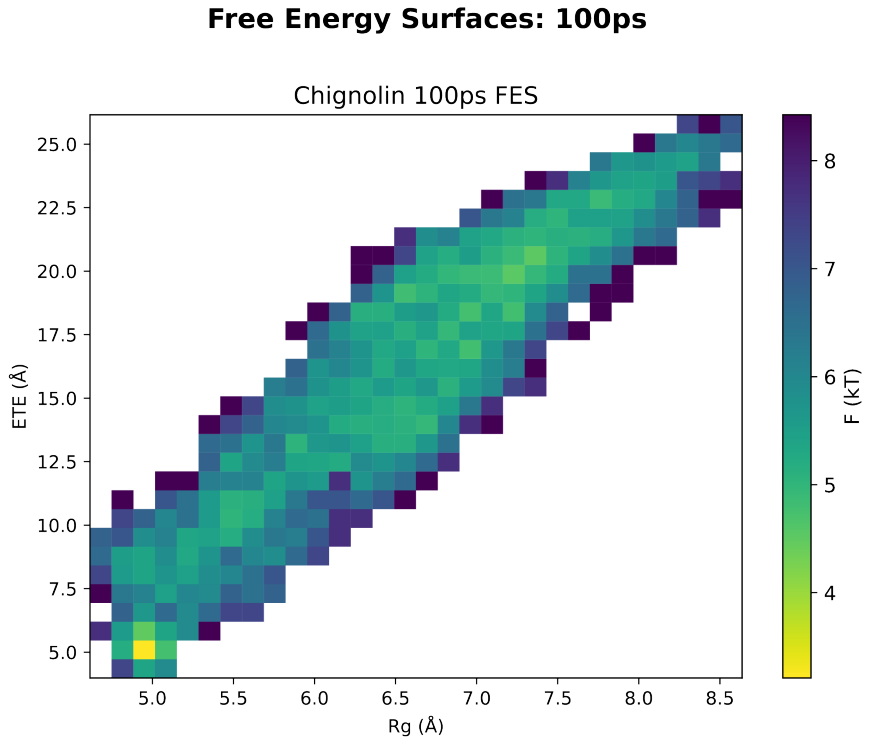


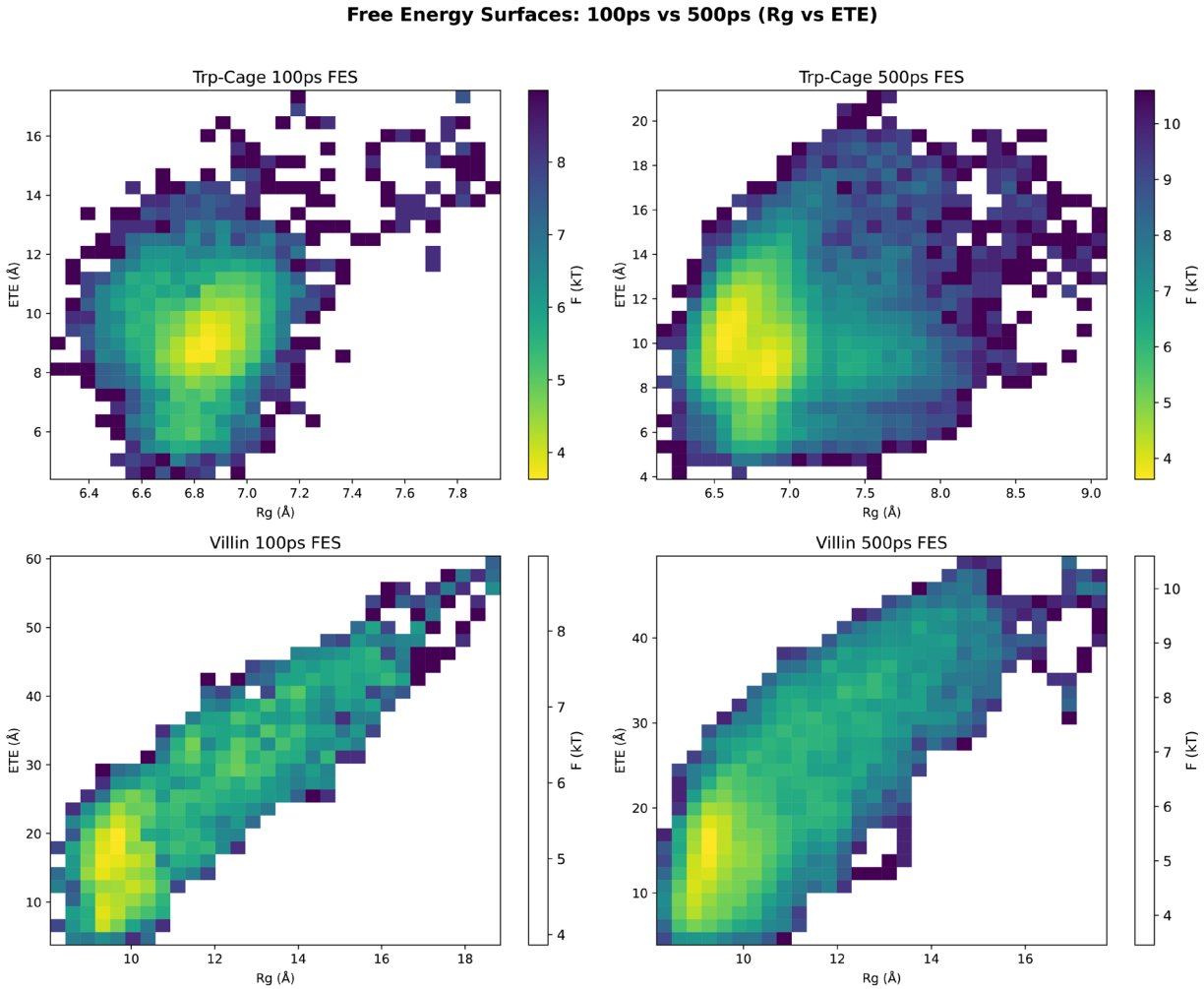




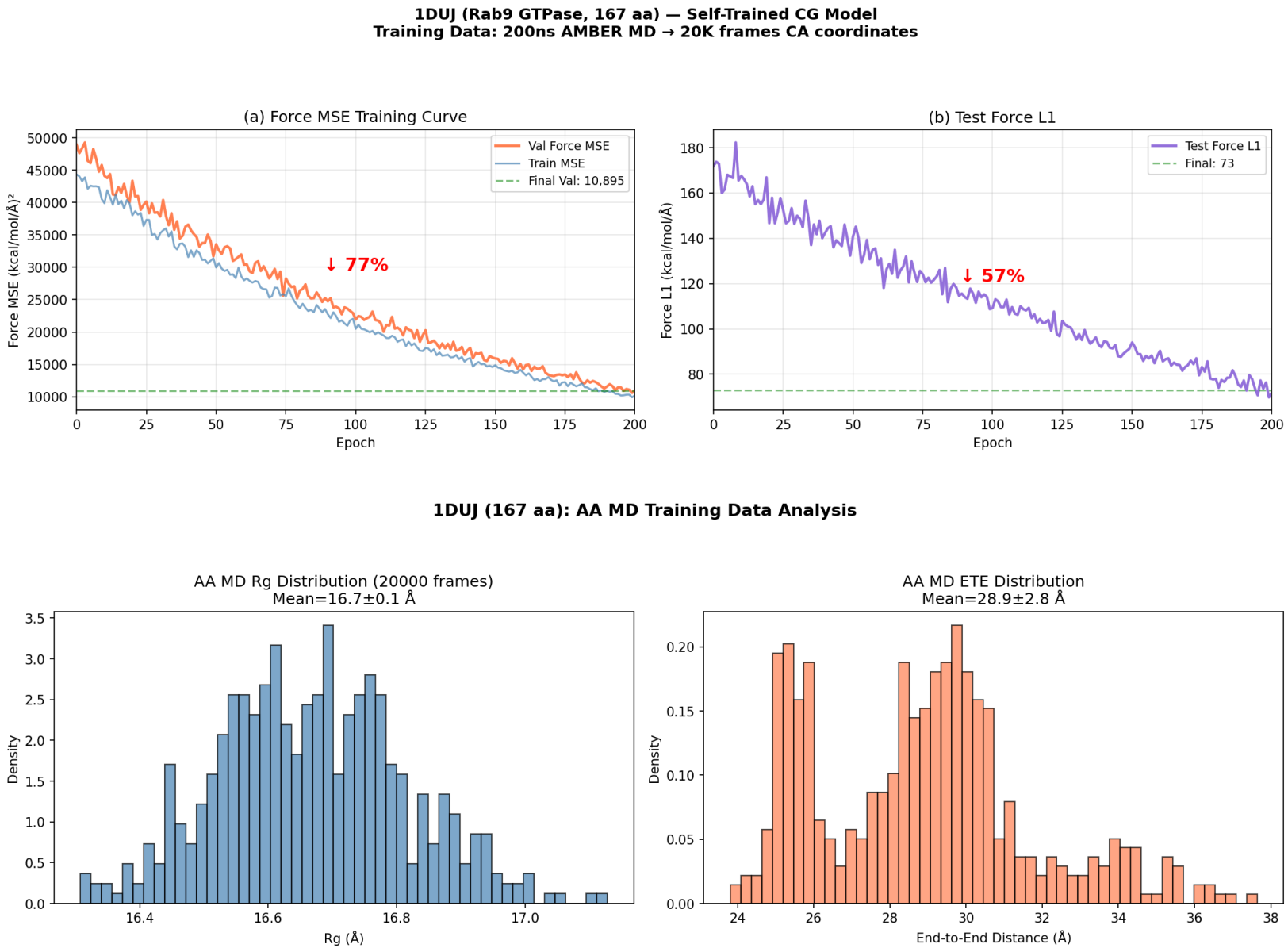


Free Energy Surface: Longer Sampling Reveals Stable States





Self-trained 1DUJ: Loss Decreases, but MD Rollout Fails



Discussion: What We Reproduced and What Remains Challenging

✓ Reproduced

- Pre-trained CG-NNP recovers native-like structures in small proteins.
- Chignolin: 0.27 Å; Trp-Cage: ~1.1 Å best RMSD.
- Longer sampling improves Villin: 3.50 Å → 2.22 Å.
- CG-MD enables faster exploration of folding-related conformations.

Evidence: RMSD, Rg, structure overlay

≈ Not fully reproduced

- Full 12-protein benchmark was beyond course-scale compute.
- No complete TICA/MSM-based free-energy surface reconstruction.
- 500 ps simulations are shorter than publication-scale CG sampling.
- Results support native-like frames, not full thermodynamic landscapes.

Scope: lightweight reproduction

! Key lesson

- 1DUJ self-training lowered loss but rollout became unstable.
- NaN appeared after ~25 ps in CG simulation monitor file.
- Lower training loss does not guarantee physical stability.
- Main bottleneck: conformational coverage and force-label quality.

Takeaway: Loss ↓ ≠ stable MD

CG-MD learns simplified forces—not static structures—and trades atomic detail for longer-timescale sampling.

AA-MD = all-atom molecular dynamics · CG-MD = coarse-grained molecular dynamics · NNP = neural network potential

Thanks for attention

Group 6

2026.5.27

Edges are defined by pairwise distances between neighboring beads

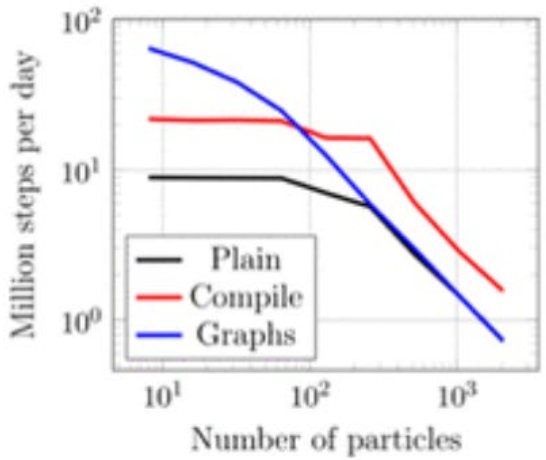
The model does not only use sequence adjacency. It builds edges according to spatial proximity, so the network can learn local structural interactions in the coarse-grained protein.

Interaction blocks perform message passing to update bead features

Through several interaction blocks, each bead collects information from its neighbors. As a result, the model learns not only bead identity, but also its structural environment.

TorchMD-Net framework for neural network potentials

TorchMD-Net 2.0



torch.compile + CUDA Graphs
Fast neighbor search
OpenMM Integration
... and more!

- Pelaez et al., 2024, TorchMD-Net 2.0: Fast Neural Network Potentials for Molecular Simulations.