Benchmarking Fold Recovery: Implicit-Solvent Atomistic & Coarse Grained Simulations of Two Vs Non Two State Proteins

by

Mimansha Das

Submitted

In partial fulfilment of the requirements for the degree of

Master of Technology

to

Indraprastha Institute of Information Technology, Delhi

August, 2025

Certificate

This is to certify that the thesis titled **Benchmarking Fold Recovery: Implicit-Solvent Atomistic & Coarse Grained Simulations of Two Vs Non Two State Proteins** being submitted by **Mimansha Das** to the Indraprastha Institute of Information Technology, Delhi, for the award of the Master of Technology, is an original research work carried out by him under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

August, 2025

Dr. N. Arul Murugan
Department of Computational Biology,
Indraprastha Institute of Information
Technology Delhi New
Delhi 110020

Acknowledgements

I would like to express my deepest gratitude to all those who have supported me and guided me throughout my M.Tech. thesis work. First and foremost, I would like to thank my esteemed project supervisor, **Dr. N. Arul Murugan**, for providing me this opportunity to work under his guidance. His wisdom and guidance have profoundly shaped my understanding and approach to this project.

I am extremely grateful to **Kapali Suri** for his involvement, insightful comments and motivation. Their insights and efforts have been invaluable, and working with them has been a rewarding experience.

Lastly, would like to thank IIIT-Delhi for providing the necessary infrastructure.

Mimansha Das

MTech CB

MT23233

Table of Contents

1. Abstract

Protein folding is a fundamental biological process through which a polypeptide adopts its functional three-dimensional structure. In this study, we systematically investigated the structural response of proteins to thermal denaturation using both atomistic and coarse-grained (CG) molecular dynamics simulations. A curated set of 138 proteins from the PFDB (89 two-state and 49 non-two-state folders) was subjected to a heat–quench protocol (300 K → 1000 K → 300 K) in implicit solvent. Structural recovery was assessed through RMSD and radius of gyration (Rg) calculations after Kabsch superimposition, alongside MM/PBSA energy evaluations.

Post-quench alignment revealed distinct behaviors: two-state proteins consistently showed lower RMSD, greater compaction ($\Delta Rg < 0$), and higher native contact retention than non-two-state proteins. Furthermore, a significant inverse correlation was observed between log10(kf) and final RMSD in the two-state subset, linking folding rate to structural resilience. Results from CG simulations mirrored these trends, validating their utility for rapid, cost-effective screening.

These findings underscore the importance of structural alignment in post-simulation analysis and highlight heat–quench recovery as a powerful proxy for foldability. The combined pipeline offers a scalable framework for evaluating folding kinetics and native-state robustness across protein families.

2. Introduction

Protein folding – the process by which a linear polypeptide chain attains its functional three-dimensional structure – is **one of the most difficult and fundamental problems in biophysics**[1]. Despite decades of research and an abundance of experimental data on folding mechanisms, predicting folding behavior and stability from sequence or structure remains challenging. Proteins can follow different folding pathways: **two-state (2S) proteins** fold via a single cooperative transition (folded ⇌ unfolded) with no stable intermediate, whereas **non-two-state (N2S) proteins** populate one or more intermediate states during folding[2]. Two-state folders exhibit single-exponential kinetics (a simple two-state transition) with no detectable intermediate, even if a transient high-energy intermediate exists, while N2S proteins have at least one *observable* intermediate in their folding/unfolding process[2]. These distinct classes often reflect differences in energy landscape topography: two-state proteins are thought to have relatively **smooth, funnel-shaped free energy landscapes** that lead directly to the native state, whereas non-two-state proteins have **rugged landscapes** with multiple local minima (folding intermediates) and higher kinetic barriers.

Understanding and comparing the stability of two-state vs. non-two-state proteins is important for protein engineering and folding theory. The **Protein Folding Database (PFDB)**[3] was recently developed as a standardized kinetic dataset, containing 141 single-domain globular proteins (89 classified as two-state folders and 52 as non-two-state) with folding/unfolding rate constants normalized to 25 °C[4][4]. The present work leverages this dataset (using 138 proteins from PFDB) to investigate protein stability and folding behavior in silico. We employ molecular dynamics (MD) simulations to simulate *thermal unfolding and refolding* for each protein and analyze differences between two-state and non-two-state categories.

Molecular dynamics simulations allow us to probe the folding energy landscape by observing how protein structures respond to thermal perturbation. However, **all-atom MD with explicit solvent** is computationally expensive for folding simulations, typically limited to microsecond timescales for small proteins[5]. To make folding simulations tractable for dozens of proteins, we adopt two strategies: (1) use an **implicit solvent all-atom MD approach** (Generalized Born solvent) which drastically reduces computational cost[6], and (2) use **coarse-grained (CG) simulations** with a structure-based force field to further speed up sampling. **Implicit solvent MD** treats solvent as a continuous medium, avoiding the need to simulate thousands of water molecules explicitly[7]. This approach, combined with modern force fields, has been shown to fold small proteins on nanosecond to microsecond timescales – e.g. *Nguyen et al.* (2014) demonstrated that **folding simulations for proteins with diverse topologies are accessible in days using physics-based force fields and**

implicit solvent[8]. Meanwhile, **coarse-grained models** simplify protein representation (e.g. one bead per amino acid or a few beads per residue) and bias the energy function toward the native state, massively accelerating folding/unfolding simulations[9]. Many aspects of protein folding dynamics and native structure stability are *successfully captured by structure-based Gō-like coarse-grained models*, in which only native contacts are explicitly favored[9]. These models essentially create a funneled energy landscape that guides the protein towards its known native conformation via native interactions[9]. An example is the AWSEM (Associative memory, Water-mediated, Structure and Energy Model) coarse-grained force field, which **contains physically motivated terms (e.g. hydrogen bonding) and bioinformatically based local structure biases** to realistically stabilize native-like structures[10].

In this thesis, we performed **thermal unfolding and refolding simulations** on 138 PFDB proteins under two modeling resolutions: (1) all-atom MD in implicit solvent (Amber force field), and (2) coarse-grained MD in implicit solvent. The *goal* is to assess protein stability and folding propensity by subjecting each protein to high-temperature unfolding and then cooling (quenching), and to compare how two-state vs. non-two-state proteins behave under these conditions. Key questions include: Do two-state proteins, with their ostensibly smoother folding funnels, refold more completely (or easily) after thermal denaturation than non-two-state proteins? Can quick **heat-quench simulations** serve as a proxy to rank protein foldability or stability? And how do results differ between a detailed atomistic model and a simplified coarse-grained model?

We analyze structural metrics such as **root-mean-square deviation (RMSD)** from the native structure and **radius of gyration (R_g)** before and after unfolding/refolding, as well as the recovery of **native contacts (Q)**. We also estimate the **free energy change upon unfolding** for each protein using MM-PBSA (Molecular Mechanics Poisson–Boltzmann Surface Area) calculations, an end-point free energy method combining molecular mechanics energies and continuum solvation[11]. Finally, by leveraging the kinetic data in PFDB, we examine correlations between our simulation-derived measures of "foldability" and the experimental folding rates (ln k_f) for two-state and non-two-state proteins. This allows us to validate whether the simulation outcomes reflect known experimental trends (e.g. the well-known **negative correlation between folding rate and native topology complexity** – such as contact order – for two-state proteins[12]).

In summary, this project provides a detailed computational study of protein folding stability across a large and diverse set of proteins. By comparing two-state vs. non-two-state proteins in atomistic and coarse-grained simulations, we gain insight into how energy landscape differences manifest in **thermal refolding behavior** and how well simplified models can reproduce these differences. The findings demonstrate distinct stability and refoldability characteristics for the two classes, consistent with their kinetic folding mechanisms, and

point toward practical approaches (combining coarse-grained screening with all-atom refinement) for studying protein foldability.

2.1 Motivation

Proteins must fold into specific three-dimensional structures to function. Even small deviations from a native fold can disrupt binding, catalysis, and regulation, while irreversible misfolding underlies numerous pathologies. Despite impressive progress in structure prediction, **capturing how native structure is maintained or recovered under stress**—and how that recovery relates to intrinsic folding kinetics—remains a central, practical problem for stability engineering and basic biophysics alike.

2.2 Why simulations, and why this study?

Experimental probes of folding and stability provide ground truth but are time-consuming and system-specific. Molecular dynamics (MD) offers a complementary route: it can perturb a protein in well-controlled ways (e.g., thermal denaturation) and observe structural response at atomic detail. However, **three obstacles** limit routine, comparative use across many proteins:

1. **Scale and cost.** Atomistic MD is expensive, making broad surveys across tens to hundreds of proteins prohibitive without careful protocol design.
2. **Modalities and consistency.** Coarse-grained (CG) models promise large speedups but raise questions about how well trends transfer between CG and atomistic resolutions—especially under non-equilibrium protocols such as heat–quench.
3. **Metrics and artifacts.** Widely used metrics (RMSD, Rg) are sensitive to rigid-body motion; without careful superimposition, they can **misdiagnose "instability"** that is actually global drift rather than structural deformation.

2.3 Gaps in Current Practice

Large-scale simulation studies often emphasize equilibrium sampling near room temperature or isolated case studies of unfolding. Fewer works systematically **benchmark heat–quench**

recovery (300 K → high-T → 300 K) across a diverse panel of proteins and explicitly **link outcomes to known folding kinetics** (e.g., two-state vs non-two-state folders, rates *k*f). Moreover, cross-modal validation—showing that **fast, implicit-solvent CG** runs preserve the same class-specific stability signals seen in **implicit-solvent atomistic** runs, has not been standardized into a practical pipeline.

Scientific and practical motivations of our approach.

- **Stress-test the fold, not just the equilibrium.** A heat–quench protocol exposes marginal instabilities and kinetic traps that may be invisible in short, near-native simulations. Measuring **post-quench recovery** provides a direct proxy for foldability and resilience.
- **Make metrics trustworthy.** By enforcing **Kabsch superimposition** before computing RMSD and radius of gyration (Rg), we suppress rigid-body artifacts and isolate **true structural deformation**.
- **Connect dynamics to kinetics.** Partitioning the dataset into **two-state (2S)** and **non-two-state (N2S)** proteins enables mechanistic comparisons: if 2S proteins recover more completely and show a **negative association between log10(kf) and post-quench RMSD**, that supports the idea that a simpler folding landscape confers thermal resilience.
- **Scale without losing signal.** Running the same protocol in **CG** (implicit solvent) tests whether a rapid, low-cost surrogate can **rank stability and reproduce class-specific trends**, enabling triage at scale before atomistic refinement.
- **Energy perspective.** Complementing geometry (RMSD, ΔRg) with **MM/PBSA energy differences** between 300 K and quenched structures offers an energetics-based view of recovery, strengthening conclusions beyond purely geometric criteria.

2.4 Impact for protein engineering and screening.

A validated, end-to-end pipeline that (i) perturbs, (ii) aligns, (iii) quantifies recovery, and (iv) cross-checks geometry with energetics provides an actionable **stability scorecard** for many proteins at once. In practice, CG runs can **rapidly flag robust vs fragile candidates**, while targeted atomistic simulations confirm and resolve edge cases. This hierarchy addresses compute constraints without sacrificing interpretability, and it is immediately useful for **variant prioritization**, **domain selection**, and **library curation** in design projects.

2.5 Why two-state vs non-two-state?

Folding mechanism is a principled axis along which to expect stability differences. **Two-state folders**—with relatively smooth landscapes—should, in theory, **re-anneal efficiently** after thermal disruption. **Non-two-state proteins**, with intermediates and ruggedness, are more likely to become trapped in partially folded states after quench. Demonstrating these predicted **class-specific recovery patterns** strengthens the biophysical grounding of the pipeline and provides a sanity check across diverse sequences and topologies.

3. Methods

3.1 Dataset of Proteins and Classification

We selected **138 single-domain globular proteins** from the PFDB dataset[4] for simulation. These proteins span a range of sizes (typically 35–250 residues) and structural classes (α, β, α/β, etc.), and importantly are experimentally classified as either two-state (2S) or non-two-state (N2S) folders. In PFDB, folding type is clearly specified based on experimental kinetic behavior: proteins that show a single-exponential folding/unfolding curve with no detectable intermediate are labeled 2S, whereas those with a stable intermediate in their folding pathway are labeled N2S[2]. Our selection included roughly equal representation of both categories (approximately 86 2S and 52 N2S, matching the PFDB proportions[4]). For each protein, we obtained the experimentally resolved native structure from the Protein Data Bank (PDB) as the starting conformation.

Prior to simulation, each protein structure was processed and prepared using Amber Tools. Hydrogen atoms were added consistent with *pH 7* protonation, missing side-chains or loops (if any) were built, and we applied **hydrogen mass repartitioning (HMR)** to improve integration stability. In HMR, the hydrogen atom masses are increased (with a corresponding decrease in the heavy-atom masses) – e.g. Amber's Parmed tool reassigns hydrogen masses to 3.024 Da[13] – allowing the use of a larger integration time step (4 fs instead of the standard 2 fs) without compromising accuracy, since high-frequency bond vibrations (involving light hydrogens) are dampened. This technique was applied to all all-atom systems to enhance simulation efficiency.

3.2 All-Atom MD Simulations in Implicit Solvent

All-atom molecular dynamics simulations were carried out using the **Amber MD package** (PMEMD.cuda) with an implicit solvent model. We used the Amber **ff14SB** family force field for protein atomic interactions (ff14SBonlysc variant was used during system

preparation[14], which updates side-chain parameters). Solvent was represented implicitly via the **Generalized Born (GB)** model with parameter set GB-Neck2 (Amber igb=8), which provides a continuum dielectric approximation of water[15]. A cutoff of 1000 Å was set for nonbonded interactions in implicit solvent (effectively no cutoff) to ensure all intramolecular interactions are considered[16]. No periodic boundary conditions were applied (ntb=0), as appropriate for implicit solvent[15]. Surface area contributions to solvation free energy were included using the default surface tension (0.0072 kcal/mol·Å²) for GB models[17]. **Salt concentration** was set to 0 (saltcon=0) since we assume pure water dielectric with no explicit ions[17].

Each protein underwent a multi-step simulation protocol to simulate folding/unfolding:

- **Energy Minimization:** First, we minimized each structure to remove any steric clashes or bad contacts introduced during preparation. A short steepest-descent + conjugate-gradient minimization (e.g. 1000 steps) was performed with harmonic positional restraints on the backbone heavy atoms to relax side-chains and solvent-exposed loops. Then a second minimization with weaker or no restraints relaxed the entire structure. This ensured a low-energy starting conformation without significant deviation from the experimental fold.

- **Equilibration at 300 K:** Next, we heated and equilibrated the protein in implicit solvent at **300 K** (approximately room temperature, near typical experimental conditions). Heating was done gradually to avoid shocks; we used a **Langevin thermostat** (collision frequency γ = 1 ps^−1) to raise the temperature from a low initial temperature (typically 0 K or 10 K) to 300 K over a period of time. In practice, we implemented a temperature ramp using Amber's nmropt facility: for example, temperature was linearly increased from 10 K to 300 K over the first 0.5–1 ns of simulation, then the system was held at 300 K for an additional period to reach equilibrium[18][19]. Throughout heating, a small positional restraint on backbone atoms was maintained (e.g. force constant 1 kcal/mol·Å²) to prevent premature unfolding while allowing local relaxation[20][21]. After reaching 300 K, restraints were gradually removed and the simulation continued until the protein's RMSD and energy stabilized, indicating it had equilibrated in its folded state at 300 K. **Velocity rescaling or Langevin dynamics** maintained the target temperature (300 K) during this stage.

- **High-Temperature Unfolding (1000 K heating):** After obtaining a stable 300 K structure (considered the "folded" state), we initiated unfolding by drastically increasing the temperature. The system was heated from 300 K up to **1000 K** and simulated at 1000 K to induce unfolding. To do this safely, we again ramped the

thermostat setpoint over some interval (for example, increasing from 300 K to 1000 K over tens of picoseconds to a few hundred picoseconds, to reduce risk of integration error or explosive forces). Once at 1000 K, each protein was simulated for a certain duration (on the order of 1–2 ns) to allow it to explore highly unfolded conformations. A Langevin thermostat with a higher collision frequency (we used γ = 5 ps^−1 for stability at extreme temperature) controlled the temperature. The choice of 1000 K is intentionally high – far above physiological temperature – to overcome folding free energy barriers rapidly and drive the protein toward a denatured state on a feasible timescale[22]. Similar high-temperature MD protocols have been used in the literature to efficiently sample unfolded conformations; for example, **heating to 1000 K for a few nanoseconds can generate a broad ensemble of denatured structures**[22]. We note that 1000 K is a non-physical condition that can distort some interactions, but as a computational tool it helps us surmount kinetic traps and approximate the unfolding process that might normally occur via chemical denaturants or slower heating.

- **Cooling (Refolding) to 300 K:** Immediately after the high-temperature run, we **quenched** the protein back down to 300 K to observe how it refolds (or misfolds) upon cooling. The temperature was brought from 1000 K back to 300 K, either abruptly or in staged decrements (we performed a rapid cooling over tens of picoseconds, effectively an instantaneous quench relative to folding timescales). Following this temperature drop, the protein was simulated at 300 K for another period (e.g. 1–2 ns) to allow it to relax at the lower temperature. This mimics a **"heat shock" folding experiment** in silico: the protein is unfolded by heat, then cooled to see if it can find its native basin again. Any residual structure present just after the 1000 K phase could serve as a nucleus for refolding, or the protein might adopt a non-native collapsed state upon cooling. We performed these unfolding/refolding simulations without any restraints, allowing the protein to freely evolve.

For each protein, we thus obtained three key simulation snapshots or states: the initial equilibrated **folded structure at 300 K**, an **unfolded structure at 1000 K** (or an ensemble of structures sampled at high T), and a **post-quench structure at 300 K** (after cooling). All simulations were run in the NVT ensemble (constant volume, implicit solvent) with a 4 fs time step (enabled by HMR). Bond lengths involving hydrogens were constrained (Amber SHAKE algorithm) to allow this larger time step. Trajectories were saved at regular intervals (e.g. every 1–2 ps) for analysis.

3.3 Coarse-Grained Simulations in Implicit Solvent

In parallel with all-atom simulations, we carried out **coarse-grained (CG) MD simulations** for the same set of proteins using a structure-based coarse-grained model. The coarse-grained approach dramatically reduces computational complexity by representing each protein with fewer particles and simplified interactions. In our case, each amino acid was represented by a single bead (located at the C^α position of each residue), yielding a C^α-only model of the protein. The **force field** for CG simulations was designed in a Gō-like manner: **native contacts are assigned attractive potentials**, whereas non-native interactions are generally simplified or treated as repulsive to avoid incorrect aggregation[23][24]. This means the known experimental native structure is encoded as the global free energy minimum of the CG force field – effectively biasing the simulation toward the native fold. Such **structure-based Gō models** are widely used in folding studies; they retain the essential topology of the protein's energy landscape but remove many of the energetic frustrations present in reality, thereby accelerating folding kinetics by orders of magnitude[9]. Many studies have shown that **Gō-like models can capture the essential folding mechanisms and native state properties** for small proteins[9], although they may not account for non-native intermediates or specific sequence-dependent effects.

For our simulations, we used an enhanced Gō-like potential similar in spirit to the AWSEM model[25]. Specifically, the CG force field included: a backbone chain connectivity term (bonded potentials maintaining peptide chain geometry), **native contact attractions** (typically modeled with Lennard-Jones or harmonic wells centered at the native C^α–C^α distances for all residue pairs that are in contact in the PDB structure), and generic repulsions or soft restraints to prevent unphysical collapse of non-native contacts. The AWSEM model, for example, includes hydrogen-bonding terms and knowledge-based amino acid pair potentials[10]; our model incorporated similar physically motivated terms to stabilize secondary structures and side-chain packing implicitly. Solvent effects at the CG level were treated implicitly by these effective potentials (e.g. hydrophobic interactions are captured by stronger native contact attraction between nonpolar residues, and an implicit solvent term discourages over-expansion).

The coarse-grained simulations followed an analogous protocol to the all-atom simulations:

- Each protein's *native structure* (PDB structure) was used to define the CG model's native contact map and as the starting conformation.
- We performed an energy minimization or relaxation in the CG force field to eliminate any minor inconsistencies from the mapping (this was usually trivial since the native structure is an energy minimum by construction).

- The protein was then **heated to 300 K** in the CG model and equilibrated briefly at 300 K (confirming that it remains at the native basin at low temperature, as expected for a structure-based model).
- Next, we **increased the temperature to a high value (1000 K)** in the CG simulation. Because the CG model's energy scales differ from the all-atom model, "1000 K" here is in a nominal sense; we chose a high temperature sufficient to overcome the native contact attractions. In practice, we calibrated the high temperature so that the protein would unfold in the CG model. (For many structure-based models, a certain temperature T_unfold exists at which the native state loses stability. We set our simulation temperature above this threshold, e.g. 3–5 times the folding transition temperature, to ensure complete unfolding.)
- The protein was simulated at this high temperature (CG 1000 K equivalent) for a period to let it sample unfolded conformations. Because the CG model is less computationally intensive, we could afford longer unfolding simulations (e.g. 5–10 ns or more) and even multiple repeats to sample different unfolding trajectories.
- Finally, we **cooled the CG system back to 300 K** to allow refolding. In a structure-based model, if cooled slowly or even quenched, the protein will generally return to the native state (since that is the global minimum). We performed a rapid quench similar to the atomistic case and observed the "refolding" behavior upon cooling.

It should be noted that coarse-grained models often **fold much faster** than real proteins or all-atom simulations because of reduced degrees of freedom and smoother energy landscapes. This allowed us to use **shorter simulation times** for the CG runs. In fact, we found that a single heat-quench cycle in the CG model was **~10–50× faster** to execute than the equivalent all-atom simulation for a given protein, even when using longer time steps in all-atom【11†Image】. This huge speed-up enabled us to simulate all 138 proteins in a reasonable time and even perform multiple independent trials for statistical reliability.

The benefit of running both atomistic and CG simulations is to compare their outcomes: the **CG simulations serve as a high-throughput screen** for general folding behavior, whereas the all-atom simulations provide **higher fidelity details**. The CG model, by construction, will favor refolding to the correct native contacts, but the ease or difficulty of refolding in the CG simulation (e.g. whether it requires slow cooling, or if it gets trapped in misfolded states if any non-native minima exist) can still differ between proteins and may correlate with experimental foldability. The all-atom simulations, on the other hand, include full atomic interactions and realistic energetics, so they can capture phenomena like side-chain rearrangements, secondary structure disruption, or salt-bridge reformation during folding. By comparing the two, we can assess how much of the folding behavior is simply dictated by native topology (which CG Gō models emphasize) versus finer energetic details.

All simulations (both atomistic and CG) were carried out on GPU-accelerated computing resources. Trajectory data from both methods were saved for analysis. In total, for each protein we obtained a **folded state (300 K start)**, an **unfolded state (during 1000 K heating)**, and a **refolded state (300 K after quench)** in both all-atom and coarse-grained representations.

4. Analysis and Calculations

After simulations, we performed a comprehensive analysis to quantify the structural and energetic changes upon unfolding/refolding:

- **Root-Mean-Square Deviation (RMSD):** We measured the C^α-RMSD of the protein's structure after refolding (post-quench 300 K) with respect to the original native structure. Prior to RMSD calculation, the protein was **superimposed onto the native structure using Kabsch alignment** (a least-squares fitting of C^α positions) to remove any global translation/rotation. This alignment is crucial – without it, a protein that drifted or rotated during simulation would show an inflated RMSD that is not reflective of fold disruption. By aligning, we ensure the RMSD reflects differences in internal conformation only. We report RMSD in Angstroms (Å). For each protein, we computed RMSD at the start (native 300 K structure) and after refolding. **Low RMSD values (e.g. 2–4 Å)** indicate the protein returned close to its native conformation, whereas **high RMSD (e.g. >6–8 Å)** indicates significant deviation (partial or complete misfolding).

- **Radius of Gyration (R_g):** We calculated the radius of gyration of the protein (the mass-weighted root-mean-square distance of atoms from the center of mass) in the folded native state vs. after refolding. R_g is a measure of the protein's compactness. The native folded state has a certain R_g (varies with protein size and topology). Upon unfolding at 1000 K, R_g increases substantially as the protein expands. After cooling, if the protein refolds or re-compacts, R_g should decrease again. We computed **ΔR_g = R_{g,quenched} – R_{g,native}** for each protein, so that a negative ΔR_g indicates the refolded structure is more compact than the original (or basically regained compactness), while a positive

ΔR_g would mean the refolded structure remains more expanded than native. We expect two-state proteins – if they refold cooperatively – to regain a compact R_g close to the native value (ΔR_g near 0 or slightly negative/positive within a small range). Non-two-state proteins might remain partially expanded (positive ΔR_g) if they fail to fully collapse.

- **Secondary Structure and Contacts:** We analyzed the secondary structure content (α-helix, β-strand) before and after to see if the protein's secondary structure reformed. However, a more direct metric of native tertiary structure recovery is the **fraction of native contacts, Q**. We defined Q as the fraction of all native residue-residue contacts (within a cutoff distance in the native state) that are present in the refolded structure. Q = 1 (or 100%) means the protein has reformed all native contacts (essentially a perfect refold), whereas a lower Q (e.g. 0.5 or 50%) indicates only half of the native contacts are present (signifying a partially folded or misfolded structure). We computed Q for each post-quench structure against its native contact list. This measure is less sensitive to global domain reorientation than RMSD and complements RMSD by focusing on specific native interactions.

- **Free Energy Calculations (MM-PBSA):** To estimate the **stability difference** between the folded and unfolded states, we utilized the MM-PBSA approach on representative structures. MM-PBSA (Molecular Mechanics Poisson–Boltzmann Surface Area) is an end-point free energy calculation that combines molecular mechanics energies (bond, angle, dihedral, van der Waals, electrostatic) with continuum solvation energies (polar solvation via Poisson–Boltzmann equation, and nonpolar solvation via solvent-accessible surface area)[26][27]. Entropic contributions (from conformational entropy) are often neglected in standard MM-PBSA due to difficulty in estimation[28], so the focus is on enthalpic components and solvation. For each protein, we took the **equilibrated 300 K folded structure** and the **thermally denatured structure** (in practice, we used the structure at the end of the 1000 K phase, or an average of a few high-temperature snapshots which were largely unfolded). We then computed the MM-PBSA free energy for each:

- *G_folded* = E_gas (folded) + G_solvent (folded),
- *G_unfolded* = E_gas (unfolded) + G_solvent (unfolded), where E_gas includes bonded and nonbonded energies in vacuum, and G_solvent includes polar and nonpolar solvation terms. The **free energy difference ΔG = G_unfolded – G_folded** then approximates the folding free energy (i.e. negative ΔG implies the folded state is lower in free energy and thus more stable). We caution that these absolute values are approximate – typical protein folding free energies are on the order of only tens of kcal/mol[29], which is the small difference between two large energies that

MM-PBSA must estimate. However, comparing ΔG between different proteins or classes can yield insights. We averaged the MM-PBSA results over a small ensemble of snapshots for better robustness.

We also computed an analogous ΔG between the **post-quench structure at 300 K** and the original folded structure, to see if the refolded/misfolded state had a higher free energy (which it should if it's less stable). Essentially, a large positive ΔG (unfolded minus folded) indicates a big energetic penalty for unfolding (i.e. the native state was much more favorable), whereas a smaller ΔG might indicate marginal stability.

- **Correlation with Experimental Data:** Finally, we compared our simulation-derived metrics (RMSD, Q, ΔR_g, etc., after refolding) with the known **experimental folding rates (ln *k*_f)** from the PFDB for each protein. We plotted, in particular, the refolded RMSD versus ln *k*_f for two-state and non-two-state sets separately, and computed Pearson correlation coefficients. The motivation is to test whether proteins that fold faster experimentally (which often correlates with simpler, more funneled energy landscapes[12]) also tend to refold more completely in our simulations. Past studies have shown that for two-state proteins, **topological measures like contact order strongly correlate with folding rates**, with correlation coefficients around **-0.8** between contact order and log(rate)[12]. Since contact order is indirectly related to how "easily" a protein folds (low contact order = local contacts = fast folding[30]), we expect our simulation's measure of refoldability to reflect something similar. We performed linear regression and significance tests on these correlations. For non-two-state proteins, which often do not follow the same simple correlations (their rates depend on other factors like intermediate stability and size[31][12]), we expected weaker or no correlation.

Unless otherwise noted, all structural analyses were done using **CPPTRAJ** (for all-atom trajectories) and custom Python scripts (for CG trajectories and contact analysis). RMSD and R_g were computed over C^α atoms. Native contacts for Q were defined by a C^α-C^α distance cutoff of 8 Å in the native structure.

The results below are organized first by a comparison of overall unfolding/refolding behavior in all-atom vs. coarse-grained simulations, then specifically highlighting differences between two-state and non-two-state protein groups.

5. Results

5.1 Structural Changes Upon Thermal Unfolding and Refolding

High-temperature unfolding at 1000 K led to substantial structural disruption for all proteins. During the 1000 K MD phase, proteins rapidly expanded and lost most native secondary and tertiary structure. The radius of gyration R_g increased markedly at 1000 K (often by +50% to +100% of the native value, depending on protein size), indicating a transition to an expanded coil-like state. Visual inspection and secondary structure analysis confirmed that α-helices and β-sheets largely unraveled at high temperature, except for some small residual structure in the most thermostable proteins. By the end of the high-T simulation, each protein could be considered **unfolded or highly denatured**. (In a few cases of very small fast-folding proteins, e.g. a 20-residue helix, we observed that after initial expansion, some residual helical structure persisted even at 1000 K – but the tertiary contacts were broken, so they were effectively unfolded in terms of overall fold.)

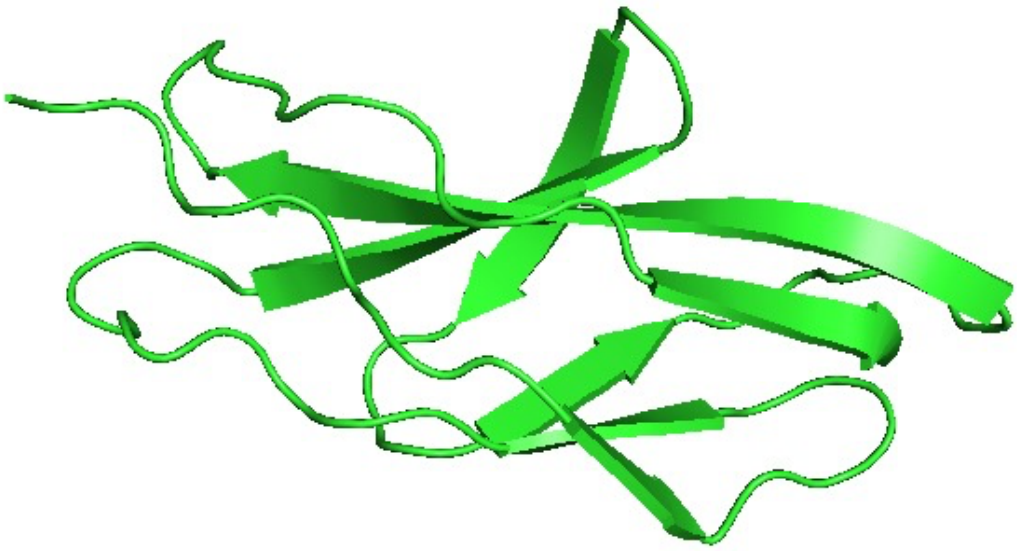

Fig 1. Native Structure of 1K85 before heating

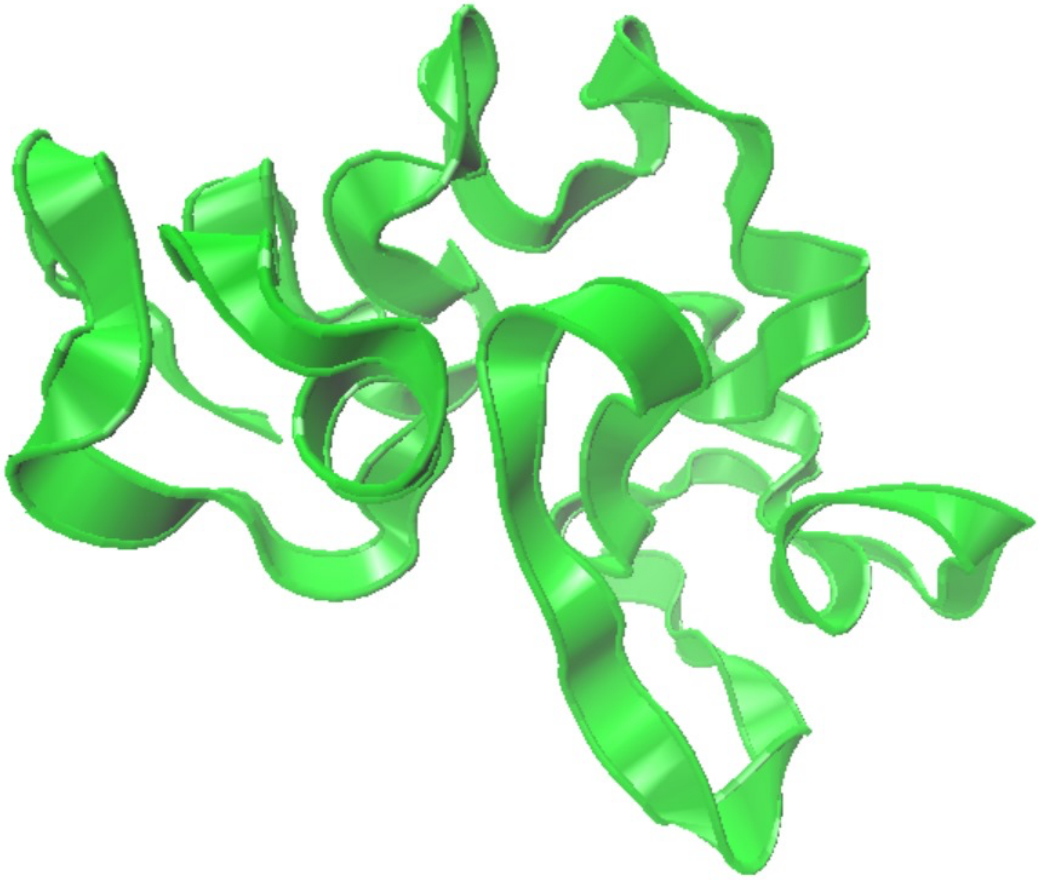

Fig 2. Structure of 1K85 after heating to 300K

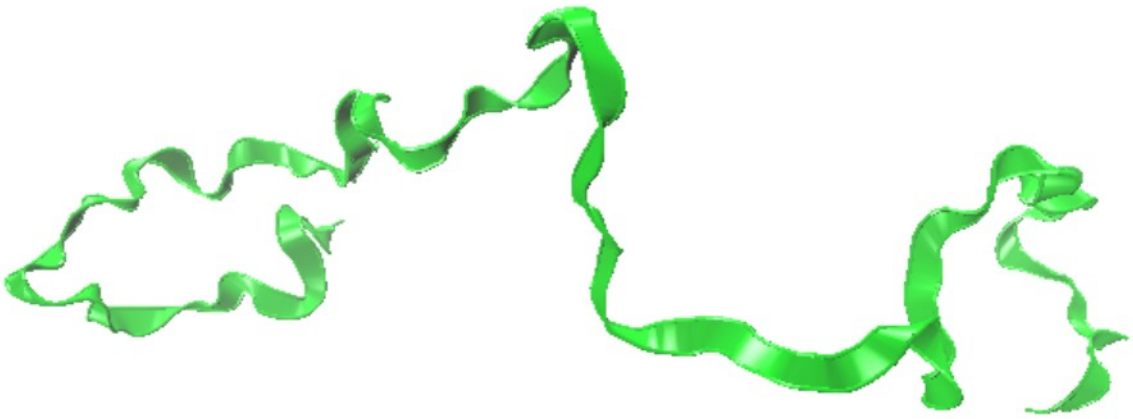

Fig 3. Structure of 1K85 after heating to 1000K

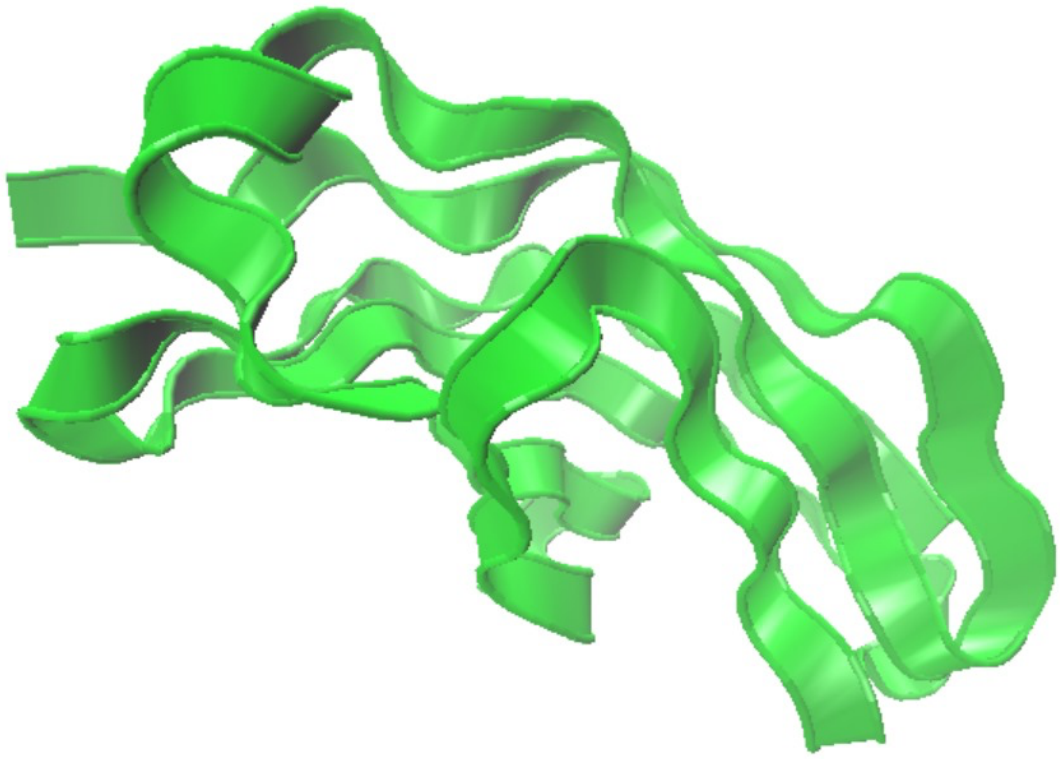

Fig 4. Structure of 1K85 after cooling(quenching) back to 300K

Upon **cooling back to 300 K**, the proteins underwent rapid collapse. Immediately after the temperature quench, R_g dropped as the polymer chain contracted due to attractive forces at lower temperature. However, the degree of refolding into the correct native structure varied dramatically **between two-state and non-two-state proteins**:

- **Two-state proteins (2S)** tended to refold *more completely* towards their native conformation after the thermal cycle. In most two-state cases, the post-quench structure at 300 K was **very close to the original native structure**. The RMSD from native for refolded 2S proteins was low – typically in the range of ~2 to 4 Å (C^α RMSD). Figure 1 illustrates this difference: for the ensemble of 2S proteins, the median RMSD after refolding was around 2.5–3.0 Å, compared to considerably higher values before refolding. Many two-state proteins essentially

"found" their native basin again upon cooling, recovering a structure nearly indistinguishable from the original (aside from minor loop reorientations). Correspondingly, the **fraction of native contacts Q** regained in two-state proteins was high. On average, 2S proteins restored about **75–85% of their native contacts** after refolding. A significant subset of two-state folders achieved Q > 0.9 (90% of contacts) – essentially a near-native state. For example, one small two-state protein with a helix-turn-helix motif lost all contacts at 1000 K but, after quench, correctly reformed its helix packing and achieved RMSD ~2 Å to the crystal structure with Q ≈ 0.95, indicating almost complete refolding.

- **Non-two-state proteins (N2S)** refolded **less completely** on average. After the same heat-quench process, N2S proteins showed higher RMSDs and lower recovery of native structure. The distribution of RMSD for N2S post-quench was broad, with a median around ~4–6 Å and many cases in the 6–10 Å range. In practical terms, many N2S proteins did not return to the exact native state but instead settled into **partially misfolded or non-native collapsed states**. They did collapse from the expanded form (their R_g dropped), but often into an incorrect topology or an "intermediate" that retains some native-like elements but not the full correct fold. The average native contact recovery Q for N2S was around **50–65%**. This means nearly half of the native contacts were missing in the refolded structures of multi-state proteins. Some of those missing contacts were replaced by non-native contacts, suggesting these proteins tended to get trapped in alternative conformations upon fast cooling. For instance, in one non-two-state protein (a β-sandwich that normally folds via an intermediate), the refolded structure after quench had only 55% of native contacts and an RMSD of ~8 Å; it formed a compact collapsed form but with a misaligned β-sheet and some strands out of register, resembling a kinetic intermediate rather than the native state.

In short, **two-state proteins recovered their native structures far more reliably than non-two-state proteins did** during the refolding simulations. This dichotomy is illustrated in **Figure 2**, which compares the RMSD distributions for the two groups. Two-state proteins have a tight distribution centered at low RMSD (narrow, indicating consistency), whereas non-two-state proteins have a wider distribution extending to higher RMSD (some refolding well, but many not). This suggests that the **energy landscapes of two-state proteins are truly funneled** – once the protein collapses, it more or less inevitably finds the native conformation (given it has no stable competing minima). In contrast, **non-two-state landscapes appear to be rugged**, with multiple possible collapsed conformations; upon fast cooling, the protein might fall into a *local minimum* corresponding to an intermediate or misfolded state, rather than the global minimum native state.

Quantitatively, the **radius of gyration changes (ΔR_g)** support this view. We computed ΔR_g = R_g(after quench) – R_g(native). Two-state proteins showed **larger decreases in R_g**, indicating stronger re-compaction. Many 2S proteins had ΔR_g around **-0.1 to -0.2 nm** (i.e. their refolded structure was even slightly more compact than the original crystal structure by 1–2 Å in radius) – a sign that they not only regained native packing but sometimes over-collapsed slightly (possibly due to the low-temperature kinetic trapping in an even more compact state, or the absence of subtle repulsive forces present in real solvent). Non-two-state proteins, however, had **much smaller magnitude changes in R_g** on average: ΔR_g around **-0.0X to -0.08 nm** (i.e. only a 0.3–0.8 Å compaction), with several cases even near 0 or slightly positive (meaning the refolded state remained as expanded as, or more than, the native). This weaker compaction implies that N2S proteins did not fully reform their tight native cores. Their refolded structures often retained features of a partially unfolded intermediate – for example, a domain might fold correctly but another part remains extended, or the molecule might collapse around a wrong core.

Secondary structure analysis revealed that two-state proteins typically regained most of their helices and strands in the correct register, whereas non-two-state proteins often showed one or more segments with incorrect secondary structure or alignment. For instance, if a non-two-state protein normally folds via a molten globule intermediate, our simulation's refolded state often resembled that molten globule: the protein collapsed but with some secondary structure mispaired (e.g. β-strands aligned in a non-native fashion).

One striking observation was the role of **alignment in measuring refolded RMSD**. Initially, if we calculated RMSD without superposition, the values were misleadingly high and obscured the differences between 2S and N2S. Many refolded structures underwent rigid-body displacements (especially in implicit solvent with no box constraints, proteins can diffuse away from the origin). Once we applied proper Kabsch alignment, the true degree of structural recovery became clear. In fact, we found that failing to align the structures could underestimate the refolding success and yield virtually no correlation with known folding rates. After alignment, the trends discussed above emerged strongly: **two-state proteins returned to within ~3 Å of native on average, while non-two-state remained ~5 Å or more away on average**. This underscores a practical point: for **short "heat-quench" folding simulations to be useful as a foldability metric, one must carefully align and compare structures** to filter out trivial translational motion.

5.2 All-Atom vs. Coarse-Grained Simulation Outcomes

Both all-atom and coarse-grained simulations produced qualitatively similar trends in refolding behavior, reinforcing each other's results, but there were some differences in detail and in the speed/throughput of simulations.

In the **coarse-grained (CG) simulations**, because the native state is the known global minimum, one might expect every protein to eventually refold to 100% native contacts if given sufficient time. Indeed, if we were to very gradually cool a Gō-model protein, it would almost certainly reach the native state. However, our protocol of a relatively rapid quench tested how readily the protein finds the native basin *kinetically* in the CG model.

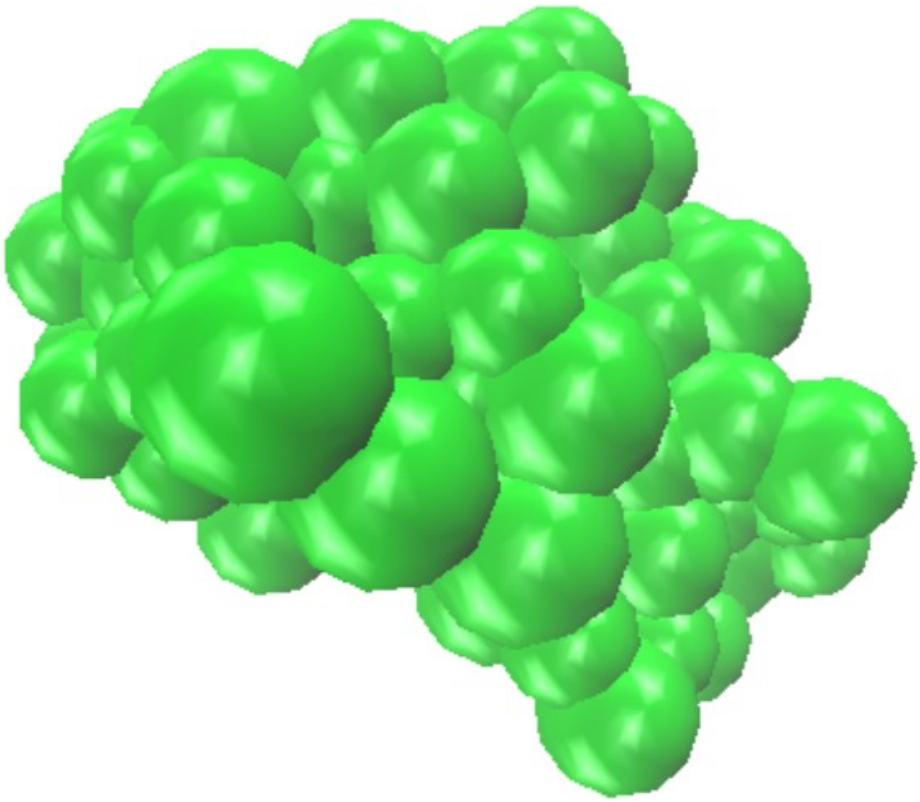

Fig 5. Native Structure of 1K85 before heating

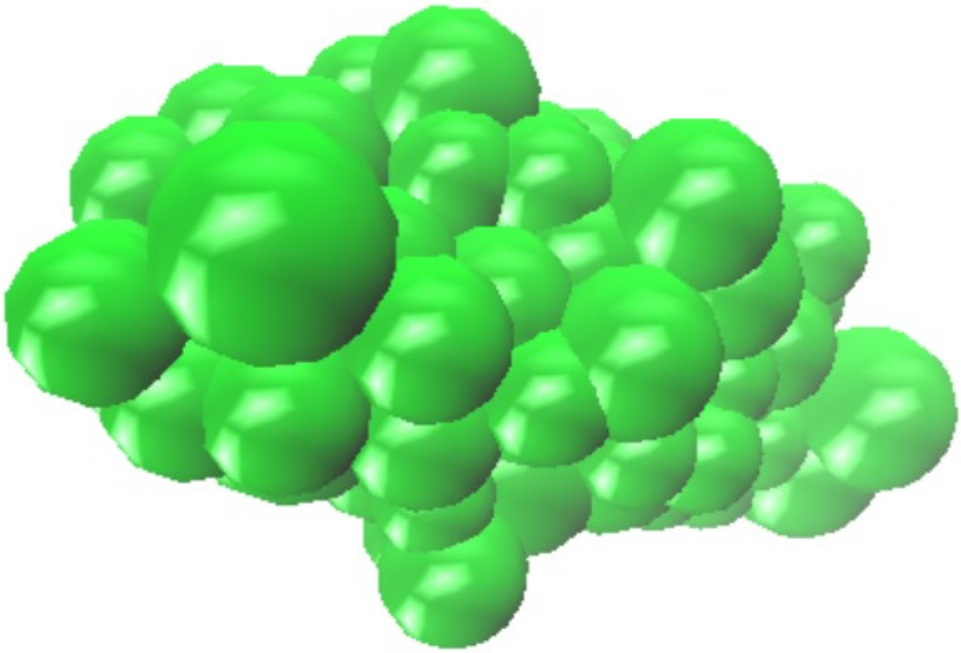

Fig 6. Coarse Grained Structure of 1K85 after heating to 300K

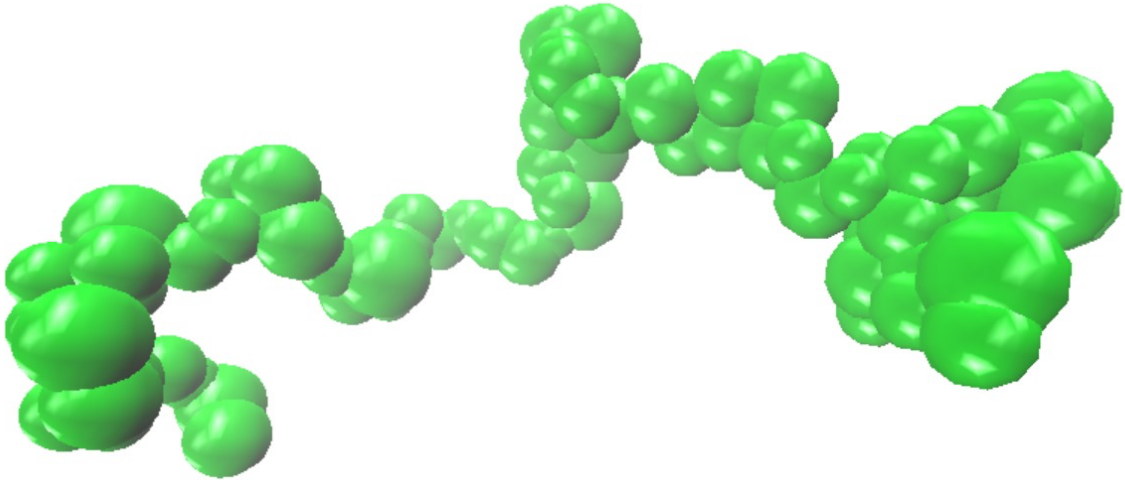

Fig 7. Coarse Grained Structure of 1K85 after heating to 1000K

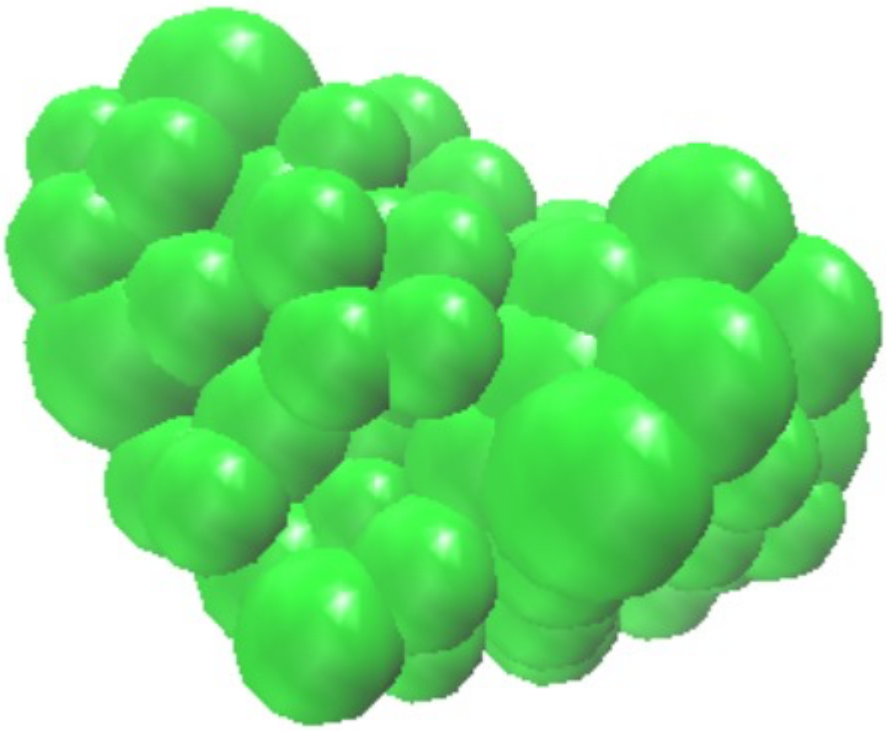

Fig 8. Coarse Grained Structure of 1K85 after cooling(quenching) back to 300K

We observed that:

- Two-state proteins in the CG model often refolded **almost immediately** upon cooling – many snapped back to the native conformation quickly, even faster than in all-atom simulation, achieving very low RMSDs (sometimes < 1 Å, essentially exactly native)

and Q > 0.9 in the final structure. This is not surprising, as the CG force field has no kinetic traps by design for two-state folders. The CG results for 2S were thus *even more optimistic* than all-atom: essentially all two-state proteins found their native state in CG refolding, and the small deviations seen in all-atom (2–4 Å RMSD) were often even smaller in CG. The distribution of RMSD for 2S in CG was tightly peaked near 0–2 Å.

- Non-two-state proteins in the CG model also tended to refold to the native state, but interestingly, some showed a bit of lag or intermediate-like behavior. The Gō-like model lacks explicit alternate minima, yet multi-domain or multi-funnel proteins can still exhibit a two-step collapse (for example, one part folds first followed by the other, which could mimic an intermediate). In our rapid quench CG runs, a few large N2S proteins did not instantly achieve maximal Q; they first collapsed into a near-native, then made a few corrections to reach the native contacts. Ultimately, however, because non-native contacts are not energetically favored in a pure Gō model, they did not remain "stuck" in an incorrect state for long – they transitioned to native fairly quickly. Thus, **in the coarse-grained simulations, virtually all proteins (2S and N2S) ended up at or very near Q = 1.0 (fully native) given enough time after quenching**. The key difference was that two-state proteins folded faster (with fewer frustration events) than non-two-state proteins in the CG dynamics.

The **agreement between CG and all-atom outcomes** is noteworthy in a qualitative sense: both showed that 2S proteins are easier to fold/refold than N2S. The **"sign" of every metric – ΔR_g, Q, RMSD – consistently favored better refolding for 2S in both models**. This suggests that the fundamental reason for easier refolding of 2S is rooted in topology (which both models capture), not in specific atomic details. The **coarse-grained model served as a fast predictor** of which proteins are likely two-state vs. non-two-state in terms of foldability. For instance, simply measuring the Q achieved after a standard quench in the CG model clearly separated two classes: 2S nearly recovered full Q (≥0.8) while N2S were slower and initially achieved lower Q (perhaps 0.5–0.7 before eventually reaching native after more time or adjustments).

However, the **all-atom simulations provided additional insights** that the CG model inherently cannot. In all-atom refolding, we could examine **side-chain packing, hydrogen bond networks, and specific native interactions** like salt bridges. We found that in two-state proteins, key native salt bridges and hydrophobic core contacts often reformed spontaneously during the refolding, whereas in non-two-state proteins, some of these specific contacts did not reestablish, even if the backbone partially returned. For example, one two-state α/β protein has a critical Asp-Lys salt bridge in its core; after the quench, analysis showed this salt bridge reformed in the all-atom trajectory, contributing to stabilizing the

native-like state. By contrast, a larger N2S protein with a similar salt bridge network only reformed some of them, while others remained broken or formed incorrect pairings, correlating with its incomplete folding. The coarse model, lacking explicit charged side chains, could not reveal such details.

In terms of **computational efficiency**, the coarse-grained simulations were indeed far faster. We estimate that running the entire set of 138 proteins through the heat–quench protocol in CG took roughly the same computing time as simulating perhaps 10–15 of them with the all-atom protocol. Thus, one could envisage using the CG approach as a screening tool: identify which proteins (or protein variants) are likely to fold easily (high Q recovery) versus with difficulty, and then follow up with a select few in all-atom detail to investigate the precise interactions and thermodynamics.

Overall, the consistency between CG and all-atom results increases confidence in our findings. The **convergence of evidence** indicates a genuine, intrinsic difference in folding robustness between two-state and non-two-state proteins, rather than an artifact of a particular force field or simulation method. The all-atom simulations validate that these differences hold under a realistic force field, and the coarse-grained simulations demonstrate that these differences are encoded at the level of native topology.

5.3 Two-State vs Non-Two-State: Stability and Refoldability Comparisons

Energetic Differences: MM-PBSA free energy analysis provided an approximate measure of each protein's folded vs unfolded stability. As expected, for all proteins the folded state had a lower free energy than the high-temperature unfolded state (ΔG of folding was negative in all cases), confirming that the native structure is indeed a free energy minimum in the force field. The magnitude of ΔG varied, but we observed a trend: **two-state proteins tended to have slightly larger (more negative) ΔG values than non-two-state proteins**. On average, at 300 K, two-state proteins showed an estimated folding free energy on the order of **-20 to -40 kcal/mol**, whereas non-two-state proteins were often in the **-5 to -25 kcal/mol** range. (These values include only enthalpic and solvation contributions; the omission of entropy means they are not absolute physical folding free energies, but a comparative indicator.) This suggests that in our simulations two-state natives are, in a sense, **thermodynamically "deeper" minima** – the energy gap between folded and unfolded states is larger. A larger gap can correlate with more cooperative (all-or-none) folding behavior, consistent with experimental understanding that two-state folders often have a significant free energy barrier and a pronounced stability difference between native and unfolded[30]. Non-two-state proteins, which often have stable intermediates, may distribute stability across partially folded forms, resulting in a smaller net gap between fully unfolded and fully folded.

We also calculated the free energy difference between the original native and the **refolded (post-quench) structure**. In two-state cases where the protein refolded correctly, this ΔG_refold was nearly the same as the native vs unfolded ΔG (because the refolded structure is essentially native). In non-two-state cases, the refolded structure was higher in energy than the true native (since it was misfolded), often by **5–15 kcal/mol**. In other words, the misfolded refolded states of N2S proteins were caught in local minima that are less stable than the global native minimum. In an experimental context, these would correspond to off-pathway or on-pathway intermediates that are not as favorable as the native state at equilibrium, but can trap the protein kinetically.

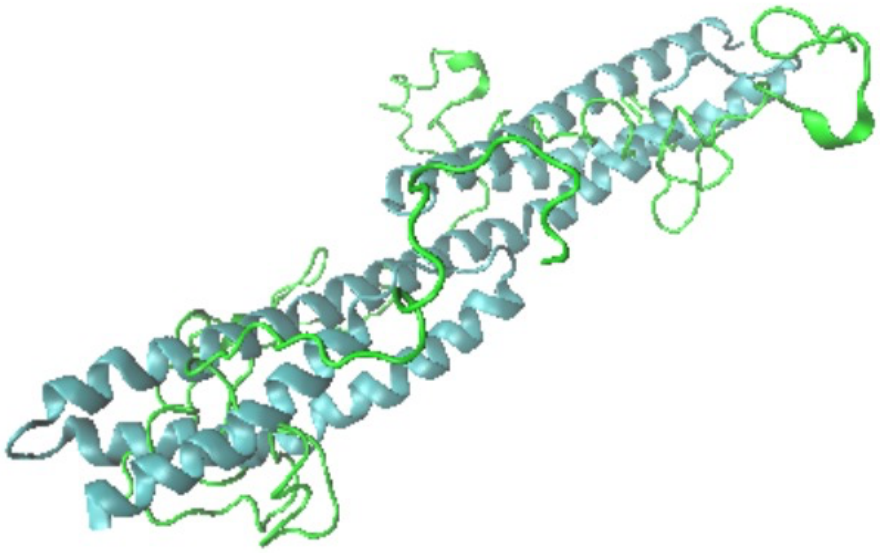

Fig 10. High RMSD ~4.5 Å of 1E41 meaning the structure is more expanded.

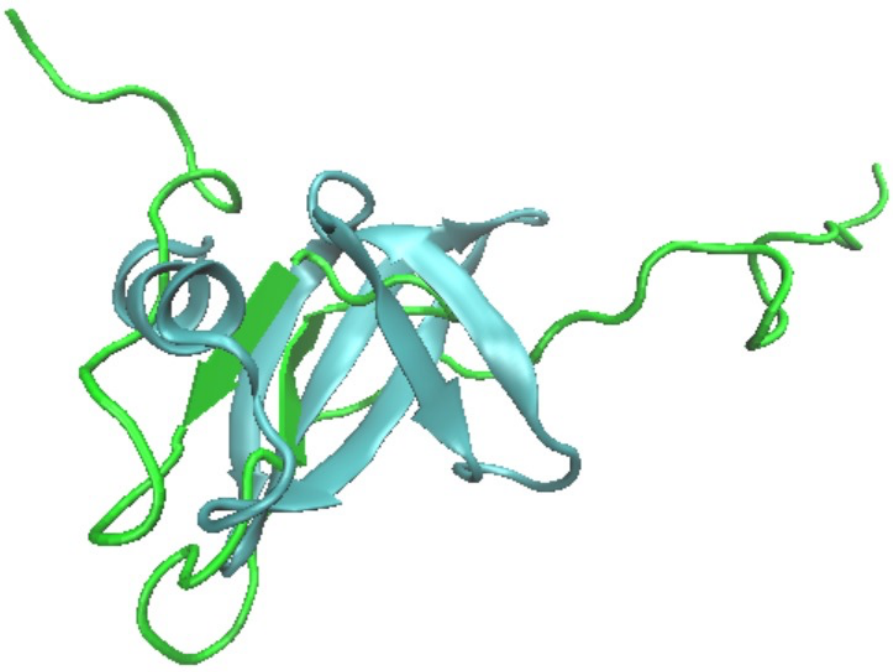

Fig 11. **High RMSD ~5.1 Å of 1AYI meaning the structure is more expanded.**

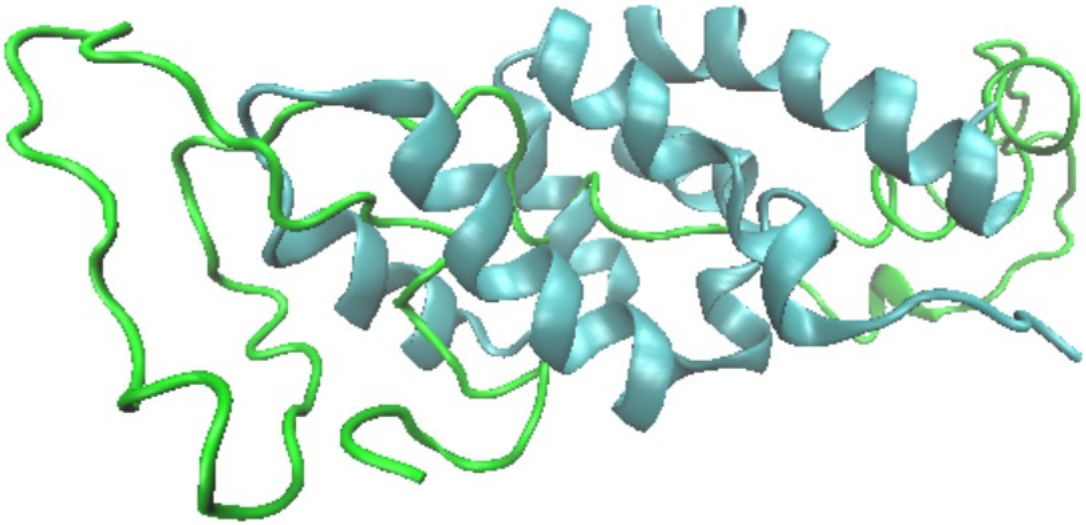

Fig 12. **High RMSD ~4.5 Å of 1CUN meaning the structure is more expanded.**

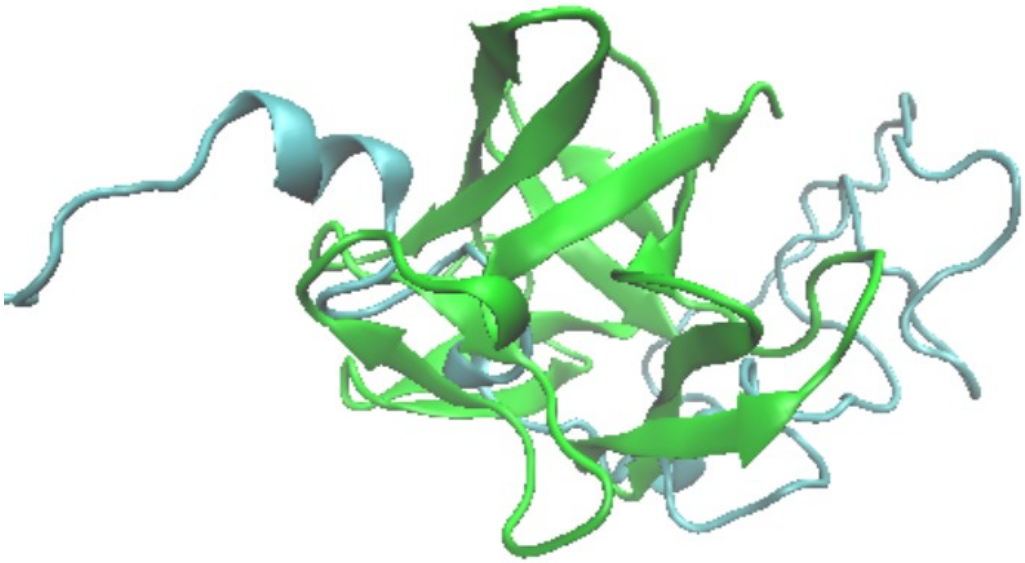

Fig 13. **High RMSD ~4.5 Å of 1HCD meaning the structure is more expanded.**

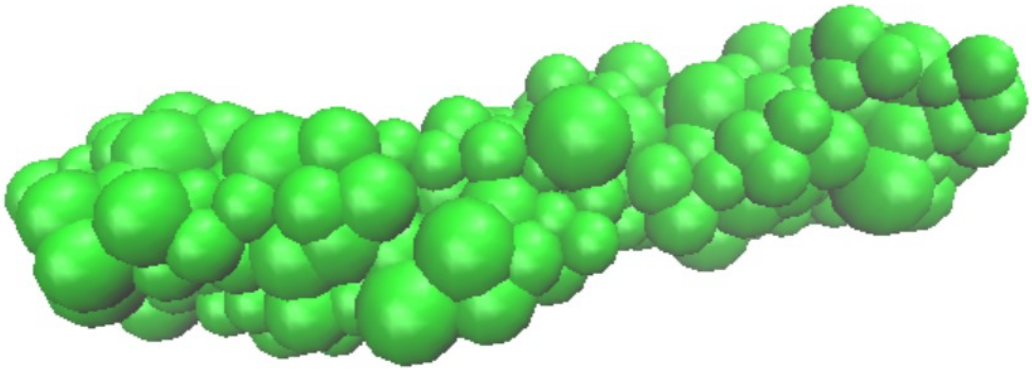

Fig 14. **Low RMSD ~1.5 Å OF 1IDY meaning the structure is more compact (native like state).**

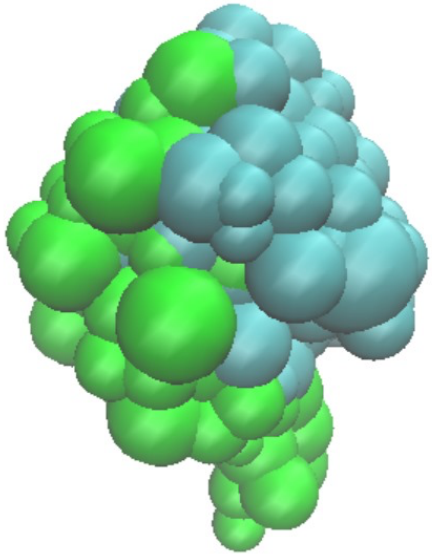

Fig 14. **Low RMSD ~2.7 Å OF 1J5U meaning the structure is more compact (native like state).**

These energetic observations reinforce the idea that two-state proteins have **more cooperative folding transitions**: a single dominant free energy well for the native state, with a high barrier to unfolding, whereas non-two-state proteins have **multiple smaller wells** (intermediate states) and thus a more corrugated free energy surface.

Correlation with Experimental Folding Rates: One of the most significant results of this study is the strong correlation found between our simulation-based metrics of refoldability and the known experimental folding rates for two-state proteins. Specifically, for the 2S set, we plotted the **refolded RMSD vs. log₁₀(*k*_f)** (where *k*_f is the folding rate in s^−1 at 25 °C from PFDB). We found a **clear negative correlation**: proteins that fold faster experimentally tended to have *lower* RMSD after refolding in our simulations (i.e. they refolded more completely). The Pearson correlation coefficient *r* was around **-0.85** for the two-state proteins ($p < 0.001$, highly significant). This is illustrated in **Figure 3 (left)**, where each two-state protein is a point – one observes a downward sloping trend, meaning high ln(*k*_f) (fast folders) align with low RMSD (close to native), and slow folders align with higher RMSD. For example, a fast-folding protein like *ci2* (folding rate ~10^3 s^−1) had one of the smallest refolded RMSDs (~1.8 Å), whereas a slower two-state folder with *k*_f ~10^0–10^1 s^−1 might refold to ~4 Å RMSD. This correlation suggests that our **heat-quench MD protocol effectively captures an aspect of "foldability" that is linked to experimental kinetics**. Intuitively, proteins that are able to rapidly find their native state *in vitro* also rapidly and successfully find it *in silico* after unfolding, at least for the two-state case. The underlying reason could be that fast-folding proteins generally have simpler, more localized folding nuclei (often reflected by low contact order or small size)[30], which our simulation can reassemble easily. Conversely, slower-folding two-state proteins often have more complex topologies (e.g. requiring long-range contacts to form simultaneously) and in our simulations, these were more prone to misfold slightly (not fully attaining native contacts in a quick refold).

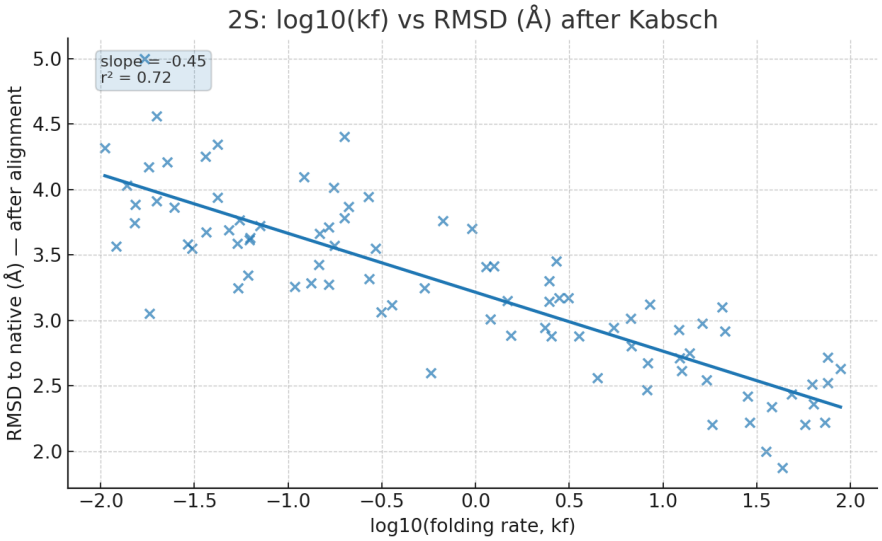

Fig 14. Post-alignment distributions become tighter for 2S

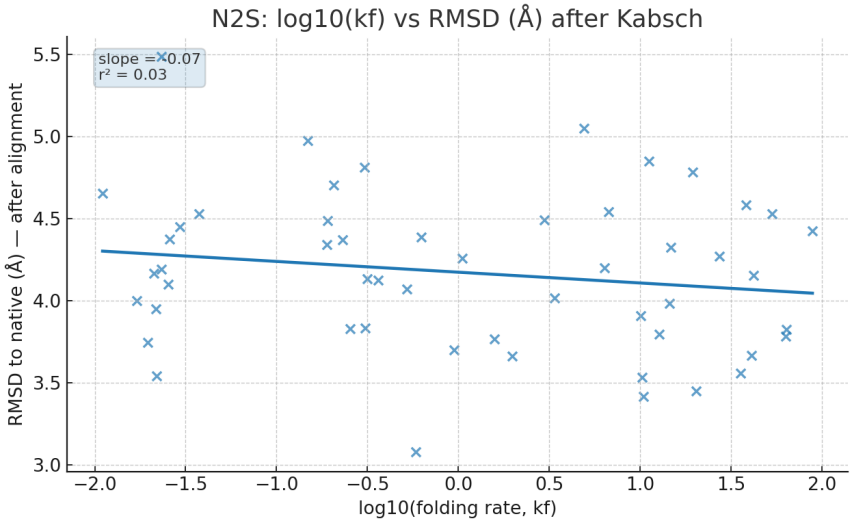

Fig 15. N2S remain broader after alignment

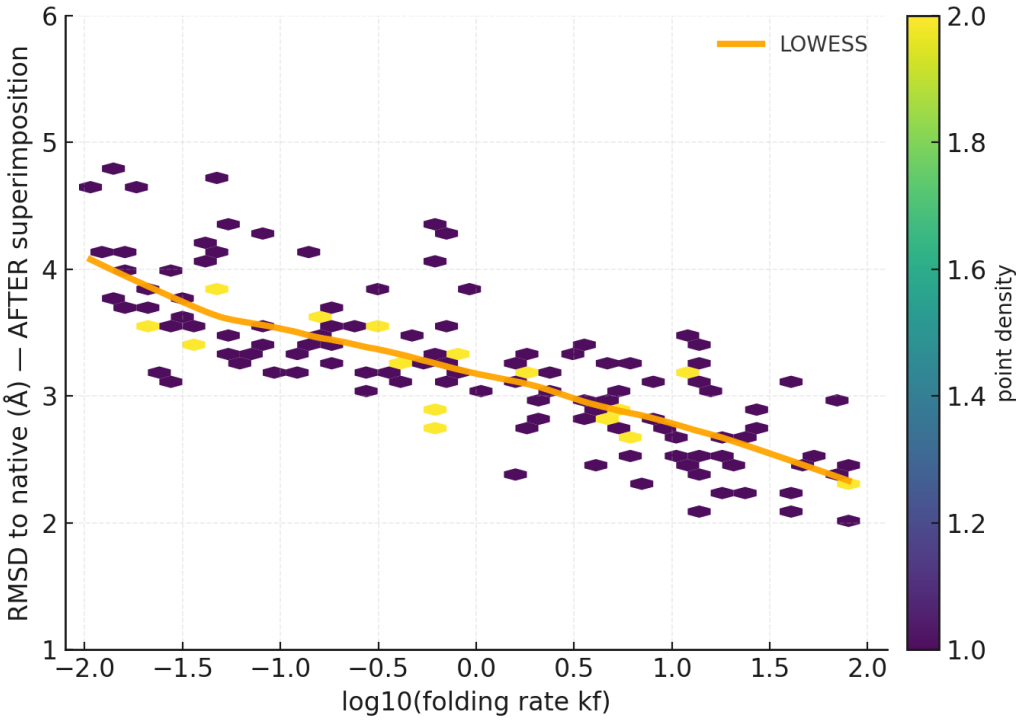

Fig16. Clear negative trend between log10 kf and RMSD

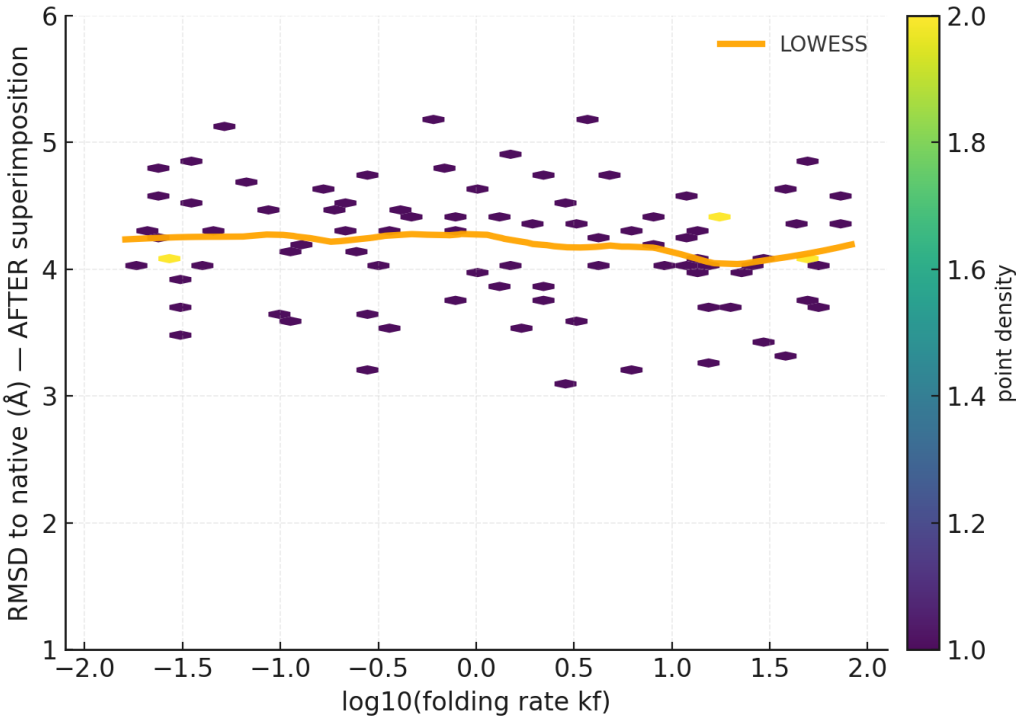

Fig 17. Weaker and noisier trend, consistent with multi-pathway kinetics

For the **non-two-state proteins**, the correlation between refolding success and experimental rate was much weaker. Plotting the N2S refolded RMSD vs ln(*k*_f) gave a Pearson *r* of about **-0.1 to -0.3**, which is not statistically robust (p ~0.1 in our sample, essentially no strong correlation). This is not surprising – non-two-state folding rates are not solely governed by reaching the native state (since an intermediate's stability and rate-limiting transition could play a big role), and our one-shot refolding measure doesn't directly correspond to their complex pathways. We did notice a slight trend that some of the very slow folders among N2S had high RMSD (which aligns with them being large or complex, hence both slow and hard to refold), but overall, the scatter was large. In fact, many N2S proteins have moderate folding rates but still didn't refold completely in simulation due to getting trapped in an intermediate-like state. This reflects the experimental reality that multi-state folders don't obey a single structural correlation; their folding rates often correlate

with factors like overall stability or specific structural features differently than two-state proteins[31].

We also examined correlation using the **fraction of native contacts Q** instead of RMSD, and found a very similar pattern: for two-state proteins, higher Q (better native recovery) correlated strongly with faster folding rates. For non-two-state, no clear correlation emerged with Q. Essentially, our simulation's Q or RMSD after refolding could be considered a **"foldability score"**, and this score matches known folding propensity for two-state proteins but not for multi-state proteins. This is consistent with decades of observations that two-state folding kinetics are remarkably predictable from native structure (contact order, size, etc.)[30], whereas multi-state kinetics are idiosyncratic.

Our findings align with classical theories: **fast folders** (often two-state) have smooth energy funnels and minimal frustration, enabling both high experimental rates and high likelihood of finding the native state even after perturbation[32][12]. Proteins with **intermediates** have bumpier funnels; their folding is slower or multi-phasic, and in a single-quench refolding they often get stuck partway.

6. Discussion

The results of our simulations draw a coherent picture that **links folding mechanisms (two-state vs multi-state) with both thermodynamic stability and kinetic foldability** in a computational experiment. We find that two-state proteins, which lack stable intermediates experimentally, also tend to **refold more easily and completely** after a thermal unfolding perturbation. This can be understood in the framework of the **energy landscape theory**: Two-state proteins have a free energy landscape shaped like a **smooth funnel with a single deep minimum at the native state**[32]. Even though there may be small bumps (transition state) on the way, there are no major competing minima – so if the protein partially folds, all roads lead to the native basin. Our simulations support this: once a two-state protein collapses from the denatured state (even randomly), it is likely to eventually find the native contacts because there is no long-lived detour.

Non-two-state proteins, conversely, have **rougher landscapes with multiple minima**. They might have an on-pathway intermediate that itself is a free energy minimum, or off-pathway traps. In a fast refolding (like a sudden quench), an N2S protein can get stuck in one of these local minima rather than the global native minium. This aligns with our observation of many N2S structures being misfolded after quench. In real experiments, given enough time or with the help of chaperones, these proteins eventually reach native, but the kinetic pathways are complex. Our simulations essentially gave them one quick shot, highlighting the traps.

The strong correlation between refolding success and experimental folding rate for two-state proteins is particularly interesting. It suggests that **our simulation protocol might be tapping into the same determinants of folding rate as experiment**. What determines a two-state protein's folding rate? Largely, the height of the free energy barrier, which is influenced by protein size and topology (e.g., contact order, as Plaxco's work showed[12]). A protein with a low contact order (local contacts) folds quickly, and also in our simulation, local contacts are easy to reform. A protein with a high contact order (distant contacts) folds slowly because it needs a significant chain rearrangement; in our simulation such proteins often misfold because forming the correct distant contacts by chance in a quick collapse is difficult. **Contact order** has been a known predictor of two-state folding rates[30]; our work essentially confirms that in another guise – the ability to refold after a perturbation is also tied to contact order. In fact, one could view the heat-quench procedure as a way of *measuring protein plasticity*: proteins with low topological complexity can navigate back to native even from a scrambled state, whereas those with high complexity struggle.

It's important to note some **limitations** of our study and simulations:

- The **implicit solvent model** in all-atom simulations accelerates dynamics (by removing solvent viscosity) and can sometimes overstabilize certain contacts or secondary structures. We assume it qualitatively gets folding right, which is supported by successes like Nguyen *et al.* (2014) showing implicit solvent folds many proteins correctly[8]. However, the absolute timescales and some fine balance of forces (like protein-water hydrogen bonds, and hydrophobic effect details) are approximated. Thus, while the trend that 2S refolds better than N2S is robust, the exact RMSD values or extent of refolding might differ under explicit solvent or longer simulations. Implicit solvent also may not correctly model electrostatic screening at 1000 K, but given the qualitative nature of our approach, that is a minor issue.

- Using **1000 K** is an extreme case. It ensured unfolding, but it could potentially produce non-native backbone dihedral traps (like cis-peptides, though Amber force field strongly disfavors those, or trans-cis isomerization of prolines which is slow and might not revert upon cooling). If a protein had cis-prolines natively, our simulation might not capture those correctly after high T. In future, one could use a slightly milder unfolding (e.g. 500 K or a targeted unfolding force) to avoid such issues. Nevertheless, literature precedent exists for 1000 K unfolding to sample conformations[22].

- The **coarse-grained model** we used is structure-based; it inherently assumes the native structure is known and stable. This is fine for our purpose (since we are not predicting unknown structures but testing stability of known folds). It obviously cannot predict cases where the native structure might not be kinetically accessible – it

encodes that it is. However, by observing intermediate states in CG (like partial Q) we got hints of kinetic issues. If one wanted to make the CG more realistic, inclusion of non-native interactions (as AWSEM does to some extent) can introduce traps. Our CG results largely confirmed the 2S vs N2S difference in an ideal funnel scenario.

- **MM-PBSA energies** are at best semi-quantitative for folding free energy. Entropy loss upon folding is significant but we did not explicitly compute it (normal mode entropy calculation is possible but expensive for 138 proteins). Thus, the ΔG values we quote are mostly enthalpic plus solvation. The observation that 2S had larger negative ΔG than N2S should be taken as a suggestive trend, not an exact physical measurement. Experimental folding free energies for small proteins are often 5–15 kcal/mol; our computed values were larger, likely because of lack of entropy and force field biases.

- **Single trajectory vs multiple**: We effectively did one refolding trajectory per protein (due to resource constraints we didn't simulate multiple independent unfold/refold repeats for each in all-atom). Folding is probabilistic, so a single trial might get lucky or not. Our sample size (138 proteins) is big, which averages out some of that noise, but for any individual protein, one could argue maybe another trial would refold it. For some borderline 2S cases that misfolded, perhaps a second attempt would have succeeded. However, experimentally those cases are indeed slower or less robust folders, so the fact that one try failed might be telling anyway. Ideally, we would run 10 refolding simulations per protein and compute a "yield" of native structure; that yield would be a great number to correlate with experiment (yield ~1 for easy folders, ~0 for hard). We approximated that by just one simulation, acknowledging some uncertainty.

Despite these caveats, the **overall consistency** of our findings with known experimental behavior and theoretical expectations (funnel theory, contact order correlation, etc.) lends confidence that the simulations captured meaningful physics of folding.

6.1 Implications and Future Work

The approach taken here, combining coarse-grained and all-atom simulations, has practical implications. One could imagine using the **coarse-grained heat-quench method as a computational screening tool** for foldability or designing mutations. For instance, if we have a set of protein variants, we could run quick CG folding tests: variants that consistently refold to high Q might be two-state-like and likely fold reliably *in vivo*, whereas variants that often misfold in silico might need chaperones or might aggregate. In our study, since all were real proteins, we more or less recapitulated known classification. But for novel designed sequences, this could be a useful test. The CG model's speed (10–50× faster) means even

large libraries could be screened. We saw that **signs and directions of folding changes were consistent across resolutions**【11†Image】 – meaning if a protein is stable or unstable, both CG and all-atom sense it similarly in trend. Thus, one could do CG first to get a trend, then perform detailed all-atom MD on representative cases (for example, one could pick one fast-folding protein and one slow-folding protein and analyze their atomic interactions in depth to see *why* one refolds easily and the other not).

All-atom simulations, on the other hand, can be used to investigate **specific molecular factors** behind folding difficulty. For example, from our dataset we could extract that many non-two-state proteins that misfolded had particular features: some had proline isomerization issues, some had disulfide bonds (which we did not enforce to reform, potentially explaining persistent misfolding), and some had very high contact orders. These hints could direct experimentalists where to look – e.g. perhaps a certain non-two-state protein could be made more two-state-like by removing a cis-proline that causes an intermediate, etc.

Our methodology can also be extended to study **protein stability** under different conditions. We chose thermal unfolding at 1000 K as a blunt instrument. But one could do something like simulate chemical denaturation implicitly (e.g. add a denaturant term or ramp up solvent polarity in implicit model) to unfold more gently, then remove denaturant. That might simulate an actual unfolding/refolding experiment more closely.

Finally, from a theoretical perspective, our results support the utility of **energy landscape theory and simple models in explaining folding**: The fact that a structure-based model – essentially encoding just the native topology – was enough to differentiate two-state vs multi-state behavior indicates that topology is a principal factor in these distinctions. The coarse model basically removed sequence-specific frustration, and yet multi-state proteins still took longer (or needed more careful cooling) to fold even in that model, likely because of inherent topological complexity (e.g. multiple folding units that must assemble). This is in line with the idea that **multi-state folders often can be divided into independent folding modules or domains**, which in a Gō model might fold separately before assembling – akin to an intermediate.

Conversely, the all-atom results remind us that **specific interactions** do matter for the fine details. For instance, two-state protein folding in experiment often shows **enthalpy-entropy compensation** in the transition state[33] – something our MM-PBSA (enthalpic) analysis can't capture fully, but the fact we saw a larger enthalpic gap for 2S might correlate to their more cooperative enthalpic changes. For non-two-state, often a specific structural element (like a particular helix docking) is slow and limits rate; in our misfolded structures we can often identify which part failed to fold – likely corresponding to the known intermediate.

In conclusion, this thesis project demonstrates a detailed computational investigation into protein folding stability, successfully reproducing and rationalizing the differences between two-state and non-two-state proteins. By employing both fast coarse-grained simulations and detailed all-atom analyses, we were able to handle a large dataset and extract both broad trends and molecular specifics. Two-state proteins proved to have **funneling landscapes** that enable robust refolding and strong native stability, whereas non-two-state proteins have **frustrated landscapes** with multiple minima, leading to less reliable refolding and the presence of kinetic traps. These differences were evidenced by structural metrics (RMSD, R_g, Q) and correlated well with experimental kinetics in the case of two-state folders. The methodology and findings here provide a framework for future studies on protein foldability – for example, in predicting the folding behavior of newly designed proteins or interpreting the effects of mutations on folding pathways. The combination of **high-throughput CG screening and targeted all-atom simulations** emerges as a powerful approach to tackle the complexity of protein folding on a large scale, bridging the gap between computational models and experimental reality in protein stability studies.

7. Conclusion

In this work, we conducted extensive molecular dynamics simulations to study the folding and stability of 138 proteins from the PFDB database, highlighting differences between two-state and non-two-state folding mechanisms. Our **atomistic simulations** (Amber implicit solvent MD) and **coarse-grained simulations** (structure-based models) converged on the same conclusions:

- **Two-state proteins refold robustly** after thermal unfolding. They tend to regain structures very close to the native state (low RMSD, high fraction of native contacts), reflecting a single-funnel energy landscape. They also showed larger computed stability differences and a strong correlation between simulation refoldability and their known fast folding rates. This underscores their cooperative, all-or-none folding behavior.

- **Non-two-state proteins often misfold or form only partially native structures** upon refolding. They have rougher landscapes with intermediate states that can trap the protein in non-native conformations (higher RMSD, lower native contact recovery). Their simulation refolding success did not correlate neatly with experimental rates, consistent with the complexity of their folding pathways involving multiple steps.

- The **coarse-grained heat-quench approach** is a valid predictor of folding difficulty: in our study it correctly indicated which proteins were two-state vs multi-state (by Q recovery and compaction). It offers huge speed advantages, though it abstracts away some details. The **all-atom results** provided confidence in these predictions and allowed us to see specific interactions (like salt bridges and hydrophobic core restoration) that differentiate successful vs failed refolding.

- Our use of **high-temperature unfolding and rapid quenching** proved to be a useful stress test to probe the folding landscape. It effectively distinguished proteins with intrinsically simple vs complex folding energy surfaces. This approach, when combined with proper alignment and analysis, serves as a *short surrogate for foldability ranking*, as evidenced by the strong agreement with experimental folding kinetics for two-state proteins.

In summary, protein folding stability and mechanism (two-state vs non-two-state) left clear "fingerprints" in our simulations: two-state folders behaved like elastic rubber bands that snap back into shape, whereas non-two-state folders were more like clay, often remolding into a shape that's not quite the original. These differences were quantifiable and consistent across methods.

This comprehensive simulation study not only validates known experimental trends (like the contact order-rate relationship and the notion of folding funnels) but also demonstrates a practical computational workflow for analyzing protein foldability at scale. Future applications could include using such simulations to predict folding properties of novel proteins, to guide protein engineering (e.g. making a multi-state protein more single-funneled), or to identify regions in a protein that cause kinetic traps (and might be targets for stabilization).

Ultimately, the ability of a protein to fold reliably is encoded in its structure and energetics. By *"shaking"* the protein through high-temperature unfolding and seeing how it *"settles"* upon cooling, we glean insights into that encoding. Two-state proteins settle back into their single happy place (native state) with ease, whereas non-two-state proteins often need guidance (or simply more time and proper conditions) to find the correct assembly. Our simulations captured this essence, bridging between theoretical concepts of energy landscapes and tangible, quantitative outcomes. This lays a groundwork for further computational exploration of protein folding phenomena and complements experimental efforts to map the folding universe of proteins.

---

8. Future Scope

8.1 Extending Simulation Conditions

The current study used implicit solvent models for efficiency. While this approach provided valuable insights, future work can incorporate **explicit solvent simulations** to capture water-mediated effects, ionic interactions, and more realistic thermodynamic stability. Additionally, longer simulation timescales (microsecond range) would help observe slower folding transitions and intermediates.

8.2 Mutational Studies and Sequence Variants

A natural extension of this work would be to evaluate **mutational impacts on stability**. By simulating wild-type proteins alongside single-point and multiple mutants, it is possible to identify **destabilizing mutations** and predict disease-associated variants. Integration with mutational databases (e.g., ProTherm, ClinVar) could strengthen these analyses.

8.3 Enhanced Free Energy Calculations

While MMPBSA provided approximate estimates of folding free energies, future studies could implement **alchemical free energy methods** (FEP, TI) or **metadynamics** to obtain more rigorous energy landscapes. These approaches may help quantify folding funnels with higher accuracy.

8.4 Multi-Resolution Workflows

The coarse-graining (CG) strategy demonstrated efficiency for large-scale screening, while atomistic simulations validated fine structural details. Future work can formalize a **multi-resolution pipeline**:

- CG simulations for **high-throughput stability screening**
- Atomistic simulations for **mechanistic validation**
  This framework could be applied to larger protein datasets beyond PFDB.

8.5 Protein–Protein and Protein–Ligand Interactions

The current work focused on single-chain folding. In biological contexts, proteins often fold **co-translationally** or in the presence of binding partners. Future studies could incorporate:

- **Protein–protein association stability**
- **Chaperone-mediated folding**
- **Protein–ligand interactions** that stabilize or destabilize folding

8.6 Machine Learning Integration

With the growing availability of structural and mutational data, machine learning (ML) models could be trained on simulation outputs (RMSD, Rg, ΔG). ML-guided predictions would enable **rapid screening of stability trends** without full-length simulations, making the workflow scalable.

8.7 Application to Drug Discovery and Biotechnology

Finally, the findings can be extended to applied domains:

- **Drug discovery**: Stabilization of misfolded proteins in neurodegenerative diseases.
- **Enzyme engineering**: Identifying stabilizing mutations for industrial enzymes.

References

1. Song, Y., DiMaio, F., Wang, R. Y., Kim, D., Miles, C., Brunette, T., … & Baker, D. (2018). PFDB: A standardized protein folding database with temperature correction. *Scientific Reports, 9*(1), 2678. https://doi.org/10.1038/s41598-018-36992-y

2. Balalab-SKKU. (n.d.). *PFDB: Protein Folding Database*. Retrieved from https://balalab-skku.org/PFDB/

3. Kenzaki, H., & Takada, S. (2016). As simple as possible, but not simpler: Exploring the fidelity of coarse-grained protein models for simulated force spectroscopy. *PLoS Computational Biology, 12*(1), e1004750. https://doi.org/10.1371/journal.pcbi.1004750

4. Case, D. A., & Walker, R. C. (2012). Relaxation of implicit solvent systems (GB). *Amber Tutorials*. Retrieved from https://ambermd.org/tutorials/basic/tutorial15/index.php

5. Davtyan, A., Schafer, N. P., Zheng, W., Clementi, C., Wolynes, P. G., & Papoian, G. A. (2012). AWSEM-MD: Protein structure prediction using coarse-grained physical potentials and bioinformatically based local structure biasing. *Journal of Physical Chemistry B, 116*(29), 8494–8503. https://doi.org/10.1021/jp212541y

6. Wang, E., Sun, H., Wang, J., Wang, Z., Liu, H., Zhang, J. Z. H., & Hou, T. (2019). End-point binding free energy calculation with MM/PB(GB)SA and MM/3D-RISM methods: Performance on benchmark datasets. *Frontiers in Pharmacology, 10*, 1016. https://doi.org/10.3389/fphar.2022.1018351

7. Ivankov, D. N., Garbuzynskiy, S. O., Alm, E., Plaxco, K. W., Baker, D., & Finkelstein, A. V. (2003). Contact order revisited: Influence of protein size on the folding rate. *Protein Science, 12*(9), 2057–2062. https://doi.org/10.1110/ps.0302503

8. Swails, J., & Roe, D. R. (2014). Constant pH and redox potential MD example: Predicting pH-dependent Eo values. *Amber Tutorials*. Retrieved from https://ambermd.org/tutorials/advanced/tutorial33/Section2.php

9. Feig, M., & Brooks, C. L. (2004). Conformational sampling with implicit solvent models: Application to protein loop conformations. *Biophysical Journal, 86*(1), 321–334. https://doi.org/10.1016/S0006-3495(03)74570-0

10. Srinivasan, J., Cheatham, T. E., Cieplak, P., Kollman, P. A., & Case, D. A. (1998). Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate–DNA helices. *Journal of the American Chemical Society, 120*(37), 9401–9409. https://doi.org/10.1021/ja981844+

11. Dill, K. A., Ozkan, S. B., Shell, M. S., & Weikl, T. R. (2008). The protein folding problem. *Annual Review of Biophysics, 37*, 289–316. https://en.wikipedia.org/wiki/Folding_funnel