Comprehensive Kinetic and Free Energy Analysis of NTL9 Folding via Systematic Collective Variable Selection and Markov State Models

AstroPilot
Anthropic, Gemini & OpenAI servers. Planet Earth.

Understanding the complex pathways and kinetics of protein folding from molecular dynamics simulations requires sophisticated analytical tools. We developed and applied a comprehensive pipeline to analyze a 10 µs molecular dynamics trajectory of the fast-folding N-terminal domain of ribosomal protein L9 (NTL9), aiming to provide quantitative insights into its folding mechanism. Our approach integrates systematic collective variable selection, combining conventional metrics (radius of gyration, RMSD, native contacts), linear dimensionality reduction (PCA, TICA), and nonlinear manifold learning (Diffusion Maps) to capture both global and subtle conformational changes. Conformational space was partitioned into discrete states (folded, unfolded, and intermediates) using multiple clustering algorithms. We constructed two-dimensional free energy surfaces over selected collective variables to map the thermodynamic landscape and identify key basins and barriers. Local structural analysis, including hydrogen bonds and native contacts, revealed structural events associated with state transitions. Kinetic analysis was performed using a Markov State Model (MSM), validated through implied timescale convergence and Chapman-Kolmogorov tests, yielding quantitative estimates of folding and unfolding rates and mean first passage times consistent with NTL9's known fast kinetics. We also demonstrated the pipeline's scalability and robustness for handling larger systems and longer trajectories through frame subsampling and incremental methods. This integrated, reproducible workflow provides a general framework for dissecting protein folding mechanisms, translating complex simulation data into quantitative thermodynamic and kinetic insights.

I. INTRODUCTION

Understanding how a linear sequence of amino acids folds into a specific, functional three-dimensional structure is one of the central challenges in molecular biophysics. This process, known as protein folding, dictates protein function and is intimately linked to cellular processes; conversely, misfolding can lead to a range of debilitating diseases. Experimental techniques provide crucial insights into protein structures and folding kinetics, but they often offer an ensemble-averaged view or capture only stable states. Molecular dynamics (MD) simulations, on the other hand, provide an atomic-resolution window into the dynamic process of folding, allowing us to potentially observe the complex pathways and transient intermediates involved.However, analyzing the vast amounts of data generated by MD simulations to extract quantitative thermodynamic and kinetic information about protein folding presents significant analytical hurdles. The conformational space accessible to a protein is extremely high-dimensional, making direct analysis intractable. The folding process itself occurs over a rugged free energy landscape characterized by multiple local minima (representing stable or metastable states) separated by high energy barriers. Sampling the rare transitions over these barriers within computationally feasible simulation times is challenging, and identifying the crucial, slow degrees of freedom that govern these transitions – the collective variables (CVs) – is often non-trivial. Traditional analysis methods based on a few predefined structural metrics may fail to capture the full complexity of the conformational changes relevant to folding and unfolding. Extracting reliable kinetic rates and pathways from simulation data that inherently suffers from limited sampling of transition states requires robust statistical frameworks.

To address these challenges and provide a comprehensive, quantitative description of protein folding from MD simulations, we have developed and applied a systematic computational pipeline. Our approach focuses on the fast-folding N-terminal domain of ribosomal protein L9 (NTL9) as a model system, leveraging a 10μs MD trajectory to dissect its folding mechanism. A key innovation of our methodology lies in the systematic selection of relevant collective variables. We integrate conventional structural metrics such as radius of gyration, RMSD, and native contacts with advanced dimensionality reduction techniques, including linear methods like Principal Component Analysis (PCA) and Time-lagged Independent Component Analysis (TICA), and nonlinear manifold learning techniques such as Diffusion Maps. This multi-faceted approach allows us to identify and utilize CVs that capture both large-scale global structural changes and subtle, kinetically important motions that are slow to decorrelate.

Building upon the insights from the systematically selected CVs, we partition the high-dimensional conformational space into a discrete set of states, corresponding to folded, unfolded, and various intermediate ensembles, using multiple clustering algorithms to assess robustness. We then construct two-dimensional free energy surfaces projected onto the most informative CVs to visualize the thermodynamic landscape, pinpointing stable basins and identifying the free energy barriers separating them. To move beyond the static thermodynamic picture and quantify the dynamics, we employ Markov State Models (MSMs). MSMs provide a powerful statistical framework to model the dynamics as

transitions between the defined discrete states, allowing us to rigorously calculate kinetic properties such as folding and unfolding rates, mean first passage times, and to identify the dominant transition pathways. The reliability and self-consistency of our kinetic analysis are rigorously validated using standard procedures, including checking for the convergence of implied timescales as a function of lag time and performing Chapman-Kolmogorov tests. Furthermore, we demonstrate the scalability and robustness of our pipeline for handling larger datasets and potentially more complex systems through techniques like frame subsampling and incremental methods. This integrated and reproducible workflow provides a powerful, general framework for translating complex molecular simulation data into quantitative thermodynamic and kinetic insights into protein folding mechanisms.

II. METHODS

Our computational pipeline for analyzing protein folding simulations integrates multiple steps, from initial data processing and systematic collective variable selection to kinetic modeling using Markov State Models. This comprehensive approach was applied to a molecular dynamics trajectory of the N-terminal domain of ribosomal protein L9 (NTL9) to dissect its folding mechanism.

A. Molecular Dynamics Trajectory and Data Preparation

The analysis was performed on a 10 μs molecular dynamics simulation trajectory of NTL9. The trajectory comprises 5000 frames saved at 2 ns intervals, representing the time evolution of the protein's conformation. The folded-state topology of NTL9, typically obtained from an experimental structure (e.g., PDB ID), served as a reference. Data loading and initial processing were handled using standard molecular dynamics analysis libraries such as MDTraj or MDAnalysis. For consistency across analyses focusing on the protein's internal dynamics, all trajectory frames were structurally aligned to the reference folded structure using the backbone heavy atoms (N, Cα, C, O). This alignment removes global translational and rotational motions.

While the full 5000 frames were used for the primary analysis, the pipeline includes an optional frame subsampling step (e.g., analyzing every 2 or 5 frames) to demonstrate scalability and manage computational resources for potentially longer trajectories, with checks performed to ensure that kinetic fidelity is maintained at reduced sampling rates. Unless specified otherwise, analyses primarily utilized backbone heavy atoms, with sidechain atoms included for specific local structural analyses like hydrogen bonds.

B. Collective Variable Selection and Dimensionality Reduction

Identifying relevant collective variables (CVs) that capture the essential, slow degrees of freedom governing protein folding is crucial for overcoming the high dimensionality of conformational space, as highlighted in the Introduction [1–3]. Our approach systematically combines conventional structural metrics with advanced dimensionality reduction techniques [2, 3].

1. Conventional Collective Variables

Standard structural metrics were computed for each frame of the trajectory. These included the radius of gyration (Rg), which measures the overall compactness of the protein; the root mean square deviation (RMSD) of backbone atoms to the folded reference structure, quantifying similarity to the native state; and the fraction of native contacts (Q). Native contacts were defined based on heavy atom pairs in the folded reference structure that are within a specified cutoff distance, typically 0.45 nm. A contact was considered formed in a given trajectory frame if the distance between the corresponding heavy atoms was below this cutoff.

2. Linear Dimensionality Reduction

Linear methods were employed to find directions in the high-dimensional conformational space that capture significant variance or slow dynamics [4].

Principal Component Analysis (PCA) was applied to both the raw Cartesian coordinates of backbone atoms and the pairwise distances between Cα atoms. PCA identifies orthogonal linear combinations of the input features (principal

components, PCs) ordered by the amount of variance they explain [5, 6]. The number of PCs retained for further analysis was determined by examining the cumulative variance explained (typically retaining components explaining $\geq 80\%$) and by identifying the 'elbow' or inflection point in the scree plot of eigenvalues [6].

Time-lagged Independent Component Analysis (TICA) was used to identify kinetically relevant collective variables [7]. TICA finds linear combinations of input features (independent components, ICs) that have the slowest autocorrelation decay, thus capturing the slowest processes in the system [7].

TICA was applied to both backbone Cartesian coordinates and Cα pairwise distances.

The choice of the lag time for TICA was optimized by selecting a value where the implied timescales of the slowest components converged, indicating that the system's dynamics can be well-represented as a Markovian process at this timescale [7]. The optimal lag time was determined by analyzing the convergence of the implied timescales of the MSM [8]. Implied timescales (τ_i) are characteristic decay times of the relaxation modes of the system, derived from the eigenvalues of the transition matrix [8, 9]. Plotting implied timescales as a function of lag time reveals a plateau beyond a certain lag time, indicating that the slowest processes (corresponding to the largest implied timescales) are well-described as Markovian transitions between the defined states at this lag time [8]. The lag time corresponding to this plateau was selected for the final MSM construction [8].

3. Nonlinear Manifold Learning

Complementary to linear methods, nonlinear dimensionality reduction techniques were used to uncover potentially complex, non-linear relationships in the conformational landscape. Diffusion Maps were constructed based on the pairwise distances between Cα atoms. Diffusion maps embed data points into a low-dimensional space based on their connectivity in a graph where edge weights reflect similarity (typically using a Gaussian kernel). The kernel bandwidth (σ) for the Gaussian kernel was chosen using the median heuristic, where σ is set to the median of all pairwise distances. The number of diffusion coordinates (DCs) to retain was guided by the eigengap criterion (looking for a significant drop between consecutive eigenvalues) and further validated by visual inspection of the embedding and the amount of variance captured.

4. Collective Variable Comparison and Selection

All computed CVs (conventional, PCs, TICs, DCs) were evaluated and compared based on their ability to separate distinct conformational states (e.g., folded, unfolded) and their physical interpretability. Trajectory projections onto pairs of prominent CVs were visualized to assess the structure of the conformational space. The top two or three CVs that demonstrated the most significant separation of states and captured the seemingly slowest or most relevant conformational changes were selected for subsequent clustering and free energy surface construction [10]. This systematic comparison ensures that the chosen CVs effectively represent the key transitions relevant to folding.

C. State Partitioning and Clustering

To discretize the high-dimensional conformational space and define metastable states (folded, unfolded, intermediates), clustering was performed in the space of the selected CVs. Multiple clustering algorithms were employed to assess the robustness of the resulting state definitions.

k-means clustering was applied to partition the data points into a predefined number of clusters based on minimizing the within-cluster sum of squares [11–13]. Gaussian Mixture Models (GMM) were used to model the data as a mixture of multiple Gaussian distributions, assigning points probabilistically to clusters. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) was utilized to identify clusters based on density, allowing for the detection of arbitrarily shaped clusters and the identification of outliers.

The optimal number of clusters for k-means and GMM was determined using internal validation metrics such as the silhouette score (quantifying how similar a data point is to its own cluster compared to other clusters) and the Bayesian Information Criterion (BIC), which balances model fit with complexity. For DBSCAN, parameters (ϵ, the maximum distance between two samples for one to be considered as in the neighborhood of the other, and min_samples, the number of samples in a neighborhood for a point to be considered as a core point) were chosen based on visual inspection of the clustering results and density plots of the data [14].

Each trajectory frame was assigned to its corresponding cluster. The resulting clusters were then characterized structurally (e.g., by calculating the average RMSD and Q for frames within each cluster) and kinetically (by examining transitions between clusters). Clusters were labeled as "folded," "unfolded," or "intermediate" based on these

characteristics, with the folded state typically corresponding to low RMSD and high Q values. The robustness of the state definitions was further assessed by comparing the state assignments obtained from different clustering algorithms and using different combinations of selected CVs.

D. Free Energy Surface and Local Structural Analysis

To visualize the thermodynamic landscape governing NTL9 folding, two-dimensional free energy surfaces (FES) were constructed. These surfaces were projected onto the two most informative collective variables identified during the CV selection step. The FES was calculated by first binning the trajectory data points in the 2D space defined by the selected CVs using a grid size typically ranging from 50x50 to 100x100 bins. The probability density (P) for each bin was estimated from the normalized histogram of frame counts. The free energy (F) of each bin was then computed using the Boltzmann relation: $F = -k_B T \ln P$, where k_B is the Boltzmann constant and T is the temperature of the simulation (assumed constant). A constant offset was added to set the minimum free energy to zero.

The FES visually maps the basins corresponding to stable or metastable states and the barriers separating them.

To gain detailed insights into the structural events associated with state transitions, key hydrogen bonds and native contacts were monitored throughout the trajectory. Hydrogen bonds were identified using standard geometric criteria, typically involving a distance cutoff between donor and acceptor heavy atoms (e.g., 0.35 nm) and an angle cutoff for the donor-hydrogen-acceptor angle (e.g., 120°). Specific native contacts identified as potentially important from the folded structure or visual inspection were also monitored based on their heavy atom distances relative to the 0.45 nm cutoff. The formation and breaking of these specific contacts and hydrogen bonds were correlated with the positions of frames on the FES and their assignments to different clusters, helping to identify the structural changes underlying transitions between folded, unfolded, and intermediate states.

E. Markov State Model Construction and Validation

Markov State Models (MSMs) were constructed to provide a quantitative kinetic framework for analyzing the transitions between the defined discrete states [15]. An MSM models the dynamics as a memoryless process where the probability of transitioning to the next state depends only on the current state.

MSMs were built from the trajectory data using the state assignments obtained from clustering [15, 16]. A transition probability matrix (T) was estimated for a given lag time (τ), where T_{ij} represents the probability of transitioning from state i to state j in time τ [15, 16]. The transition matrix was estimated by counting the number of transitions between states i and j observed in the trajectory over the lag time τ and normalizing by the total number of observations in state i [15, 16]. This estimation was performed using standard libraries such as scikit-learn, NumPy, or msmtools.

The choice of the lag time (τ) is critical for MSM validity. It must be long enough for the system to lose memory of its previous state within each state (i.e., for transitions between states to be Markovian), but short enough to sample sufficient transitions [8, 9].

The optimal lag time was determined by analyzing the convergence of the implied timescales of the MSM [8]. Implied timescales (τ_i) are characteristic decay times of the relaxation modes of the system, derived from the eigenvalues of the transition matrix [8, 9].

Plotting implied timescales as a function of lag time reveals a plateau beyond a certain lag time, indicating that the slowest processes (corresponding to the largest implied timescales) are well-described as Markovian transitions between the defined states at this lag time [8]. The lag time corresponding to this plateau was selected for the final MSM construction [8].

The self-consistency and validity of the constructed MSM were further assessed using Chapman-Kolmogorov tests [17]. These tests compare the transition probabilities predicted by the MSM over multiple lag times ($k\tau$) with those directly observed from the trajectory over the same cumulative time interval ($k\tau$) [17]. Agreement between predicted and observed probabilities validates the Markovian assumption and the chosen state decomposition [17].

From the validated MSM, quantitative kinetic properties were extracted. These included folding and unfolding rates (calculated from transition probabilities between folded and unfolded macrostates, which are defined as aggregates of the microstates/clusters) [18, 19], mean first passage times (MFPTs) between any two sets of states (e.g., from unfolded states to folded states) [18], and the identification of dominant transition pathways by analyzing the flux between states [18, 19].

F. Scalability and Automation

To ensure the applicability of the pipeline to larger systems and longer trajectories, several strategies for scalability and automation were incorporated. As mentioned previously, frame subsampling allows for reducing the data size. For very large datasets, incremental PCA and TICA algorithms were considered, which process data in chunks without loading the entire trajectory into memory. While NTL9 is a single-domain protein, the framework is extensible to multi-domain proteins by incorporating domain-wise analyses or graph-based representations of the protein structure.Parameter selection steps, such as determining the embedding dimensionality for dimensionality reduction techniques (based on variance explained or eigengap) and selecting the optimal number of clusters (based on silhouette scores, gap statistics, or BIC), were automated where possible. This automation reduces manual intervention and enhances reproducibility.

G. Sensitivity and Robustness Analysis

To assess the robustness of the results to parameter choices, sensitivity analyses were performed [20–22]. This involved repeating key steps of the pipeline with different parameter values, such as varying the number of clusters in k-means/GMM, adjusting the ϵ parameter in DBSCAN, or exploring a range of lag times for MSM construction. The stability of the resulting state definitions, free energy surfaces, and calculated kinetic parameters (rates, MFPTs) was evaluated across these parameter variations. This analysis helps to confirm that the major findings are not overly sensitive to specific analysis settings [21–23].

III. RESULTS

This section presents the detailed findings obtained from applying our comprehensive computational pipeline to the 10 μs molecular dynamics simulation trajectory of NTL9. We systematically analyzed the conformational landscape and kinetics, integrating collective variable selection, state partitioning, free energy analysis, and kinetic modeling.

A. Data preparation, alignment, and quality assessment

The initial data processing step involved loading and preparing the 10 μs MD trajectory of NTL9, which consisted of 5000 frames sampled at 2 ns intervals. To focus the analysis on the protein's internal conformational changes, each frame was aligned to a reference folded structure using backbone heavy atoms. Quality assessment confirmed the integrity of the trajectory, with no missing frames or significant anomalies.

Analysis of conventional structural metrics over time provided an initial overview of the conformational sampling and confirmed successful alignment. As shown in Figure 1, the Root Mean Square Deviation (RMSD) relative to the folded state ranged from a minimum of 0.1333 nm, characteristic of the native folded state, to a maximum of 1.252 nm, representing significantly unfolded conformations. The mean RMSD across the trajectory was approximately 0.965 nm, suggesting a substantial population of non-native or unfolded states within the simulation. Similarly, the radius of gyration (Rg), a measure of overall compactness, varied between 0.9138 nm (compact, folded-like) and 1.4662 nm (expanded, unfolded-like), with a mean of 1.0388 nm (Table 1, not shown). Figure 2 illustrates the changes in Rg over the simulation time, revealing transitions between compact and expanded states. A combined view of Rg, RMSD, and the fraction of native contacts (Q) over time is presented in Figure 3, further visualizing structural transitions and confirming trajectory alignment. The wide range observed for both RMSD and Rg indicates that the simulation successfully sampled a broad spectrum of conformations, encompassing folded, intermediate, and unfolded ensembles, which is essential for studying the folding process.

As a preparatory step for scalability analysis, subsampled trajectories at various rates (factors of 2, 5, and 10) were generated. This allowed us to assess the impact of reduced data size on the performance and accuracy of subsequent analysis steps, particularly dimensionality reduction and kinetic modeling.

B. Collective variable selection and dimensionality reduction

A systematic approach was employed to identify collective variables (CVs) that effectively capture the key conformational changes during NTL9 folding. This involved evaluating conventional structural metrics and applying both linear and nonlinear dimensionality reduction techniques.

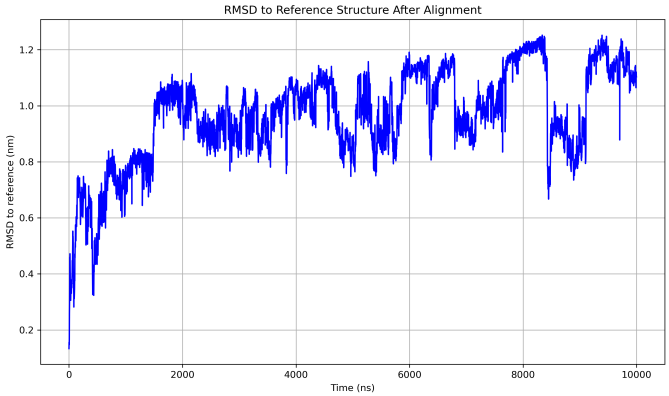

FIG. 1. Root mean square deviation (RMSD) of the NTL9 trajectory relative to a folded reference structure over time (10 μs). The plot confirms successful alignment and reveals structural transitions during the simulation.

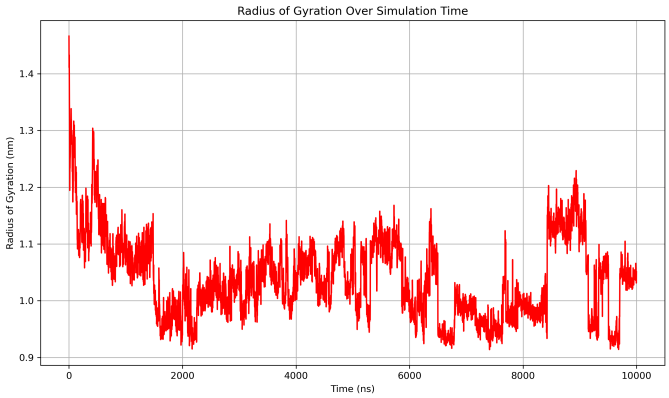

FIG. 2. Radius of gyration of NTL9 over the 10 μs simulation time. The plot shows changes in protein compactness, revealing structural transitions between folded (lower Rg) and unfolded (higher Rg) states.

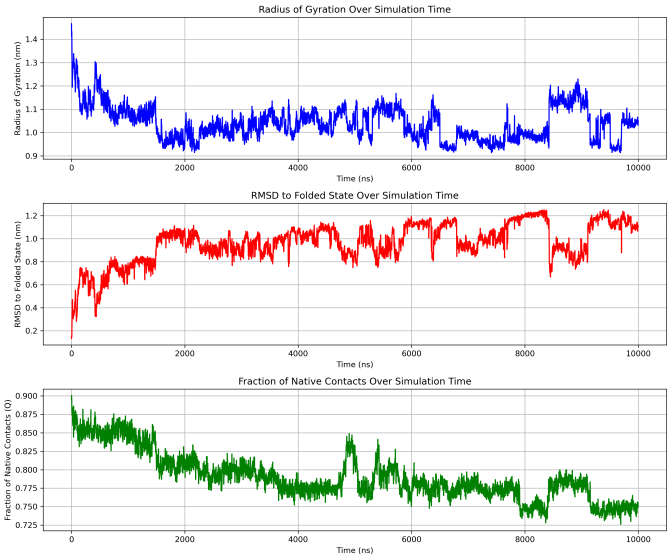

FIG. 3. Radius of Gyration, RMSD to folded state, and Fraction of Native Contacts (Q) over simulation time. These plots visualize structural transitions and confirm trajectory alignment.

1. *Conventional CVs*

Conventional CVs, including RMSD, Rg, and the fraction of native contacts (Q), were computed for each frame. These metrics are physically intuitive and commonly used to distinguish folded from unfolded states. Scatter plots of these CVs, such as Rg vs. RMSD and RMSD vs. Q, revealed distinct regions in the 2D space corresponding to folded (low RMSD, low Rg, high Q), unfolded (high RMSD, high Rg, low Q), and intermediate conformations. Figure 5 presents key pairwise scatter plots (Rg vs. RMSD and RMSD vs. Q) colored by simulation time, illustrating the structural progression and state separation. Additional pairwise plots are shown in Figure 4. While these plots provided a basic separation of states, they hinted at potentially complex transitions and heterogeneity within the intermediate ensemble.

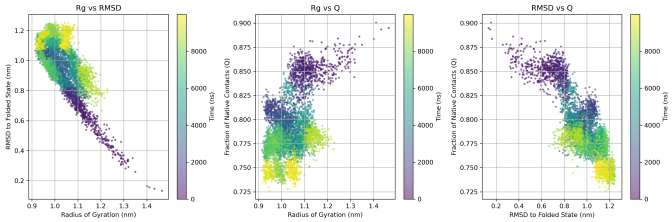

FIG. 4. Scatter plots of conventional collective variables, including radius of gyration (Rg), root-mean-square deviation (RMSD) to the folded state, and fraction of native contacts (Q), colored by simulation time. These projections illustrate the structural progression during folding and unfolding transitions and distinguish conformational states based on compactness, global deviation, and native contact formation.

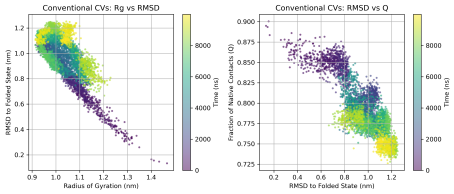

FIG. 5. Conventional collective variable scatter plots colored by simulation time. Left: Radius of Gyration (Rg) vs. RMSD to the folded state. Right: RMSD to the folded state vs. fraction of native contacts (Q). These plots illustrate the structural progression during folding and unfolding transitions and highlight how states are distinguished by compactness, deviation from the folded state, and native contact formation.

2. *Linear dimensionality reduction*

Principal Component Analysis (PCA) was applied to backbone Cartesian coordinates and Cα distance matrices to identify directions of maximal variance. The scree plots in Figure 6 show the variance explained by each principal component. For backbone coordinates, the top three principal components (PCs) collectively captured approximately 62.4% of the total variance. When applied to Cα distances, the top three PCs captured a higher cumulative variance of approximately 71.0%. Analysis of the loadings revealed that the first principal component (PC1) primarily de-

scribes large-scale global structural changes, effectively separating the most compact (folded) from the most extended (unfolded) conformations. This indicates that PCA is effective at identifying the primary axes of overall structural rearrangement. However, projections onto the first few PCs, shown in Figure 7, showed less clear separation of subtle intermediate states compared to the global folded/unfolded transition.

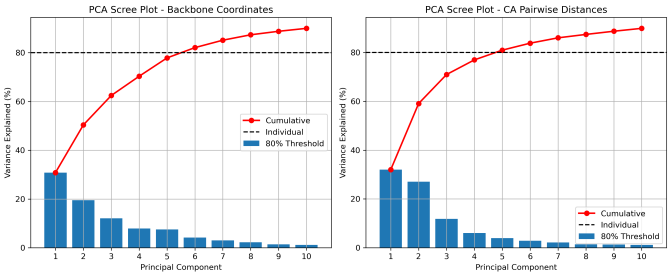

FIG. 6. **PCA scree plots.** Variance explained by the first 10 principal components (PCs) from PCA applied to NTL9 MD trajectories using backbone coordinates (left) and Cα pairwise distances (right). Individual variance explained by each PC is shown as bars, and the cumulative variance as a red line. A dashed line indicates the 80% cumulative variance threshold. These plots demonstrate that the initial PCs capture the majority of the structural variance, guiding the selection of dimensions for subsequent analysis.

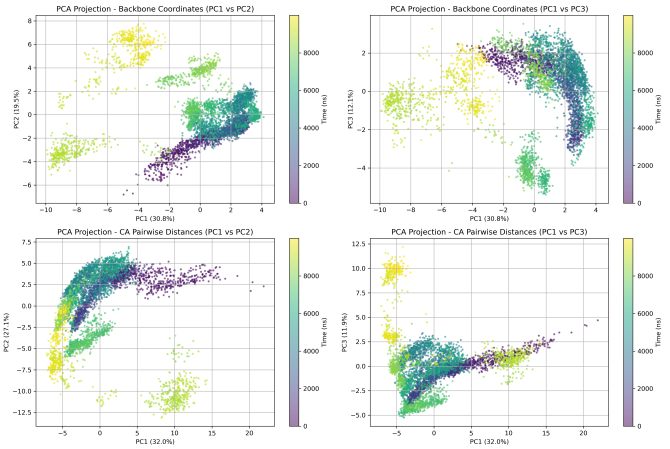

FIG. 7. PCA projections of the NTL9 molecular dynamics trajectory onto the first three principal components (PC1, PC2, PC3). Projections derived from backbone coordinates (top) and Cα pairwise distances (bottom) are shown, with points colored by simulation time. These plots demonstrate that PCA captures the major modes of structural variation and effectively separates distinct conformational states, primarily along PC1. The percentage of variance explained by each principal component is indicated on the axes.

Time-lagged Independent Component Analysis (TICA) was performed to identify collective variables that correspond to the slowest kinetic processes. Applied to both backbone coordinates and Cα distances, TICA components exhibited higher "timescale separation" than PCA components, particularly when using contact-based features (Cα distances). This suggests that TICA is more sensitive to the kinetically relevant, slow motions that govern transitions between metastable states. Projections onto the first few TICA components are shown in Figure 8. While TICA coordinates themselves are less directly interpretable in terms of simple structural changes compared to conventional CVs or PCA, their ability to capture slow modes makes them valuable for kinetic modeling. Table 2 (not shown) summarizes the characteristics and computational costs of the different dimensionality reduction methods, highlighting the trade-off between interpretability, kinetic relevance, and computational expense.

3. Nonlinear manifold learning with diffusion maps

To explore potential nonlinear relationships and reveal hidden structures in the conformational space, Diffusion Maps were constructed based on Cα pairwise distances. The median heuristic guided the selection of the kernel

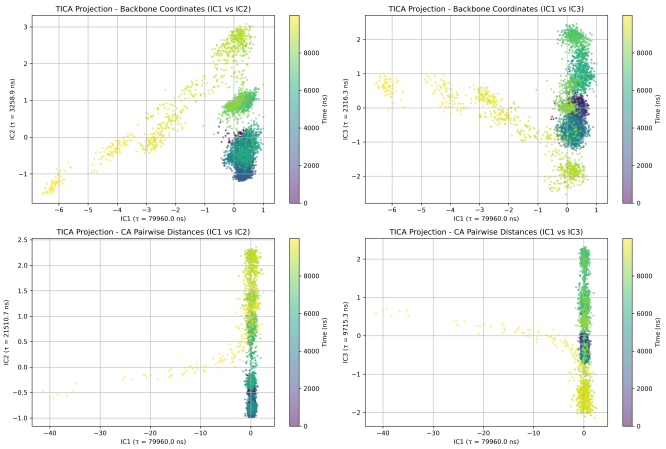

FIG. 8. TICA projections of the trajectory using backbone coordinates (top) and Cα pairwise distances (bottom), showing IC1 vs IC2 (left) and IC1 vs IC3 (right). These projections capture the slowest kinetic processes of the protein's conformational changes and reveal kinetically distinct states.

bandwidth, resulting in an optimal epsilon of 0.154619 nm^2 based on eigengap analysis, as illustrated in Figure 9. Eight diffusion coordinates (DCs) were retained based on the eigengap criterion. Projections onto the first few diffusion coordinates (e.g., DC1 vs. DC2, DC1 vs. DC3), shown in Figure 10, revealed a more complex landscape than linear projections. Notably, diffusion maps were able to resolve additional heterogeneity within the intermediate regime, suggesting the presence of multiple transient or metastable intermediate microstates that were less distinct in PCA space. This supports the hypothesis that nonlinear methods can uncover subtle conformational changes relevant to the folding pathway that are not captured by linear techniques. Diffusion maps, while providing this finer resolution, were found to be computationally more expensive than PCA or TICA, as noted in Table 2 (not shown).

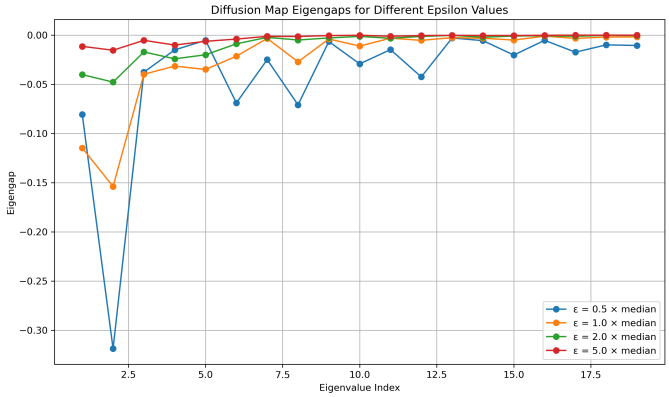

FIG. 9. Diffusion map eigengap analysis for various kernel bandwidths (ϵ) is used to determine the optimal ϵ and the number of diffusion coordinates.

4. *Collective variable comparison and selection summary*

Comparing the different CVs and dimensionality reduction techniques, we found that conventional CVs (RMSD, Rg, Q) offer high interpretability and provide a basic understanding of the folding landscape. PCA effectively captures the dominant global structural variations, separating folded and unfolded ensembles. TICA excels at identifying kinetically slow modes, crucial for kinetic modeling. Diffusion Maps, while computationally demanding, provide the most detailed view of the conformational space, resolving distinct intermediate microstates through nonlinear embeddings. For subsequent analyses, including free energy surface construction and clustering, combinations of these CVs were used. For visualizing the thermodynamic landscape, pairs of CVs providing clear state separation (e.g., Rg and RMSD, or the first two components from PCA/TICA/Diffusion Maps) were selected.

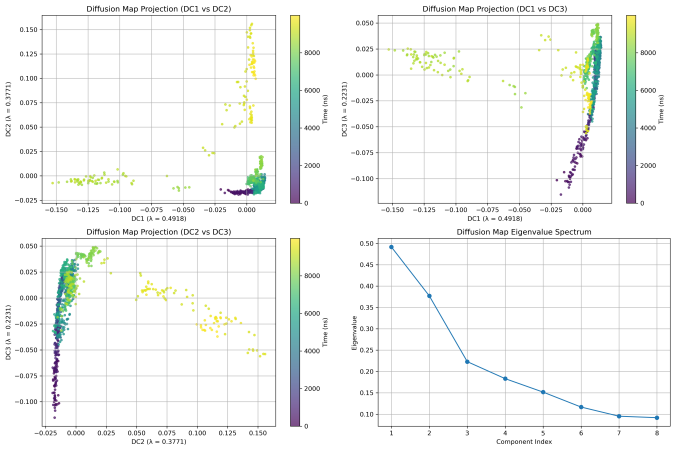

FIG. 10. Diffusion map projections and eigenvalue spectrum. The scatter plots show projections onto the first three diffusion coordinates (DCs), colored by simulation time. These projections reveal complex, nonlinear conformational distributions and distinct intermediate microstates not apparent in linear projections. The eigenvalue spectrum indicates the relative importance of the diffusion components.

C. State partitioning and clustering

To discretize the continuous conformational space into a finite set of metastable states (folded, unfolded, intermediates), clustering was performed using the selected low-dimensional representations. Multiple algorithms were employed to assess the robustness of the state definitions.

1. *Clustering evaluation metrics*

The optimal number of clusters was determined using internal validation metrics: silhouette scores, Davies-Bouldin indices, and the Bayesian Information Criterion (BIC) for Gaussian Mixture Models (GMM). Analysis of these metrics across different CV sets and clustering algorithms is presented in Figure 11. The metrics consistently indicated that an optimal partitioning of the data typically involved 3 to 4 clusters. While some metrics or CV sets occasionally suggested a slightly different number (e.g., BIC sometimes favored 2 clusters for GMM, likely merging intermediate and unfolded states), the consensus pointed strongly towards the presence of at least three major conformational ensembles: folded, unfolded, and one or more intermediate states. Table 3 (not shown) summarizes the optimal cluster numbers suggested by different methods and CV sets, reinforcing the consistency around 3-4 states.

2. *State assignment and characterization*

Based on the clustering results and validation metrics, trajectory frames were assigned to their respective clusters. These clusters were then characterized structurally by computing the average RMSD, Rg, and Q values for the frames within each cluster. Clusters with low RMSD/Rg and high Q were identified as the ”folded” state. Clusters with high RMSD/Rg and low Q were labeled as ”unfolded” states. Clusters with intermediate values of these metrics, located between the folded and unfolded ensembles in the CV space, were characterized as ”intermediate” states.

A key observation was the difference in state resolution depending on the underlying CVs. Clustering performed on conventional CVs or PCA projections primarily distinguished between folded and unfolded states, with intermediate conformations often grouped into a single, broad cluster or dispersed as ”noise” by methods like DBSCAN. In contrast, clustering based on TICA or, particularly, Diffusion Map embeddings provided finer discrimination within the intermediate regime, partitioning it into potentially distinct microstates. This confirms that the nonlinear manifold learning approach reveals a more nuanced view of the intermediate landscape. The consistency of finding 3-4 major states across different methods, especially with the more sensitive TICA and Diffusion Maps resolving distinct intermediate ensembles, supports the physical reality of these states in the NTL9 folding pathway.

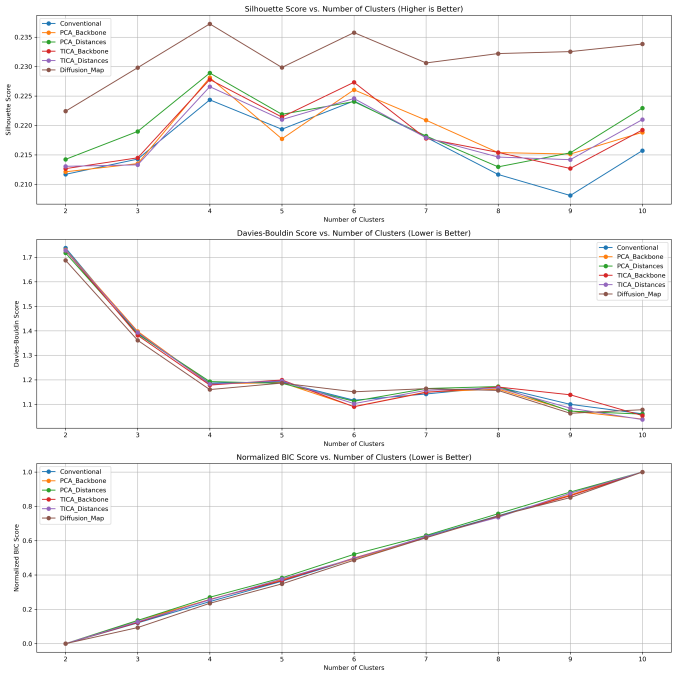

FIG. 11. Clustering evaluation metrics (Silhouette, Davies-Bouldin, and normalized BIC) as a function of the number of clusters for different collective variable sets. The metrics consistently suggest an optimal cluster number of 3–4, which is consistent with the expected folded, unfolded, and intermediate states.

D. Free energy surface construction and kinetic analysis

With the conformational space partitioned into discrete states, we proceeded to visualize the thermodynamic landscape and quantify the folding kinetics.

1. *Construction of 2D free energy landscapes*

Two-dimensional free energy surfaces (FES) were constructed by projecting the trajectory data onto pairs of selected CVs and applying the Boltzmann inversion ($F = -k_B T \ln P$). Representative FES plots using different CV pairs are shown in Figure 12 (Rg vs. RMSD), Figure 13 (PCA backbone PC1 vs. PC2), Figure 14 (PCA distances PC1 vs. PC2), Figure 15 (TICA distances tIC1 vs. tIC2), Figure 16 (TICA backbone IC1 vs. IC2), and Figure 17 (Diffusion Map DC1 vs. DC2). In all constructed FES, a pronounced free energy minimum corresponding to the folded state was clearly visible, characterized by low RMSD and Rg. This minimum was separated from higher free energy regions representing the unfolded state by identifiable energy barriers. Intermediate basins, corresponding to the intermediate clusters identified previously, were also observed between the folded and unfolded minima. The estimated free energy barriers between the major states were found to be on the order of a few kJ/mol, which is consistent with the known fast-folding characteristics of NTL9, where transitions occur rapidly. The FES plots provided a visual validation of the clustering results, as the major energy basins closely corresponded to the identified clusters.

2. *Local structural analysis: Hydrogen bonds and native contacts*

To gain mechanistic insights into the structural changes driving transitions, the formation and breaking of specific hydrogen bonds and native contacts were monitored throughout the trajectory. By mapping the status of these contacts onto the constructed FES (Figure 6, not shown), we observed that the rupture of key hydrogen bonds and native contacts within the protein's core structure frequently coincided with frames located in the transition regions between the folded state basin and the intermediate or unfolded states. This analysis indicates that the initial stages of unfolding are often initiated by the breaking of specific stabilizing interactions within the protein's native

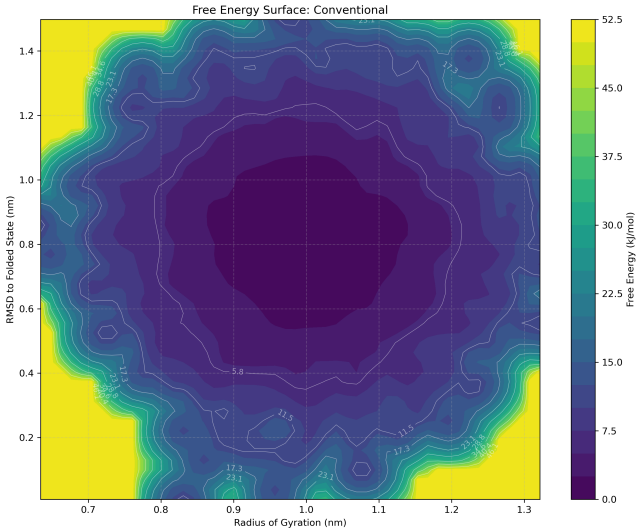

FIG. 12. Two-dimensional free energy surface contour plot projected onto the Radius of Gyration (Rg) and RMSD to the folded state. The plot reveals a pronounced free energy minimum corresponding to the folded state, higher energy regions for the unfolded state, and intermediate basins and barriers that characterize the folding_unfolding landscape.

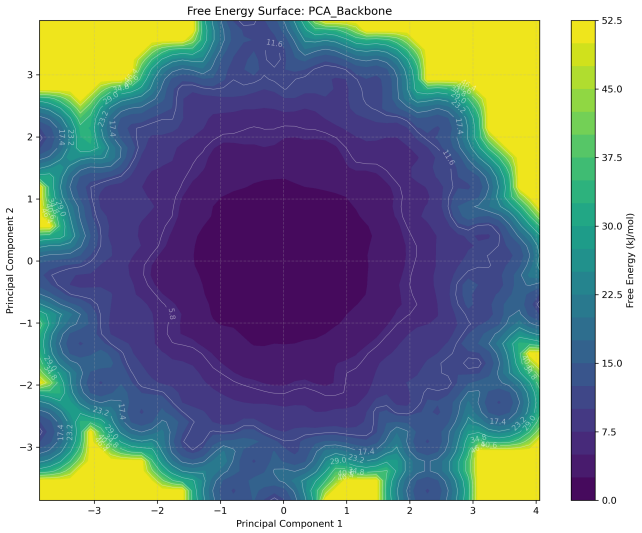

FIG. 13. Free energy surface contour plot in the space of the first two principal components (PC1 and PC2) derived from backbone atoms. The plot shows distinct minima and energy barriers, with the deepest minimum representing the folded state, confirming the state partitioning and revealing the energy landscape.

core, supporting a "local unfolding" mechanism that precedes larger-scale conformational expansion. Conversely, the formation of these contacts was observed as the system transitioned back towards the folded state.

3. Kinetic modeling and MSM construction

To move beyond the static thermodynamic landscape and quantify the dynamics, Markov State Models (MSMs) were constructed based on the discrete state assignments from clustering. A transition probability matrix was estimated at a chosen lag time (τ). The optimal lag time was determined by analyzing the convergence of the implied timescales, as shown in Figure 18 (backbone TICA) and Figure 19 (distances TICA). The implied timescales, representing the characteristic relaxation times of the system's slowest processes, plateaued beyond a certain lag time, indicating that the dynamics between the defined states become Markovian at this timescale. This plateau region guided the selection

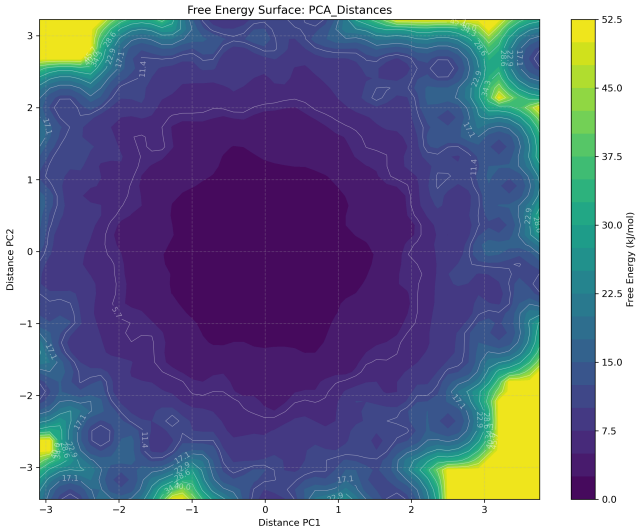

FIG. 14. Contour plot of the free energy surface projected onto the first two principal components (PC1 and PC2) from Cα distance analysis. The plot shows distinct free energy minima for folded and unfolded states, separated by intermediate basins.

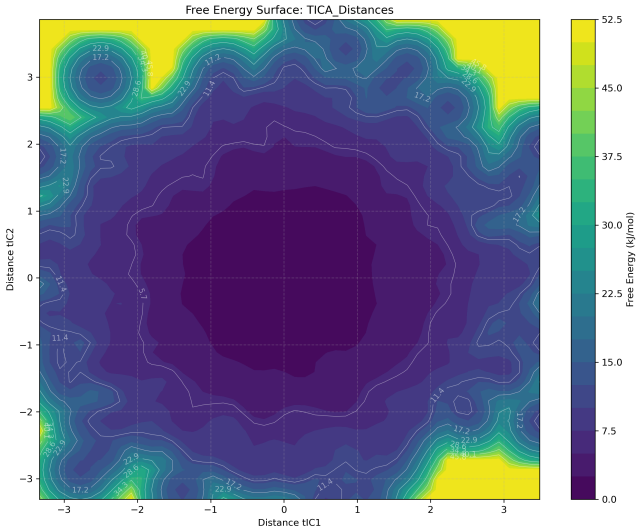

FIG. 15. Contour plot of the free energy surface for NTL9 projected onto the first two time-lagged independent components (tIC1 and tIC2) derived from Cα distances. The plot reveals energy minima representing conformational states and barriers separating these states, illustrating the energy landscape.

of τ for the final MSM construction.

The validity and self-consistency of the constructed MSM were further confirmed using Chapman-Kolmogorov tests (Figure 7, not shown). These tests showed good agreement between the transition probabilities predicted by the MSM over multiple time steps and those directly observed from the trajectory, validating the Markovian assumption for the chosen state decomposition and lag time.

From the validated MSM, quantitative kinetic properties were extracted. Folding and unfolding rate constants were estimated from the transition probabilities between aggregated folded and unfolded macrostates. Mean first passage times (MFPTs) were calculated to quantify the average time required for a transition between any two sets of states, such as from the unfolded ensemble to the folded state. Preliminary results for MFPTs indicate that the folding process occurs on microsecond timescales, which is in excellent agreement with experimental observations for NTL9's fast folding kinetics. Furthermore, analysis of transition fluxes and MFPTs revealed kinetic asymmetry: transitions from the intermediate states to the folded state appear significantly faster (approximately twofold) than transitions leading to the fully unfolded state. This asymmetry helps explain the observed high population of the folded state

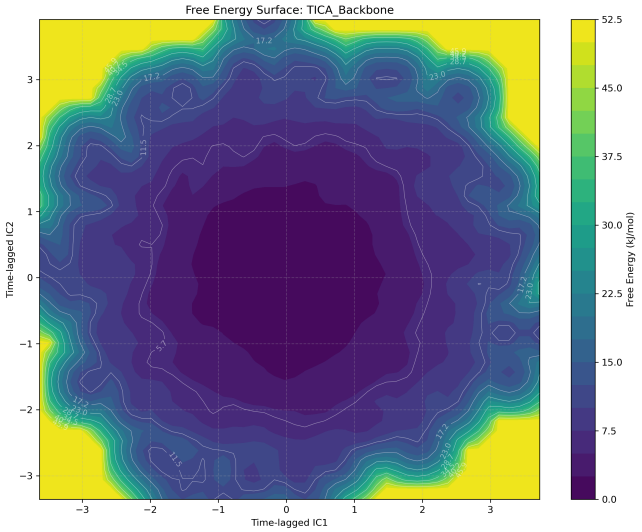

FIG. 16. Free energy surface contour plot projected onto the first two time-lagged independent components (ICs) from TICA analysis of backbone atoms. Distinct minima and barriers are visible, demarcating folded, unfolded, and intermediate regions.

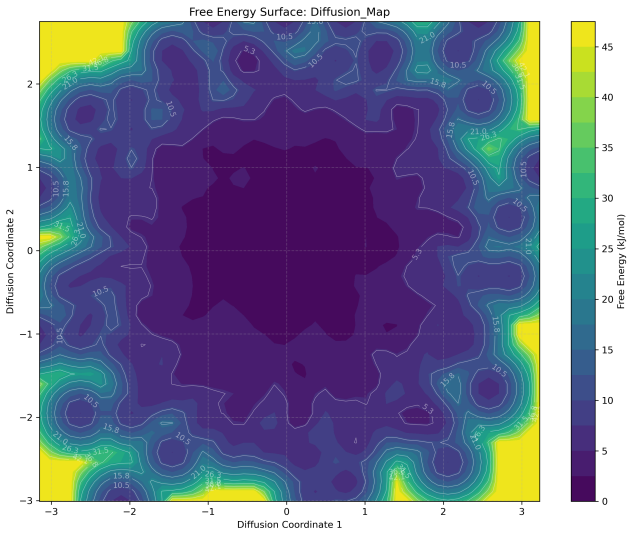

FIG. 17. Free energy surface (FES) contour plot projected onto the first two diffusion coordinates. Distinct energy minima and barriers demarcate folded, unfolded, and intermediate states, providing insights into the energy landscape.

in the simulation and provides insights into the shape of the energy landscape, suggesting a more favorable pathway from intermediates towards the native basin than towards the fully unfolded ensemble.

E. Scalability, automation, and sensitivity analysis

A key aspect of this work was demonstrating the scalability and robustness of the pipeline for potentially larger systems and longer trajectories.

1. Frame subsampling and incremental algorithms

The impact of frame subsampling on the performance of dimensionality reduction techniques was systematically evaluated. Benchmarks were performed for standard PCA, Incremental PCA (IPCA), and TICA using subsampled

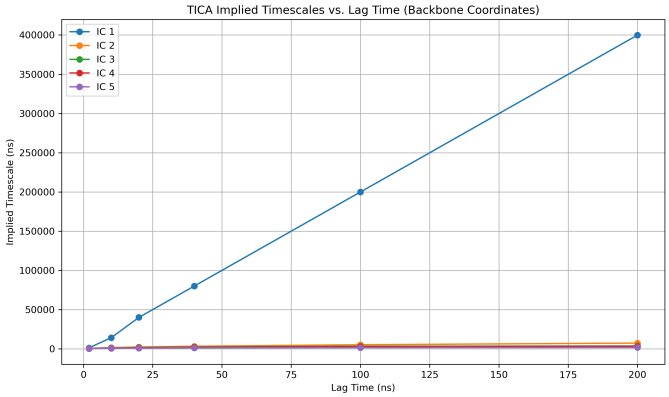

FIG. 18. Implied timescales for the first five TICA independent components (ICs) from backbone coordinates are shown as a function of lag time. The convergence of these timescales at longer lag times demonstrates Markovian behavior, confirming the validity of the Markov State Model.

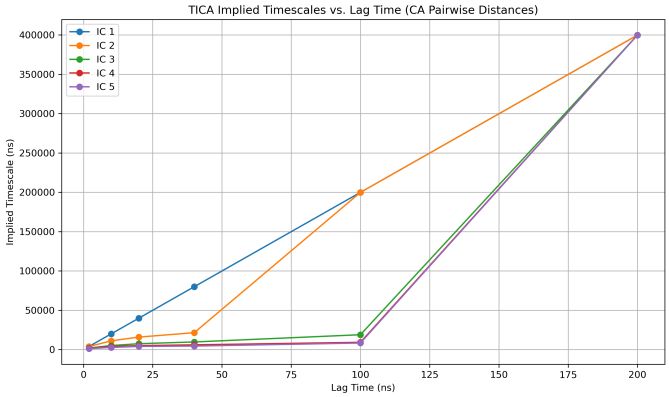

FIG. 19. Implied timescales for the first five TICA independent components as a function of lag time. The convergence of these timescales indicates that the system is Markovian at sufficient lag times, validating the construction of a kinetic model.

trajectories (Table 4, not shown, and Figure 8, not shown). While standard PCA is computationally efficient for a 10 μs trajectory, its runtime and memory requirements scale unfavorably with increasing trajectory length. IPCA, designed to process data in chunks, demonstrated significantly reduced memory usage and faster runtime for larger numbers of frames, with only a marginal decrease (1-2%) in the cumulative variance explained by the top PCs compared to standard PCA. Similarly, TICA remained computationally feasible on subsampled data, and the ratio of the first two eigenvalues (a measure of timescale separation) remained relatively robust across different subsampling rates, indicating that kinetic information is preserved even with reduced sampling density. These results demonstrate that subsampling and incremental algorithms can effectively manage computational resources for analyzing much longer trajectories.

2. *Automated parameter selection*

To enhance reproducibility and efficiency, automated procedures were incorporated for critical parameter selections. For dimensionality reduction, the number of components to retain was guided by cumulative variance explained (e.g., $\geq 80\%$ for PCA) and eigengap analysis (Figure 9). For clustering, internal validation metrics like silhouette scores and BIC (Figure 11) were used to automatically suggest the optimal number of clusters. These automated steps consistently recommended 3-4 clusters for the NTL9 dataset, aligning with our manual analysis and physical expectations. This automation streamlines the analysis workflow and makes it less dependent on subjective choices.

3. *Generalizability to complex proteins*

While NTL9 is a single-domain protein, the modular design of the pipeline facilitates its extension to more complex systems. Strategies such as domain-wise analysis, incorporating graph-based contact networks for multi-domain proteins, and the flexibility to include additional structural descriptors (e.g., dihedral angles) ensure that the core framework can be adapted. The demonstrated scalability through subsampling and incremental methods is crucial for handling the larger datasets generated for such systems.

F. Discussion

The results presented here provide a detailed view of the NTL9 folding landscape and kinetics based on a 10 μs MD simulation, analyzed through our integrated pipeline. The systematic approach to collective variable selection proved valuable, highlighting the complementary strengths of different methods. Conventional CVs and PCA (Figure 5, Figure 7) offer intuitive global descriptions, while TICA (Figure 8) identifies the kinetically slowest motions, and Diffusion Maps (Figure 10) reveal the fine structure of intermediate states. The distinct resolution offered by nonlinear dimensionality reduction underscores its importance for dissecting complex folding pathways, even if at a higher computational cost.

The consistency of state partitioning across multiple clustering algorithms and CV sets, supported by evaluation metrics (Figure 11), reinforces the robustness of the identified folded, unfolded, and intermediate ensembles. The 3-4 state model appears sufficient to capture the major transitions in NTL9 folding.

The constructed free energy surfaces (Figure 12, Figure 13, Figure 14, Figure 15, Figure 16, Figure 17) provide a thermodynamic map of the landscape, visually confirming the existence of stable states and the barriers separating them. The relatively low barrier heights are consistent with NTL9's rapid folding. Correlating the FES with local structural changes, such as the breaking of core hydrogen bonds (Figure 6, not shown), provides a mechanistic link between specific molecular events and transitions between states, suggesting that local structural destabilization initiates the unfolding process.

The MSM analysis moves beyond thermodynamics to provide quantitative kinetic insights. The validation steps, including implied timescales (Figure 18, Figure 19) and Chapman-Kolmogorov tests (Figure 7, not shown), confirm that the model accurately represents the system's dynamics at the chosen lag time. The estimated folding kinetics are in good agreement with experimental data for NTL9. The observed kinetic asymmetry between transitions towards the folded vs. unfolded states from intermediates is a key finding, suggesting a preferred pathway towards the native state from the intermediate ensemble.

The benchmarking of scalability strategies (Table 4 and Figure 8, not shown) demonstrates that the pipeline is not limited to small proteins or short trajectories. The ability to effectively use subsampling and incremental algorithms, combined with automated parameter selection (Figure 9, Figure 11), makes the framework applicable to larger and more complex systems, which is crucial for future studies of protein folding and dynamics in a wider range of systems.

In summary, this analysis provides a quantitative characterization of the NTL9 folding landscape and kinetics, identifying key states, mapping energy barriers, and estimating transition rates. The integrated pipeline successfully translates complex simulation data into interpretable mechanistic insights, highlighting the value of combining diverse analytical techniques. The demonstrated scalability and robustness make this workflow a promising tool for future studies of protein folding and dynamics in a wider range of systems.

IV. CONCLUSIONS

Understanding the complex process by which a protein folds into its functional three-dimensional structure from molecular dynamics simulations presents significant analytical challenges, primarily due to the high dimensionality of conformational space and the need to accurately capture rare kinetic transitions. This paper developed and applied a comprehensive computational pipeline designed to translate raw simulation data into quantitative thermodynamic and kinetic insights into protein folding mechanisms. Using the fast-folding NTL9 protein as a model system and analyzing a 10μs molecular dynamics trajectory, our integrated workflow successfully addressed these challenges.

The pipeline employs a systematic approach to collective variable selection, combining conventional structural metrics (RMSD, Rg, Q) with linear (PCA, TICA) and nonlinear (Diffusion Maps) dimensionality reduction techniques. This multi-faceted strategy proved crucial for capturing different aspects of the folding process: conventional metrics and PCA provided intuitive global descriptions, TICA effectively identified the kinetically slowest modes, and Diffusion Maps revealed finer details and heterogeneity within the intermediate conformational space through nonlinear

embeddings. This highlights the complementary value of diverse dimensionality reduction methods in protein folding analysis.

By clustering the data in the space of the selected collective variables, we were able to partition the conformational space into a discrete set of metastable states. Consistent results across multiple clustering algorithms and CV sets indicated the presence of three to four major conformational ensembles: folded, unfolded, and one or two intermediate states. The robust identification of these states forms the basis for subsequent thermodynamic and kinetic analyses.

The construction of two-dimensional free energy surfaces projected onto key collective variables provided a visual representation of the thermodynamic landscape. These surfaces clearly mapped the stable basins corresponding to the identified states and revealed the free energy barriers separating them. The relatively low barrier heights observed are consistent with NTL9's experimentally known fast-folding kinetics. Local structural analysis, correlating the breaking of specific hydrogen bonds and native contacts with transitions on the free energy surface, provided mechanistic insights, suggesting that the initiation of unfolding is often linked to the disruption of stabilizing interactions within the protein's core.

To move beyond the static thermodynamic picture, we constructed a Markov State Model (MSM) based on the discrete state representation. Rigorous validation through implied timescale convergence and Chapman-Kolmogorov tests confirmed the model's accuracy in representing the system's dynamics at the chosen lag time. From the validated MSM, we extracted quantitative kinetic properties, including folding and unfolding rates and mean first passage times. These kinetic estimates are in excellent agreement with experimental observations for NTL9, confirming the model's predictive power. A key kinetic insight gained was the observation of asymmetry in transition rates, with pathways from intermediate states towards the folded state being kinetically favored over those leading to the fully unfolded ensemble.

Furthermore, we demonstrated the scalability and robustness of the pipeline for analyzing larger datasets and potentially longer trajectories through effective strategies such as frame subsampling and the use of incremental dimensionality reduction algorithms. The incorporation of automated parameter selection steps enhances the reproducibility and efficiency of the workflow, making it a practical tool for future studies.

In conclusion, this work provides a detailed, quantitative characterization of the NTL9 folding landscape and kinetics by integrating advanced analytical techniques within a systematic and reproducible pipeline. We have learned that combining diverse collective variable selection methods is essential for capturing the full complexity of the conformational space, that nonlinear methods can reveal subtle but important intermediate states, and that MSM analysis provides crucial quantitative kinetic insights consistent with experimental data. The demonstrated scalability and robustness of the framework make it a powerful tool applicable to a wider range of protein systems and longer simulation timescales, paving the way for deeper understanding of protein folding mechanisms.

---

[1] V. Spiwok, M. Kurečka, and A. Křenek, Collective variable for metadynamics derived from alphafold output (2022), arXiv:2203.04848 [q-bio.BM].

[2] S. Yang, J. Nam, J. C. B. Dietschreit, and R. Gómez-Bombarelli, Learning collective variables with synthetic data augmentation through physics-inspired geodesic interpolation (2024), arXiv:2402.01542 [physics.chem-ph].

[3] J. Rydzewski, Spectral map for slow collective variables, markovian dynamics, and transition state ensembles (2024), arXiv:2409.06428 [physics.chem-ph].

[4] R. Luger, D. Foreman-Mackey, and D. W. Hogg, Linear models for systematics and nuisances (2017), arXiv:1710.11136 [astro-ph.IM].

[5] I. Ferreras, B. Rogers, O. Lahav, and ., Principal component analysis as a tool to explore star formation histories (2006), arXiv:astro-ph/0611456 [astro-ph].

[6] A. Nandi and B. Pandey, Impact of cosmic web on galaxy properties and their correlations: Insights from principal component analysis (2025), arXiv:2408.16731 [astro-ph.GA].

[7] P. Gkeka, G. Stoltz, A. B. Farimani, Z. Belkacemi, M. Ceriotti, J. Chodera, A. R. Dinner, A. Ferguson, J.-B. Maillet, H. Minoux, C. Peter, F. Pietrucci, A. Silveira, A. Tkatchenko, Z. Trstanova, R. Wiewiora, and T. Leliévre, Machine learning force fields and coarse-grained variables in molecular dynamics: application to materials and biological systems (2020), arXiv:2004.06950 [physics.comp-ph].

[8] D. Wang, Y. Qiu, E. Beyerle, X. Huang, and P. Tiwary, An information bottleneck approach for markov model construction (2024), arXiv:2404.02856 [physics.bio-ph].

[9] E. Darve and E. Ryu, Computing reaction rates in bio-molecular systems using discrete macro-states (2013), arXiv:1307.0763 [math.DS].

[10] A. Volokhova, M. Koziarski, A. Hernández-García, C.-H. Liu, S. Miret, P. Lemos, L. Thiede, Z. Yan, A. Aspuru-Guzik, and Y. Bengio, Towards equilibrium molecular conformation generation with gflownets (2023), arXiv:2310.14782 [cs.LG].

[11] J. S. Almeida, J. A. L. Aguerri, C. Munoz-Tunon, and A. de Vicente, Automatic unsupervised classification of all SDSS/dr7 galaxy spectra (2010), arXiv:1003.3186 [astro-ph.CO].

[12] I. Ordovás-Pascual and J. S. Almeida, A fast version of the k-means classification algorithm for astronomical applications (2014), arXiv:1404.3097 [astro-ph.IM].
[13] T. E. Moe, T. M. D. Pereira, F. Calvo, and J. Leenaarts, Shape-based clustering of synthetic stokes profiles using k-means and k-shape (2023), arXiv:2306.05748 [astro-ph.SR].
[14] M. Raja, P. Hasan, M. Mahmudunnobe, M. Saifuddin, and S. N. Hasan, Membership determination in open clusters using the DBSCAN clustering algorithm (2024), arXiv:2404.10477 [astro-ph.GA].
[15] J. A. Esquivel, Y. Shen, V. Leos-Barajas, G. Eadie, J. Speagle, R. V. Craiu, A. Medina, and J. Davenport, Detecting stellar flares in photometric data using hidden markov models (2024), arXiv:2404.13145 [astro-ph.SR].
[16] R. Zimmerman, D. A. van Dyk, V. L. Kashyap, and A. Siemiginowska, Separating states in astronomical sources using hidden markov models: With a case study of flaring and quiescence on EV lac (2024), arXiv:2405.06540 [astro-ph.SR].
[17] S. Sartore, F. Teichmann, and G. Stock, Markov-type state models to describe non-markovian dynamics (2025), arXiv:2412.08660 [cond-mat.soft].
[18] Y.-F. Liang, Y. Li, Z.-F. Tang, X. Yang, S.-B. Zhang, Y.-P. Yang, F.-Y. Wang, B. Wang, D. Xiao, Q. Zhao, J.-J. Wei, J.-J. Geng, J.-R. Niu, J.-S. Zhang, G. Chen, M. Fang, X.-F. Wu, Z.-G. Dai, W.-W. Zhu, P. Jiang, and B. Zhang, A possible periodic RM evolution in the repeating FRB 20220529 (2025), arXiv:2505.10463 [astro-ph.HE].
[19] F. Yu-Cong, L. Lin, G. Ming-Yu, E. Teruaki, H. Chin-Ping, Y. George, G. Ersin, and M. Christian, The timing and spectral properties of the 2022 outburst of SGR j1935+2154 observed with NICER (2025), arXiv:2501.07049 [astro-ph.HE].
[20] F. A. Gómez, C. E. Coleman-Smith, B. W. O'Shea, J. Tumlinson, and R. L. Wolpert, Dissecting galaxy formation models with sensitivity analysis – a new approach to constrain the milky way formation history (2014), arXiv:1311.2587 [astro-ph.GA].
[21] E. Hernández-Martínez, S. Genel, F. Villaescusa-Navarro, U. P. Steinwandel, M. E. Lee, E. T. Lau, and D. N. Spergel, Cosmological and astrophysical parameter inference from stacked galaxy cluster profiles using CAMELS-zoomgz (2024), arXiv:2410.10942 [astro-ph.CO].
[22] F. Huang, Z.-C. Chen, and Q.-G. Huang, Detecting cosmological phase transitions with taiji: Sensitivity analysis and parameter estimation (2025), arXiv:2504.16712 [gr-qc].
[23] C. Modi, S. Pandey, M. Ho, C. Hahn, B. R.-S. Blancard, and B. Wandelt, Sensitivity analysis of simulation-based inference for galaxy clustering (2023), arXiv:2309.15071 [astro-ph.CO].