Understanding Chromatin Organization and Dynamics with Coarse-Grained Modeling

by

Shuming Liu

B.S. Chemistry, Peking University, 2019

Submitted to the Department of Chemistry
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN CHEMISTRY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2024

Authored by:	Shuming Liu Department of Chemistry May 14, 2024
Certified by:	Bin Zhang Pfizer-Laubach C.D. Associate Professor of Chemistry, Thesis Supervisor
Accepted by:	Adam P. Willard Professor of Chemistry Graduate Officer, Department of Chemistry

This doctoral thesis has been examined by a Committee of the Department of Chemistry as follows:

Professor Adam P. Willard...

Thesis Committee Chair

Professor of Chemistry

Professor Bin Zhang...

Thesis Supervisor

Pfizer-Laubach C. D. Associate Professor of Chemistry

Professor Arup K. Chakraborty...

Thesis Committee Member

J.M. Deutch Institute Professor

Departments of Chemical Engineering, Physics, and Chemistry

Core faculty member and former Founding Director, Institute for Medical Engineering & Science, MIT

Founding Steering Committee Member, Ragon Institute of MGH, MIT, & Harvard

Understanding Chromatin Organization and Dynamics with Coarse-Grained Modeling

by

Shuming Liu

Submitted to the Department of Chemistry
on May 14, 2024 in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN CHEMISTRY

ABSTRACT

The genome is the blueprint of human life, and it is crucial to understand its organization. The genome organization is hierarchical with different principles dominating at different scales. At the near-atomistic level, nucleosomes are organized as ordered chromatin fibers or disordered chromatin arrays. Furthermore, chromatin and related proteins can function within condensate environments. Computational modeling provides valuable insights into such complex biological processes. Considering the complexity of chromatin and biomolecular condensates, coarse-grained (CG) modeling is essential to achieve the biologically relevant timescales. We have developed CG models and toolkits to facilitate modeling chromatin and related proteins. We have also applied CG protein and DNA models to study chromatin folding and phase separation.

In Chapter 1, we begin with an overview of the hierarchical scales of genome organization. We also introduce CG modeling as a powerful tool to understand the chromatin structures and dynamics. In Chapters 2 and 3, we demonstrate the development of CG simulation force fields and toolkits. In Chapter 2, we present novel CG force fields trained with contrastive learning. We have achieved a new set of hydropathy parameters trained with a99SB-*disp* all-atom force field trajectories of intrinsically disordered proteins, which accurately reproduces their average radius of gyration. In addition, we have developed a unified force field that captures the average radius of gyration of both ordered and disordered proteins in the training set. In the future, we will focus on benchmarking our models and existing CG models with condensate simulations, which enables more appropriate selections of CG models based on specific conditions. In Chapter 3, we introduce OpenABC, a versatile toolkit designed to streamline the setup of CG simulations, especially condensate simulations. OpenABC incorporates diverse CG force fields within an extensible framework and is built on a simulation platform that supports GPU acceleration, thus speeding up CG simulations.

In Chapters 4 and 5, we shift our focus to the applications of CG simulations. In Chapter 4, we discuss the force extension and inter-chain contacts of chromatin fibers. Our CG simulations reveal that the chromatin fiber behaves like an elastic spring under forces no more than 3 pN, while it dramatically unstacks and unwraps at approximately 4 pN. Meanwhile, inter-chain contacts can help unfold the native two-start fibril-like structures. The study demonstrates that biologically relevant pN-level forces and crowding environments contribute to the absence of 30-nm fibers in vivo. In Chapter 5, we apply Markov state models and non-Markovian dynamics models to study the folding dynamics of tetra-nucleosomes. The tetra-nucleosome with $10n + 5$-bp linkers shows more diverse structures without dominant native structures, while $10n$-bp linkers lead to funnel-shaped free energy landscape with a strong folding trend. Within the condensate, the transition rates slow down, while the unfolding and folding rates are comparable. These two studies highlight that the intrinsic physical chemistry properties of chromatin are fundamental to the genome organization in cells.

Thesis supervisor: Bin Zhang

Title: Pfizer-Laubach C.D. Associate Professor of Chemistry

Acknowledgments

Considering the past five years, it has been a long and challenging journey, I am fortunate to have support from many people so that I can go through all the challenges.

First and foremost, I must attribute the most significant credit to my advisor, Professor Bin Zhang. I admire his diligence and scientific insight. He instructs me to think critically and practise research skills. Moreover, he treats students as his own children and provides a supportive environment. I have witnessed his promotion from an assistant professor to an associate professor, and he deserves the credit. I have truly learned a lot from him.

I am also grateful to my committee members, Professor Adam Willard, and Professor Arup Chakraborty. Their instructions and advice are crucial complements to my research career. It is my honor to have such these distinguished professors on my committee.

I would also like to thank all those colleagues and friends. I am thankful to everyone in the Zhang Group: Xingcheng Lin, Xinqiang Ding, Andrew Latham, Yifeng Qi, Wenjun Xie, Kartik Kamat, Zhongling Jiang, Joe Paggi, Amogh Sood, Cong Wang, Zhuohan Lao, Greg Schuette, Advait Athreya, Justin Airas, Camryn Carter, Ivan Riveros, Yumeng Zhang, Jared Zheng. The Zhang Group is like my home at MIT. We have a good time and I really enjoy the group environment. I also want to express my gratitude to my collaborator Professor Xuhui Huang and Yunrui Qiu at the University of Wisconsin–Madison for their contribution to our project.

I sincerely appreciate the good time with my friends. Sometimes the research journey can be tough and I have to face failure, but those enjoyable and relaxing moments with friends help alleviate stress and remind me of the happiness in my life. It is also because of your understanding and encouragement that I can go through those bad moments and restore motivation.

Finally, I want to express my deepest gratitude to my family members and my girlfriend, your support means everything to me. I want to express my deepest gratitude to my parents and other family members for their understanding and support, and I apologize for my absence for years. To my girlfriend Yingjie Qu, it is my fortune to have you in my life.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The genome is the blueprint of human life. It is remarkable that the extensive 2-meter-long DNA is intricately packed within the tiny 10-μm cell nucleus. Since the genome is organized hierarchically across multiple scales, different physical principles dominate at different scales [1], [2]. The nucleosome is the most fundamental unit of the eukaryotic genome. It is a protein-DNA complex with 147-bp DNA wrapped around a protein octamer called histone [3], [4]. It has a heterogeneous distribution of electrostatic potentials, leading to a sticky molecule with abundant valency and binding sites [5]. At the near-atomistic level, it is vital to understand the structure and dynamics of chromatin, which is the polymer chain composed of nucleosomes connected by DNA linkers. Understanding chromatin structure and dynamics helps unravel the biologically relevant processes such as transcription and chromatin remodeling [6], [7]. There is a notable contradiction related to the chromatin structures: early in vitro experiments proved the existence of the 30-nm fiber with a zig-zag topology and two stacks of nucleosomes [8], [9]; however, in vivo experiments do not support the prevalence of 30-nm fibers, despite its significant stability in vitro [10]–[12]. Instead, the disordered chromatin topology called the 10-nm array dominates in cells [13]. This discrepancy suggests that certain biologically relevant factors may be neglected in some in vitro experiments.

The phase separation has been demonstrated as a widespread phenomenon in cells and governs various biological processes [14]. These membraneless aggregates control biochemical reactions by regulating the partition and dynamics of different molecules within distinct phases [15]. Considering the Flory-Huggins theory and the polymeric nature of chromatin, it is perhaps not surprising that chromatin can also undergo phase separation [2], [16]. Reconsidering the chromatin organization from the perspective of phase separation may help decipher the role of chromatin phase separation in genome organization. In vitro experiments have shown that chromatin can phase separate with liquid-like properties under various biologically relevant conditions [17], [18]. However, the situation becomes more complicated in vivo. The chromatin condensate shows more liquid-like properties locally, for example, at 50-150 nm length scales [19], while more solid- or gel-like at larger scales [19], [20]. Such scale-dependent properties of the chromatin condensate can be attributed to the viscoelastic properties of polymer condensates [2], [21]. The complexity of the phenomena calls for methods to reveal the mechanism and improve comprehension.

Computational modeling plays an indispensable role in studying biomolecules and understanding biological processes. Molecular dynamics (MD) simulations, which describe the motions of molecules based on interactions between molecules and the fundamental laws of mechanics, illustrate molecular motions with complete details [22], [23]. For example, the breathing and unwrapping of a single nucleosome can be explored through 15-μs all-atom explicit solvent simulations. However, the chromatin folding and phase separation, which involve multiple nucleosomes, occur on millisecond or second time scales [17], [24], [25]. These processes may surpass the capacity of state-of-the-art simulation engines, even supercomputers specifically designed for MD simulations [26]. To overcome the time-scale barriers, one workaround is to apply the coarse-grained (CG) modeling. CG modeling increases sampling efficiency by simplifying molecular representations, thus reducing the degrees of freedom and smoothing the effective free-energy landscape. The modifications accelerate dynamics and permit larger integration timesteps [27]. In all, CG modeling extends MD simulations to achieve biologically meaningful timescales and improves comprehension.

There are three aspects to consider when training CG models: the CG mapping, the CG potential, and the learning method. All these aspects are evaluated based on the specific problems. There are many systematic methods to determine the optimal CG mappings [28], [29]. However, we use more intuitive residual-level CG mappings since biopolymers such as proteins and DNA are composed of residues. The residual-level CG mappings preserve sequence information and support transferability to new sequences. Regarding the functional forms of the CG potential, ideally, the CG potential should be the PMF of the given CG mapping, which is typically a complex multi-body potential and hard to parameterize. Inspired by all-atom force fields, traditional CG potentials approximate PMF with bond, angle, dihedral, and two-body non-bonded interactions [30]. There are also attempts to explicitly include multi-body effects into the CG potential as analytical functionals [31], [32]. Modern neural networks are promising ways to capture multi-body effects [33]–[40]. However, the neural networks may be difficult to train, and the inference is much slower than analytical functionals, hindering practical usages in large biomolecular systems. To train the CG potential with the given resolution and functional form, we need to further select the most appropriate learning method. The learning methods can be classified as bottom-up and top-down methods. Bottom-up methods aim to reproduce the microscopic distributions of the fine-grained models [30], [41], while the top-down methods seeks to capture the macroscopic properties measured by experiments. In this thesis, we focus on the bottom-up methods. Theoretically, the best bottom-up CG potential with the given CG mapping should be the potential of mean force (PMF), which consistently reproduces the distributions of CG coordinates based on fine-grained trajectories [42]. However, since the analytical expression of PMF is a high-dimensional integration and impossible to solve, we need some alternative methods to effectively achieve PMF. The existing bottom-up learning methods can be classified as three families: match gradients, match distributions, and compare distributions. The examplary methods of the three families are force matching [42], relative entropy [43], and contrastive learning [44]–[46]. Although these methods are all rigorous and theoretically lead to the unique solution of PMF, they have different features when training in practice. Force matching requires net forces of the fine-grained model, which are not available in most datasets; moreover, atomistic forces are extremely noisy and the mean force is hard to estimate. Relative entropy demands iterations of sampling with the CG model during training. Contrastive learning needs additional efforts for noise sampling. Researchers have to wisely select the training method based on the available data, CG mappings, and CG functional forms.

There are many existing CG models for biopolymers [27], [41] that provide insights into various biological processes related to genome organization and function [1], [2], [47]. Near-atomistic residual-level CG models are particularly suitable for studying the behavior of chromatin since they balance accuracy and efficiency. For instance, residual-level CG models have been utilized to investigate chromatin folding [46], [48] and chromatin force extension [49]. Additionally, many proteins play functional roles in mediating genome organization and functions. Residual-level CG models are also appropriate for studying the interactions between such proteins and nucleosomes or chromatin [50], [51]. Importantly, many related proteins such as linker histones, HP1, and transcription factors have both ordered and disordered domains [52]–[54], necessitating modern CG models that work well for both ordered and disordered proteins [55]. Residual-level CG models have been widely used to study the phase separation of various proteins with disordered regions [55]–[59]. Since the nucleosome concentration in cells is comparable to the nucleosome concentration in condensates measured in vitro [60], it is promising to extend the phase separation simulations to chromatin and decipher the motions of chromatin in condensates.

In the following chapters, we will introduce our studies related to CG modeling of biopolymers, including CG model optimization, toolkit development, and applications. In Chapter 2, we present a series of novel CG protein models trained with the contrastive learning as a bottom-up approach [44]–[46]. We begin by deriving a new hydrophobic scale based on a99SB-*disp* all-atom force field trajectories of intrinsically disordered proteins. This new set of parameters accurately reproduces the average sizes of training set trajectories and approaches the state-of-the-art CG models in matching the experimental average sizes. To generalize the applicability, we extend our training set by incorporating ordered proteins and train a comprehensive model with more balanced non-bonded interactions compatible with both ordered and disordered proteins. We will further improve our model by including multi-body potentials and calibrate our model with existing models in the context of condensate simulations.

In Chapter 3, we present OpenABC [61], a CG model toolkit that incorporates many state-of-the-art CG models into a unified and extensible framework. OpenABC is built on OpenMM [62], which is a versatile Python library optimized for GPU-accelerated simulations. OpenABC streamlines the setup of CG simulations, particularly condensate simulations, which are typically complicated to initialize. Meanwhile, OpenABC utilizes GPU acceleration supported by OpenMM to significantly expedite CG simulations, especially condensate simulations. Together with comprehensive tutorials and documentation, these advantages make CG simulations accessible to a broader community, even experimentalists with limited computational experience. The research article related to the study has been published in *PLoS Computational Biology* 19.9 (2023): e1011442.

In Chapter 4, we shift our focus to the applications of CG models in chromatin organization. We analyze the unfolding and inter-chain interactions of chromatin fibers [49]. Our protein-DNA CG model successfully reproduces the experimental force extension curve of the chromatin fiber, which validates the accuracy of our CG model. Furthermore, the MD simulation reveals the unfolding mechanism. Below 4 pN, the nucleosomal stacking is only partially broken by the shear motions. However, at about 4 pN, some stackings are completely broken apart and the chromatin fiber splits into multiple small oligomers separated by unwrapped linker DNA. We further compute the inter-chain contact free energy and highlight the strong inter-chain contacts. The results suggest that pN-level forces and high concentration contribute to the unfolding of chromatin in cells, which are biologically relevant factors but are sometimes overlooked by in vitro experiments. This explains the stability of the chromatin fiber observed in vitro [9], while it is less prevalent in vivo [10], [12], [13]. The research article related to the study has been published in *Nucleic acids research* 50.17 (2022): 9738-9747.

In Chapter 5, we explore the dynamics of the chromatin by investigating the tetra-nucleosome system. We combine Markov state models (MSMs) and non-Markovian dynamics modeling [63] with chromatin CG modeling to study two biologically important factors: local nucleosomal concentration and linker lengths. These two factors are fundamental in chromatin and are important to mediate the chromatin organization [17], [18], [20], [21], [64]. The tetra-nucleosome with $10n$-bp linkers displays a funnel-shaped free energy landscape with the stable fibril-like structure as the native structure. Within the condensate, the folding and unfolding rates decrease, while the unfolding rate becomes comparable to the folding rate, indicating the reduction in popluation of fibril-like strctures due to the long-range contacts with other nucleosomes. Inserting an additional 5-bp into the $10n$-bp linkers induces additional twist and misaligns the nucleosomes. The tetra-nucleosome with $10n+5$-bp linkers exhibits a more flat free energy landscape, indicating a wide spectrum of partially unfolded structures with similar stability. This study serves as a pioneering example of applying non-Markovian dynamics modeling to study complex biological systems featuring slow dynamics and non-negligible memory effects.

Chapter 2

Contrastive Learning Optimized Coarse-grained Force Field for Ordered and Disordered Proteins

2.1 Introduction

Molecular dynamics (MD) simulations are indispensable tools to study the motions of complex biomolecules [22], [23]. The large size of biomolecules and the long-time scales of biologically significant motions are intrinsic challenges to computational modeling. To overcome the challenges and relieve computational burden, coarse-grained (CG) modeling, which simplifies the system by grouping atoms into CG beads and implicitly capturing solvent effects, is essential for studying large biomolecular systems. It facilitates much simpler representations with smoother free energy landscapes [27], [30], [41]. The integration of modern simulation platforms [26], enhanced sampling techniques [65], and CG modeling advances the understanding of complex biomolecular behaviors.

There are a series of existing methods for training CG models, which can be categorized as bottom-up and top-down methods. Bottom-up methods aim to fit the microscopic distributions given the data from more accurate models, typically of higher resolution [30], [42]. In contrast, top-down methods try to fit the macroscopic properties measured by experiments. The two families of methods are suitable for different situations, such as different CG resolutions and different available data. As CG modeling can be formulated as learning an energy that ideally captures the potential of mean force (PMF) [42], many machine learning algorithms from the energy-based model (EBM) community can be applied to training CG models as bottom-up methods [66]. Such methods include maximum likelihood estimation, minimizing contrastive divergence [67], score matching [68], and contrastive learning (also known as noise contrastive estimation) [44], [45]. These methods share similarities with CG model training methods known to the computational chemistry community. For example, the relative entropy method [43] can be derived from the maximum likelihood principle, while force matching [42] corresponds to score matching as both aim to match gradients of PMF. Among these methods, contrastive learning stands out as an unbiased method that avoids iterations or sampling during training. It also does not require force recordings or high-order derivatives. As a trade-off, noise samples generated with a known energy function are required as a baseline, and the training task is to distinguish data samples from noise samples as logistic regression. Contrastive learning strikes a balance between accuracy and efficiency. In practice, more noise samples and a good overlap between noise distribution and data distribution help to achieve better training results [44], [45]. In particular, the contrastive learning framework has been extended so that the noise samples can come from known potentials instead of known normalized distributions (e.g. Gaussian distribution). Enhanced sampling methods can be further applied to explore diverse noise samples, facilitating better overlap between noise and data distributions [69]. Several CG models have been successfully trained with contrastive learning and demonstrate the power of this method [40], [70].

In the present work, we used the contrastive learning method to train a series of transferable CG force fields based on long-time all-atom explicit solvent simulation trajectories performed by D. E. Shaw Research [71]–[73] and in-house simulation trajectories as our training data. Our training data includes both intrinsically disordered proteins (IDPs) and ordered proteins (OPs). We parameterized a new set of hydropathy parameters based on the a99SB-*disp* force field trajectories of IDPs [72]. We also trained a unified model for both IDPs and OPs with trajectories of IDPs and OPs. It balances the non-bonded interactions within ordered and disordered domains, which should represent the protein-protein interactions more faithfully. We compared our results with existing CG models developed with top-down and other data-driven methods. The new set of hydropathy parameters precisely reproduces the average R_g in the training set. These results highlight the effectiveness of contrastive learning as an emerging bottom-up method in the computational chemistry community and provide a series of powerful models that can be readily used to study phase behaviors of diverse protein sequences.

2.2 Methods

2.2.1 Train coarse-grained force field with contrastive learning

Contrastive learning converts an unsupervised learning task of learning data distribution to a supervised learning task of distinguishing data samples from noise samples. Our goal is to learn the distribution of Cα atom trajectories mapped from the all-atom trajectories as a reduced potential, which is the reduced potential of mean force (PMF) along the Cα positions. This goal can be achieved as a logistic regression task. With N_0 noise samples generated with a known distribution or reduced energy, and N_1 data samples, the log-likelihood is

$$l = \frac{1}{N}\left[\sum_{i=1}^{N_0} \log P(x_i^{(0)} \in X_0 | x_i^{(0)}) + \sum_{i=1}^{N_1} \log P(x_i^{(1)} \in X_1 | x_i^{(1)})\right] \tag{2.1}$$

where X_0 and X_1 represent the noise and data collections, respectively, and $N = N_0 + N_1$. The noise samples are $x_i^{(0)}$, and data samples are $x_i^{(1)}$. In our case, the noise samples come from a known reduced potential $u_0(x)$, thus the noise distribution is $p_0(x) = \exp(-u_0(x)+f_0)$, where f_0 is the reduced free energy. On the other hand, the reduced potential $u_1(x;\theta)$ includes parameters θ that remain to be optimized, and the data distribution is intended to be captured as $p_1(x) = \exp(-u_1(x;\theta^*) + f_1)$, where θ^* represents optimal parameters and f_1 is the reduced free energy. The probability of recognizing a sample x as data or noise is

$$\begin{aligned} P(x \in X_0|x) =& \frac{1}{1 + \nu^{-1} p_1(x)/p_0(x)} \\ P(x \in X_1|x) =& \frac{1}{1 + \nu p_0(x)/p_1(x)} \end{aligned} \tag{2.2}$$

with $\nu = N_0/N_1$. The proof of equation 2.2 is provided in Appendix A section *Proof of contrastive learning loss* based on the Bayes' theorem. The contrastive loss $\mathcal{L}$ is the negative log-likelihood. We can view all the noise and data samples as a mixed ensemble including N samples $\{(x_i, y_i)\}$ $(i = 1, \ldots, N)$, where x_i is the sample configuration and $y_i \in \{0, 1\}$ is the class label. We can formulate $\mathcal{L}$ as binary cross-entropy

$$\mathcal{L}(\theta, \Delta f) = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log \sigma(\alpha_i) + (1 - y_i) \log(1 - \sigma(\alpha_i))] \tag{2.3}$$

with sigmoid $\sigma(\alpha) = 1/(1 + \exp(-\alpha))$, logit $\alpha_i = -\log \nu + u_0(x_i) - u_1(x_i, \theta) + \Delta f$ and $\Delta f = f_1 - f_0$. Maximizing log-likelihood is equivalent to minimizing contrastive loss over θ and Δf. Importantly, although f_0 and f_1 depend on $u_0(x)$ and $u_1(x;\theta)$, respectively, in Appendix A section *Prove the feasibility of viewing Δf as an independent variable*, we prove that we can effectively view Δf and θ as independent parameters to be optimized without losing rigor. An intuitional explanation is that arbitrary constants can be added to u_0 or u_1, thus tuning Δf without changing p_0 or p_1. In the Appendix A section *A sufficient condition for the convexity of the contrastive loss*, we also prove that a potential as a linear function of θ is a sufficient condition for the convexity of $\mathcal{L}$, thus leading to a unique optimal solution $(\theta^*, \Delta f^*)$. In this study, all training parameters θ satisfy this condition. Additionally, we added regularizer to control the scale of parameters θ. The regularizer has the form

$$\mathcal{R}(\theta) = \frac{\zeta}{2}\text{mean}(\theta^2) \tag{2.4}$$

where coefficient ζ is the hyperparameter that controls the strength of the regularizer. The regularizer is a convex function of θ, and the overall loss as $\mathcal{L} + \mathcal{R}$ remains convex.

2.2.2 Coarse-grained force field for disordered and ordered proteins

Here, we introduce the model that we trained for intrinsically disordered proteins (IDPs) and ordered proteins (OPs). Since we only optimize the parameters of non-bonded contacts, while parameters of other terms are all fixed, we only provide details of non-bonded contacts while briefly introducing other potential terms. More details of all potential terms can be found in Appendix A section *Force field definitions*.

The IDP CG model is based on a well-known CG Cα model called hydrophobic scale (HPS) model. It is composed of harmonic bond, non-bonded contact, and Debye-Hückel potentials

$$U_{\text{IDP}} = U_{\text{bond}} + U_{\text{AH}} + U_{\text{elec}} \tag{2.5}$$

Our training target is to optimize the parameters of the non-bonded contact (i.e. parameters of U_{AH}), which has the Ashbaugh-Hatch (AH) functional form [74]

$$U_{\text{AH}}(r) = \sum_{i<j} \begin{cases} U_{\text{LJ}}(r_{ij}) + (1-\lambda_{ij})\epsilon_{\text{LJ}} & \text{if } r_{ij} \leq 2^{1/6}\sigma_{ij} \\ \lambda_{ij} U_{\text{LJ}}(r_{ij}) & \text{otherwise} \end{cases} \tag{2.6}$$

where U_{LJ} is the Lennard-Jones (LJ) potential

$$U_{\text{LJ}}(r_{ij}) = 4\epsilon_{\text{LJ}} \left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6} \right] \tag{2.7}$$

with $\epsilon_{\text{LJ}} = 0.2$ kcal/mol. λ_{ij} are the hydropathy parameters that depend on the amino acid types. We can view each λ_{ij} as an independent parameter, leading to 210 independent parameters λ_{ij}. Alternatively, we apply the mixing rule $\lambda_{ij} = (\lambda_i + \lambda_j)/2$, which results in 20 independent parameters λ_i. Most existing HPS models use the scheme of 20 independent parameters λ_i [56], [57], [59]. The AH potential is a linear function of λ_{ij} or λ_i, thus contrastive learning leads to a unique solution λ_{ij}^* or λ_i^* given data and noise samples. More details about the IDP models can be found in Appendix A section *Force field of intrinsically disordered proteins.*

We also trained some unified models that can be applied to both IDPs and ordered proteins (OPs). For unified models, IDPs and OPs share the harmonic bond, non-bonded contact, and Debye-Hückel potential, while the OPs have additional harmonic angle, periodic dihedral, and native pair potentials which are defined based on the reference native structures. Thus, the OP potential is

$$U_{\text{OP}} = U_{\text{bond}} + U_{\text{angle}} + U_{\text{dihedral}} + U_{\text{native pair}} + U_{\text{AH}} + U_{\text{elec}} \tag{2.8}$$

The reference native structures are located from the all-atom trajectories by clustering. Native pairs are essentially bonds between atoms that have close contacts in the reference structure and are from the same continuous secondary structure domains. Importantly, the native pairs are only kept between Cα atom pairs within the same continuous secondary structure domains, while native pairs between Cα atoms from different secondary structure domains are removed. Our mission is to correctly specify the contact strengths between different secondary structure domains or disordered regions with the optimized λ_{ij} or λ_i parameters, thus correctly calibrating the non-bonded interaction strengths. More details about ordered protein models and the selection of reference structures can be found in Appendix A section *Force field of ordered proteins* and *Details of noise simulations for ordered proteins.*

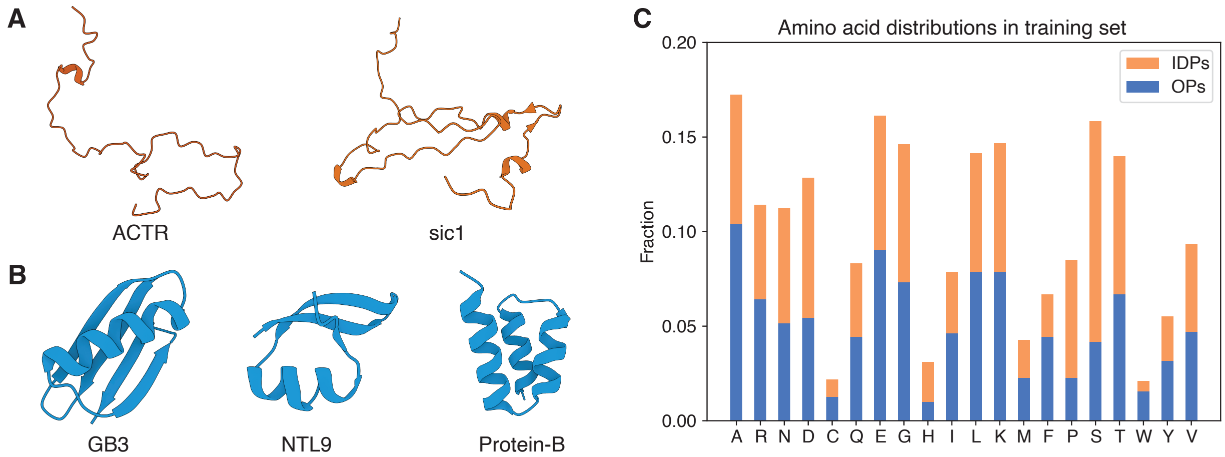


Figure 2.1: **Example structures and non-bonded potential functional form.** (A) Representative structures of two IDPs. (B) Reference structures of three ordered proteins. (C) Fraction of amino acids of training set sequences. The fractions are normalized over all the IDPs or OPs.

2.2.3 Produce noise with existing coarse-grained models

We produced noise samples with existing CG models since we need to evaluate the energy with the noise potential. For IDPs, noise samples were generated with the optimal scale of the HPS-Urry model (i.e. the normalized Urry scale shifted by -0.08) [57] with minor modifications on the bond parameters and electrostatic interaction cutoffs to align with the trained model. For each IDP, multiple independent umbrella simulations [75] were performed with umbrella bias on the radius of gyration (R_g) to enhance sampling. For OPs, the noise samples came from the same model as IDPs with additional harmonic angle, periodic dihedral, and native pair terms defined based on the reference native structures. Note that these angle, diheral, and native pair potentials are the same as the ones in the trained model for OPs (i.e. $U_{\text{angle}}, U_{\text{dihedral}}$, and $U_{\text{native pair}}$ in equation 2.8). Umbrella samplings were performed with umbrella bias applied to the root mean square deviation (RMSD) relative to the same reference native structures. Samples from different umbrella simulations were reweighted with the MBAR equation as a generalized ensemble [76]. More details about noise simulations and reweighting noise samples are provided in Appendix A sections *Details of noise simulations* and *Generate the mixed noise ensemble*. Importantly, with a small training set, we demonstrated that our results are not sensitive to the exact parameters of the noise potential (Figure A.1), indicating that the trained model learns the distribution of the data samples without being overly affected by the noise distribution.

2.3 Results

2.3.1 Hydrophobic scale model trained with a99SB-*disp* trajectories

We started by training a new set of hydropathy parameters with 20 independent parameters λ_i for the HPS model and IDPs. We used the Cα trajectories extracted from a99SB-*disp* all-atom explicit solvent simulation trajectories [72] of 41 IDPs as our training data (Figure 2.1A, 2.1C). a99SB-*disp* is the state-of-the-art all-atom explicit solvent force field for IDPs [72], [77]. By training the model with contrastive learning, we obtained a new set of hydropathy parameters (Figure 2.2A) that can accurately reproduce the R_g of proteins in the training set (Figure 2.2B). We also compared with two other popular Cα IDP force fields, which are HPS-Urry and Mpipi force fields [57], [58]. Previous comparisons have shown that these two force fields perform the best in predicting the average experimental R_g of many IDP sequences [58]. Our comparisons show that HPS-Urry performs similarly to our model in the training set in reproducing the all-atom trajectory average R_g (Figure 2.2B). In the test set, our model approaches the other two models in matching the experimental average R_g (Figure A.3). It is worth mentioning that some CG IDP force fields parameterized with top-down methods, such as the HPS-Urry model, were designed to directly match the experimental average R_g [55], [57], [59]. In contrast, our method aims to learn the distribution as a bottom-up method instead of directly matching R_g. As the model only contains 20 hydropathy parameters and is unlikely to overfit, the discrepancy of the accuracy on the training set and the test set may come from: (1) the all-atom force field is not perfect and

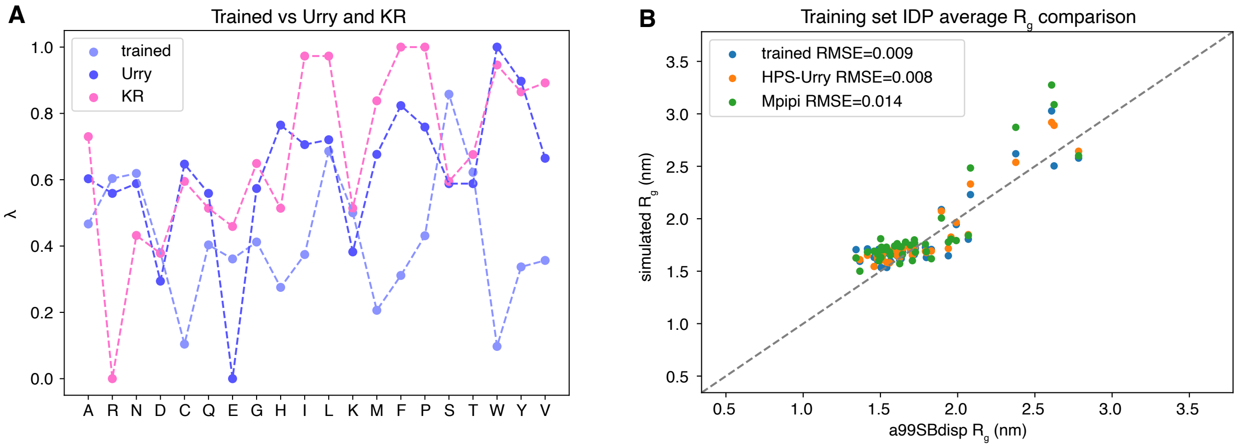


Figure 2.2: **20 hydropathy parameter model trained with a99SB-*disp* all-atom IDP trajectories reproduce the average sizes.** (A) The comparison between the new hydropathy parameter trained with all-atom trajectories with the normalized KR and Urry scales. (B) The comparison of average R_g between several CG models and all-atom trajectories on training set IDPs. The root mean square errors (RMSEs) are provided.

there is intrinsic disparity between all-atom models and experiments; (2) samplings of all-atom simulations may not be sufficient; (3) the sequence diversity is not adequate enough to represent the sequence diversity in the test set. The first two points are challenges of most bottom-up CG force field training methods and call for more accurate all-atom force fields and advanced simulation platforms. Since our results are upper-bounded by the accuracy and sampling ergodicity of the training data, reproducing the data distribution is the main target, while matching experimental average R_g of test set is of second priority. Although we try to balance the abundance of different types of amino acids (see Appendix A section *Details of all-atom simulations* for more details), cystine and tryptophan are lacking in the training set (Figure 2.1C), indicating the intrinsic scarcity of certain amino acids in IDPs [78]. In all, we learned a new set of hydropathy parameters from a99SB-*disp* IDP trajectories and the new parameters lead to results with some state-of-the-art IDP Cα force fields. With increasing amounts of high-quality all-atom data, our method can further improve the CG force field.

2.3.2 Unified model trained with ordered and disordered protein trajectories

We move forward by including 19 ordered proteins into our training set. The ordered protein trajectories were reported by D. E. Shaw Research and simulated with different force fields, including CHARMM22*, a99SB-*disp*, and DES-Amber force fields [71]–[73]. Because 20 independent parameters λ_i can hardly satisfy both ordered and disordered proteins, we proceed with the 210 independent parameters λ_{ij}. Since the results of the 210 parameters λ_{ij} are closely related to the occurrence of each type of contacts in the data and noise samples, while some contacts may be deficient, we train the parameters formulated as a prior parameter plus correction $\lambda_{ij} = \lambda_{ij}^0 + \Delta\lambda_{ij}$, where the prior λ_{ij}^0 comes from the Miyazawa-Jernigan (MJ) contact potential parameters e_{ij} [79]. MJ contact potential parameters e_{ij} are derived based on the contacts within abundant protein crystal structures, thus capturing the driving force of protein folding. As the original e_{ij} values are all negative and the more negative ones mean stronger attractions, while the more positive λ_{ij} values indicate stronger attractions, we set $\lambda_{ij}^0 = \alpha(-e_{ij} + \beta)$, which is effectively shifted and scaled $-e_{ij}$. During training, we fixed α and β as hyperparameters while only optimizing $\Delta\lambda_{ij}$ and Δf with regularizer on $\Delta\lambda_{ij}$. By testing different sets of hyperparameters, we selected the optimal one that achieves best results on both ordered and disordered proteins. We notice that this new set of parameters are within range about 0-2.5 (Figure 2.3A), which is at a larger scale relative to the hydropathy scales for IDPs (Figure 2.2A). Moreover, the new model shows a clear distinction between hydrophobic residues and hydrophilic ones, with very strong attraction when hydrophobic residues are involved. When comparing to training set IDPs, the trained model has similar RMSE to HPS-Urry and Mpipi (Figure 2.3B), while the comparison on training set OPs shows that HPS-Urry and Mpipi tend to overstretch proteins. Note that for a fair comparison on OPs, we added the same structure-based potentials (i.e. angle, dihedral, and native pair potentials) to HPS-Urry and Mpipi as our trained model to

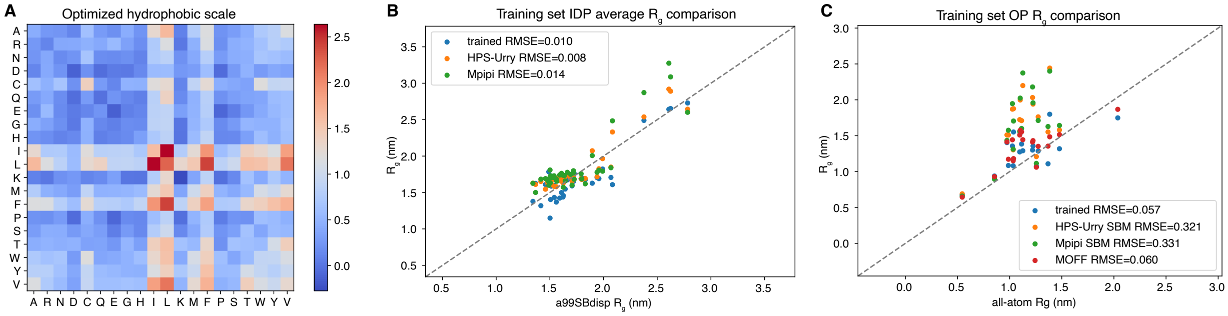


Figure 2.3: **210 hydropathy parameter model trained with ordered and disordered protein all-atom trajectories reproduce the average sizes.** (A) The 210 hydropathy parameters as heat map. (B) The comparison of average R_g between several CG models and all-atom trajectories on training set IDPs. (C) The comparison of average R_g between several CG models and all-atom trajectories on training set OPs. The root mean square errors (RMSEs) are provided.

stabilize secondary structures. We also compared with structure-based model called MOFF [55], which also includes the same structure-based potentials. MOFF achieves great results on reproducing R_g of ordered proteins. The results showed that the inter-residue attraction strengths of state-of-the-art IDP force fields are too weak for OPs, while our trained model achieves a more balanced result for IDPs and OPs.

2.4 Conclusions and Discussion

In this study, we applied contrastive learning to train Cα CG models for IDPs and OPs with all-atom explicit solvent trajectories. For the model optimized only for IDP, a new set of hydropathy scale parameters are derived based on a99SB-*disp* all-atom trajectories (Figure 2.2A). This new set of hydropathy parameters can excellently reproduce the average R_g of all-atom trajectories in the training set (Figure 2.2B). The accuracy on the training set is similar to one of the best top-down Cα CG models, namely HPS-Urry, although our trained model has very different parameters relative to HPS-Urry (Figure 2.2A). We point out that our method is a bottom-up method that learns the data distribution without directly matching R_g. It is a pure bottom-up method that achieves results comparable to state-of-the-art top-down methods, which is an uncommon achievement. However, in the test set, where we compare new IDP sequences with experimental average R_g, HPS-Urry and Mpipi reproduce average R_g more precisely. Such different performance in training and test sets indicates the imperfection of all-atom data [80] and scarcity of certain amino acids in IDP. Although the sequence diversity can be enhanced by accumulating more data, the upper bound of our bottom-up method is the intrinsic quality of the data trajectories. We call for more accurate all-atom models and sufficient data as the foundation for training more accurate CG models with bottom-up methods.

We also trained some unified models to balance the non-bonded interactions within ordered and disordered proteins. With the model parameters based on the scaled MJ potential, we achieved a set of parameters with 210 independent λ_{ij} parameters (Figure 2.3A). This new set of parameters are obviously spanning a larger scale than those IDP-only parameters (Figure 2.2A). The introduction of ordered proteins and MJ as prior clearly changes the interaction pattern and attribute hydrophobic residues as stickers. Such parameters obtain excellent results on both ordered and disordered proteins on the training set, while the discrepancy to the experimental average R_g of IDPs is a bit larger. Considering the abundance of data we used, such 210 parameters model are unlikely to overfit as well, so the gap is likely coming from data quality and insufficient co-occurences of different types of amino acids.

In the future, we aim to train and test some potential with multi-body effects. Such multi-body effects are important for the CG protein models [31]. We will also perform some condensate simulations to compare the capability of different models in reproducing the experimental phase separation behaviors.

Chapter 3

OpenABC enables flexible, simplified, and efficient GPU accelerated simulations of biomolecular condensates

3.1 Introduction

Biomolecular condensates underlie the organization of many cellular processes, such as speckles for RNA splicing, nucleoli for ribosomal RNA processes, and P granule for stress response, etc. [14], [15], [81]–[92]. They are also termed membrane-less organelles due to the lack of enclosure and exhibit liquid-like properties. Intrinsically disordered proteins (IDPs) and RNA molecules are enriched inside the condensates [14], [15], [83], [88]. These molecules promote promiscuous, multivalent interactions, leading to spontaneous phase transition and condensate formation [16]. The nature of the molecular interactions that drive phase separation, the microenvironment of the condensates, and their dynamical relaxation, are under active investigation.

Computational modeling can prove invaluable for studying biomolecular condensates by providing detailed structural and dynamic characterizations [36], [48], [49], [56]–[59], [69],

[93]–[111]. Particle-based coarse-grained modeling approaches are promising since their computational efficiency enables long-timescale simulations to promote large-scale reorganization for structural relaxation [112]–[115]. Such simulations may predict condensate physical properties de novo, elucidating the connection between molecular sequences and emergent properties [106], [116]. However, the reduced resolution of these coarse-grained models could be insufficient to describe the complex microenvironment of the condensate interior [117]–[119]. Atomistic simulations with explicit representation of solvent molecules and counter ions can be necessary to further characterize physicochemical interactions that produce the selective partition of small molecules within condensates [72], [118]–[121]. Combining the two modeling approaches at different resolutions could be particularly powerful since they enable long-timescale simulations for structural relaxation while preserving the fine-resolution details.

While many computational models and force fields have been introduced for simulations of IDPs and biomolecules, software engineering has yet to catch up. There is an urgent need to build user-friendly tools to set up and execute condensate simulations. Preparing biomolecular simulations can be rather involved. Even creating initial configurations for such simulations is often non-trivial. Much-dedicated software has been introduced to prepare atomistic simulations [122]–[125], and existing molecular dynamics (MD) simulation packages are highly optimized for computational efficiency [122], [125]–[127]. However, existing tools are not immediately transferable for setting up coarse-grained condensate simulations. Furthermore, coarse-grained force fields are often implemented into disparate simulation engines not necessarily best suited for condensate simulations, hindering cross-validation and the unleashing of full modeling potential. Further software development can significantly reduce the entry barrier for in silico studies, allowing more researchers to experience the usefulness of computational modeling. They could facilitate comparing and benchmarking various force fields, driving continuous improvement.

We introduce a software package termed OpenABC for "**Open**MM GPU-**A**ccelerated simulations of **B**iomolecular **C**ondensates". The package is flexible and implements multiple popular coarse-grained force fields for simulating proteins and nucleic acids. It dramatically simplifies the simulation setup: only a few lines of Python scripts are needed to carry out condensate simulations starting from initial configurations of a single protein or DNA. The package is integrated with OpenMM, a GPU-accelerated MD engine [62], enabling efficient simulations with advanced sampling techniques. Finally, we include tools that convert coarse-grained configurations to atomistic structures for further condensate modeling with all-atom force fields. Tutorials in Jupyter Notebooks are provided to demonstrate the various capabilities. We anticipate that OpenABC will greatly facilitate the application of existing computer models for simulating biomolecular condensates and the continued force field development.

3.2 Methods

3.2.1 Details of molecular dynamics simulations

We performed temperature replica-exchange simulations [128] with MOFF to determine the conformational ensembles of HP1α and HP1β dimers. Atomistic protein structures were predicted with RaptorX [129] and used to initialize simulations. Details on modeling HP1 proteins to preserve the tertiary structure of folded domains are provided in Appendix B *Setting up MOFF HP1 system.* Six independent replicas were simulated to maintain temperatures at 300 K, 315.79 K, 333.33 K, 352.94 K, 375.00 K, and 400 K, respectively, with the Langevin middle integrator [130] and a friction coefficient of 1 ps^{-1}. Each replica lasted for 200 million steps with a timestep of 10 fs. Exchanges between neighboring replicas were attempted every 1000 steps. More details on the replica exchange simulations are attached in Appendix B *Implementation of the temperature replica exchange algorithm.* We discarded the first 100 million steps as equilibration and used the remaining data for analysis.

We carried out slab simulations to evaluate the stability of condensates formed by HP1α

and HP1β dimers. Initial configurations of these simulations were prepared as follows. First, we randomly placed 100 copies of protein dimers into a cubic box of length 75 nm. Then we performed 5-million-step constant pressure and constant temperature (NPT) simulations at one bar and 150 K to compress the system with a timestep of 10 fs. Control of pressure and temperature was achieved by coupling the Monte Carlo barostat with the Langevin middle integrator [130]. The length of the compressed cubic box was about 25 nm. Then we fixed the compressed configuration and extended the box size to 25 × 25 × 400 nm^3. The rectangular geometry leads to the creation of a dense-dilute interface along the z-axis. Simulation results are expected to be independent of the exact box lengths and we chose 400 nm to be long enough to support phase coexistence (Figure B.2). Starting from this initial configuration, we gradually increased the temperature from 150 K to a target value in the first 0.1 million steps. We then performed 200-million-step production simulations at constant volume and constant temperature using the Nosé-Hoover integrator [130] with a collision frequency of 1 ps^{-1} and a timestep of 5 fs. Compared to the Langevin thermostat, the Nosé-Hoover integrator allows faster diffusion of protein molecules in the dilute phase to facilitate the equilibration of slab simulations.

Following similar protocols outlined above, we performed slab simulations for disordered regions of protein DDX4 and FUS with the HPS model using parameters derived from the Urry hydrophobicity scale [131]. Detailed amino acid sequences of the two proteins are provided in Appendix B. For each protein, we first obtained an equilibrium configuration from a 0.1-million-step constant temperature simulation initialized with a straight C$_\alpha$ chain. We placed 100 replicas of the equilibrium configurations into a cubic box of length 75 nm. Upon compression by a 5-million-step NPT compression at 1 bar and 150 K with a timestep of 10 fs, the system reaches a cubic box with a size of about 15 nm. We then performed slab simulations with an elongated box of size 15 × 15 × 280 nm^3 and a 10 fs timestep. Nosé-Hoover integrator was again applied with a collision frequency of 1 ps^{-1} to maintain the temperature.

3.2.2 Computing phase diagrams from slab simulations

To determine the concentration of dense and dilute phases from slab simulations, we first identified the largest cluster in a given configuration as the largest connected component of the protein-contact network. Two monomers were defined as in contact if their center-of-mass distance was less than 5 nm, though the computed phase diagrams are rather insensitive to this specific cutoff value (Table B.15). Subsequently, we translated the system so that the center of mass of the largest cluster coincides with the box center, which was located at $z = 0$. We recognized the region with $|z| < 5$ nm for HPS simulations and $|z| < 10$ nm for MOFF simulations as the dense phase, while the region with $|z| > 50$ nm as the dilute phase. The threshold values were chosen to be consistent with prior literature [55], [56] and to roughly follow the size of the condensate as revealed in the density profiles (Figures B.2, B.3). The concentrations were determined as the average density value in specified regions using the second half of the simulation trajectories. We fitted the concentration values at various temperatures using the following equation to determine the critical temperature

$$\rho_{\mathrm{H}} - \rho_{\mathrm{L}} = A(T_{\mathrm{c}} - T)^{\beta}. \tag{3.1}$$

ρ_{H} and ρ_{L} are the densities at the concentrated and dilute phases. Parameter $\beta = 0.325$ is the critical exponent corresponding to the universality class of 3D Ising model [132]. T_{c} is the critical temperature and A is the coefficient.

3.3 Results

3.3.1 Flexible force field selections for Biomolecular simulations

OpenABC implements several existing force fields for coarse-grained (CG) modeling of protein, RNA, and DNA molecules (Fig 3.1). Single-bead per amino acid force field for proteins

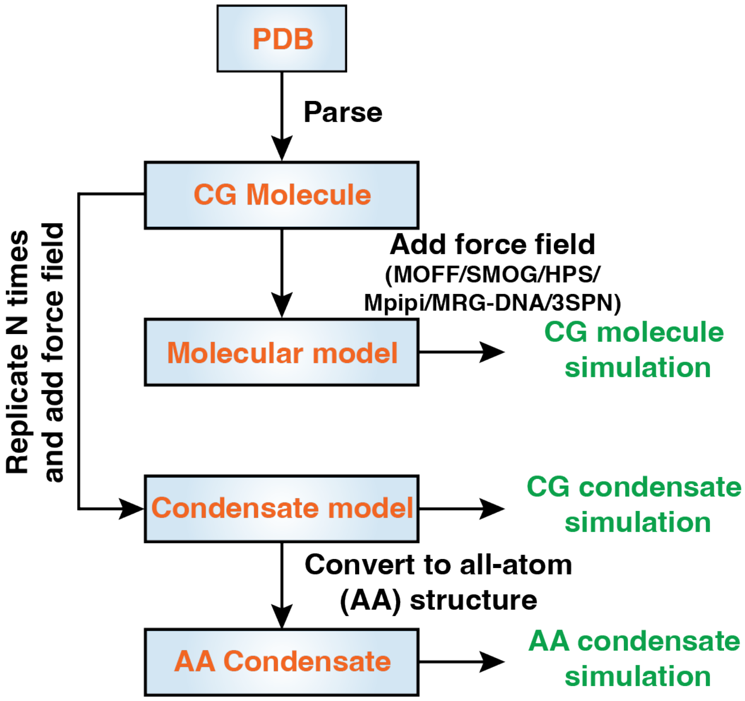


Figure 3.1: **OpenABC facilitates coarse-grained and atomistic simulations of biomolecular condensates with multiple force fields.** The diagram illustrates the workflow and various functionalities of OpenABC. To set up condensate simulations, the users must provide a configuration file in the PDB format for the molecule of interest. OpenABC parses topological and structural information from the PDB file to build a molecule object. Specifying force field options allows direct simulations of individual molecules. On the other hand, the molecule object can be replicated for condensate simulations. In addition, OpenABC allows the conversion of CG configurations to atomistic structures for simulations with all-atom force fields.

include the hydropathy scale (HPS) models [56], [57], the Mpipi force field [58], a generalized structure-based model [46], [49], [133], and the maximum entropy optimized force field (MOFF) [55]. HPS models define interactions between different pairs of amino acids based on various hydrophobicity scales [56], [57]. Recent studies have attempted to improve the accuracy of HPS models with systematic optimizations of the hydrophobicity scale to match experimental observations of IDP monomers [59], [99]. They have been used to study the phase behaviors of numerous proteins [134]–[136], revealing the contribution of charge distribution patterns, cation-π interactions, and the balance between hydrophobic and electrostatic interactions [135], [137] to the stability of condensates.

The Mpipi force field was parameterized using data from all-atom simulations and bioinformatics analysis with a careful calibration of π-π and π-cation interactions [58]. These interactions play significant roles in the formation of biomolecular condensates. The force field was shown to accurately capture the radius of gyration and critical temperatures of diverse protein sequences.

SMOG was originally introduced for studying folded proteins using interaction potentials derived from initial input configurations. We generalized the model to describe proteins with disordered domains and leveraged the Miyazawa-Jernigan statistical potential [79] for protein-protein interactions [46], [49].

MOFF was parameterized with the maximum entropy algorithm [138], [139] and the protein folding energy landscape theory [140] to provide consistent descriptions of both folded and disordered proteins [55], [141]–[143]. It was shown to reproduce the radius of gyration for a collection of proteins, including both ordered and disordered proteins [55], [144]. The balanced interactions among amino acids have proven beneficial in describing complex contacts among phase-separating proteins, including those with both ordered and disordered domains [55], [116], [143].

In addition to protein models, we implemented several force fields for nucleic acids. For example, the molecular renormalization group coarse-graining (MRG-CG) DNA model was initially introduced for simulations with explicit ions to reproduce the salt-dependent DNA persistence length [145]. We adopted it for implicit ion modeling with the Debye-Hückel approximation for electrostatic interactions. We rescaled the strength of bonded interactions to ensure the accuracy of the implicit-ion model in reproducing DNA persistence length at the physiological salt concentration [116]. We further incorporated the DNA model 3SPN [146], [147] into OpenABC for studying sequence specific properties. Unlike MRG-CG DNA that only uses one bead to represent each nucleotide, 3SPN adopts three beads to differentiate sugar, base, and phosphate. Finally, the Mpipi force field can be used to simulate RNA molecules.

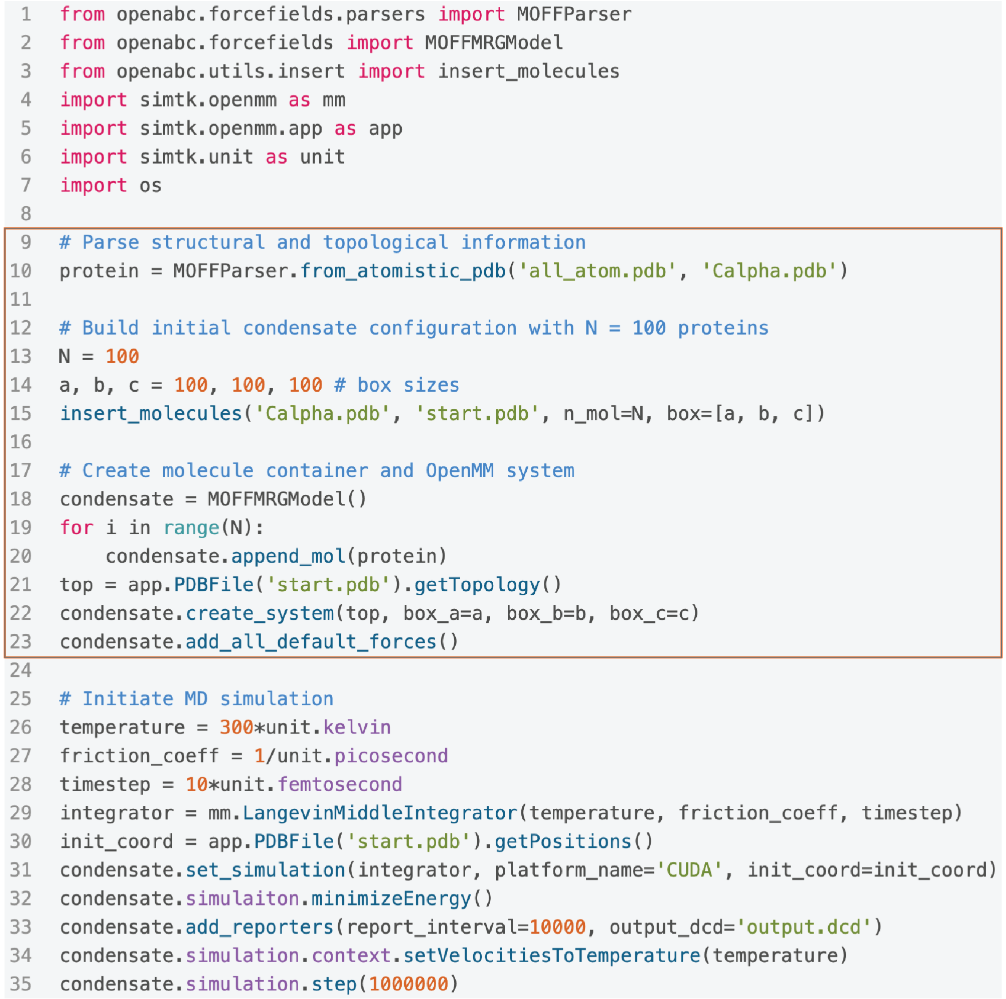
While one can in principle combine different force fields for simulating complex systems with both proteins and nucleic acids, care needs to be taken when modeling cross interactions. Previous studies have carried out systematic validations of protein-DNA and protein-RNA interactions and we implemented them into OpenABC, with combinations that include SMOG-3SPN [46], [49], [50], [102], [147]–[150], MOFF-MRG-CG DNA [116], and Mpipi Protein-RNA [58]. These combinations account for both excluded volume effect and electrostatic interactions. Detailed expressions of all the force field potentials are provided in Appendix B *Force Field Definitions*, with the parameters provided in Tables B.1-9.

3.3.2 Simplified Setup of Condensate Simulations

OpenABC leverages the MD simulation engine, OpenMM [62], to offer simulation setup with Python scripting, thus dramatically simplifying the workflow. The software treats each molecule as an object and appends such objects into a container-like class. This class allows the incorporation of various force field options and integration schemes for MD simulations.

An illustration of the typical workflow for condensate simulations is provided in Fig 3.1. OpenABC first parses a configuration file in the PDB format supplied by users to create a molecule object. The object contains topological and structural information extracted from the input file. Upon introducing interactions defined in various force fields, the molecule object can be used to simulate individual biomolecules. On the other hand, the molecule object can also be replicated N times for condensate simulations consisting of N molecules. As demonstrated in an example code in Fig 3.2, setting up an entire MD simulation of a protein condensate with default parameters only requires about 20 lines of code.

To enhance conformational sampling of individual molecules and condensates, we provide an implementation of the temperature replica exchange algorithm [128] with PyTorch [151] as part of the package (see Appendix B *Implementation of the temperature replica exchange algorithm* for details). Furthermore, we introduce utility functions to reconstruct atomistic structures from coarse-grained protein configurations with only α carbons. This functionality



```
from openabc.forcefields.parsers import MOFFParser
from openabc.forcefields import MOFFMRGModel
from openabc.utils.insert import insert_molecules
import simtk.openmm as mm
import simtk.openmm.app as app
import simtk.unit as unit
import os

# Parse structural and topological information
protein = MOFFParser.from_atomistic_pdb('all_atom.pdb', 'Calpha.pdb')

# Build initial condensate configuration with N = 100 proteins
N = 100
a, b, c = 100, 100, 100 # box sizes
insert_molecules('Calpha.pdb', 'start.pdb', n_mol=N, box=[a, b, c])

# Create molecule container and OpenMM system
condensate = MOFFMRGModel()
for i in range(N):
    condensate.append_mol(protein)
top = app.PDBFile('start.pdb').getTopology()
condensate.create_system(top, box_a=a, box_b=b, box_c=c)
condensate.add_all_default_forces()

# Initiate MD simulation
temperature = 300*unit.kelvin
friction_coeff = 1/unit.picosecond
timestep = 10*unit.femtosecond
integrator = mm.LangevinMiddleIntegrator(temperature, friction_coeff, timestep)
init_coord = app.PDBFile('start.pdb').getPositions()
condensate.set_simulation(integrator, platform_name='CUDA', init_coord=init_coord)
condensate.simulaiton.minimizeEnergy()
condensate.add_reporters(report_interval=10000, output_dcd='output.dcd')
condensate.simulation.context.setVelocitiesToTemperature(temperature)
condensate.simulation.step(1000000)
```

Figure 3.2: **OpenABC simplifies simulation setup with Python scripting.** The example code includes all the steps necessary for setting up and performing MD simulations of a protein condensate with MOFF and default settings in a cubic box of length 100 nm. The ten lines included in the highlight box correspond to the creation of the condensate system by parsing topological information from an initial PDB file, building a configuration file by inserting molecules into a box and incorporating the molecular objects, *protein*, into a container class, *condensate*, with appropriate force fields. The rest of the code includes standard simulation setups generic to OpenMM. We chose the Langevin middle integrator to perform simulations at 300 K with a friction coefficient of 1 ps^{-1} and a timestep of 10 fs.

relies on the software “reconstruct atomic model from reduced representation (REMO)” [152] and can facilitate downstream all-atom simulations. More tutorials in Jupyter Notebook format are available online at the OpenABC GitHub repository.

3.3.3 Efficient simulations with GPU-enabled MD engine

A significant advantage of integrating with OpenMM comes from its native support of GPU acceleration. Simulating implicit solvent coarse-grained condensates on GPUs can be particularly beneficial due to the inhomogeneous distribution of particles arising from implicit solvation [147]. CPU parallelization, which often relies on the spatially-based, domain decomposition strategy, is often less effective because the inhomogeneity in particle density between the condensate and dilute phases produces an imbalanced workload between CPUs.

To demonstrate the efficiency of GPU-enabled simulations, we studied five independent condensate systems. The first four systems consist of N_1 HP1α dimers and N_2 200-bp-long dsDNA randomly distributed in a cubic box of length 200 nm with periodic boundary conditions. In the fifth system, 100 HP1α dimers in a compact configuration were placed at the center of an elongated box of size $25 \times 25 \times 400$ nm^3 (Fig 3.3A). This rectangular setup is typical for the so-called slab simulations to produce a dilute and dense interface along the z-axis for computing co-existence curves and phase diagrams [56], [153], [154]. MOFF and MRG-CG force fields were used to describe the interactions among coarse-grained particles. We simulated each system for one million steps using the Langevin middle integrator [130] to control the temperature at 300 K, with a friction coefficient of 1 ps^{-1} and a time step of 10 fs. For comparison, we simulated the same systems with a closely related integrator using GROMACS, a leading MD engine with state-of-the-art performance on CPUs [122], [125]. More simulation details are provided in Appendix B *Benchmarking the performance of condensate simulations.*

As shown in Fig 3.3B, OpenMM single GPU performance matches GROMACS with hundreds of CPUs in the first four systems. While GROMACS achieved nearly linear scaling for the first four systems, introducing more CPUs did not lead to any significant speedup in the last system with a dense-dilute interface. As mentioned above, the presence of vacuum regions in slab simulations hinders the efficacy of domain decomposition. On the other hand,

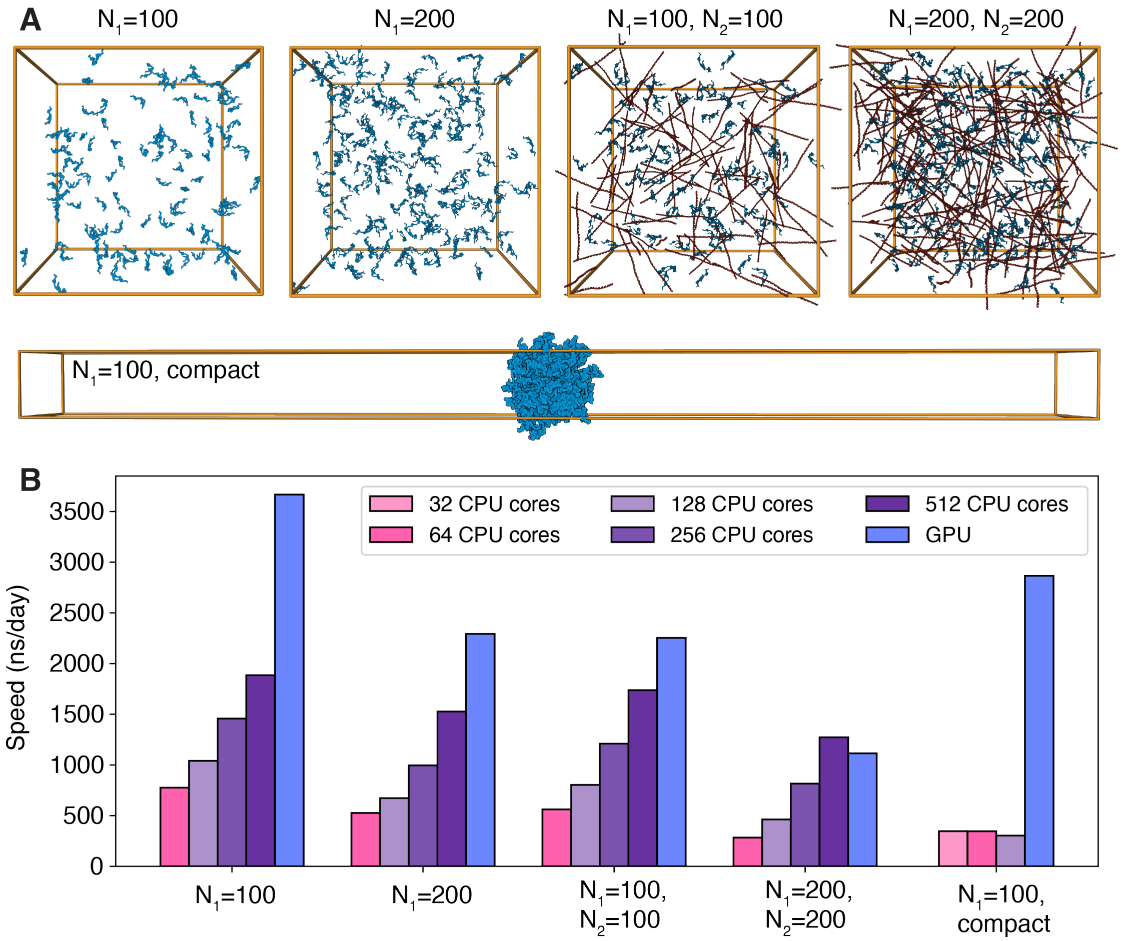


Figure 3.3: **OpenABC integrates with OpenMM for GPU-accelerated MD simulations.** (A) Snapshots of the five systems used to benchmark simulation performance. The systems consist of N_1 HP1α dimers (blue) and N_2 200-bp-long dsDNA (red, $N_2 = 0$ if not specified). The first four systems adopt homogeneous density distributions in cubic boxes of length 200 nm, while the last exhibits a dense-dilute interface in an elongated box of size $25 \times 25 \times 400$ nm^3. (B) The five data sets compare the performance of CPU simulations using GROMACS with single GPU simulations using OpenMM. The different colors indicate the number of CPUs in GROMACS simulations, as shown in the legends. The benchmarks were performed with Intel Xeon Gold 8260 CPUs and Nvidia Volta V100 GPUs.

OpenMM is less sensitive to the simulation setup and retains superior performance.

The performance of GROMACS depends on our implementation of the CG force fields and may not reflect the theoretical upper limit of the software. In particular, our use of tabulated potentials for the Debye Hückel potential and domain decomposition for parallelization may significantly affect the simulation speed. While performance improvement is possible with additional software engineering, the advantage of CG simulations of condensates on GPUs remains given the differences shown in Fig 3.3B.

3.3.4 Application: Validating force field implementations in OpenABC

Before applying the software for extensive simulations, we validated our implementations of various force fields with existing ones. We generated ten configurations for an HP1α dimer with MOFF and GROMACS through an NVT simulation. As shown in Table B.10, the potential energies evaluated using MOFF from OpenABC match those reported by GROMACS. Similar comparisons with a protein-DNA complex produce nearly identical energy values as well, as shown in Table B.11. The protein-DNA complex is formed by an HP1α dimer with a 200-bp-long dsDNA, and MOFF and MRG-CG DNA were used to quantify their interactions. The minor differences between OpenMM and GROMACS energies are mainly caused by using tabulated functions for nonbonded interactions in GROMACS.

We further evaluated the potential energies defined by the HPS model on ten configurations of a disordered protein, DDX4, using both OpenMM and HOOMD-Blue [155]. As shown in Table B.12, the two sets of energies match exactly, supporting the correctness of our force field implementation. We also validated the Mpipi force field using interaction energies evaluated with OpenMM and LAMMPS [127] for a protein-RNA system, as shown in Table B.13.

In addition to energy comparisons, we examined the conformational ensembles of HP1α and HP1β dimers using MOFF with temperature replica exchange simulations [128]. Consistent with our previous study [55], the force field succeeds in resolving the difference in their conformational distribution between the two homologs (Fig 3.4). The radii of gyration for the two dimers at 300 K are 3.33 ± 0.19 nm, and 4.27 ± 0.09 nm, respectively. These values match the previously reported values computed using GROMACS quantitatively, reproducing experimental trends. Therefore, OpenABC produces consistent results with other software despite differences in integration schemes.

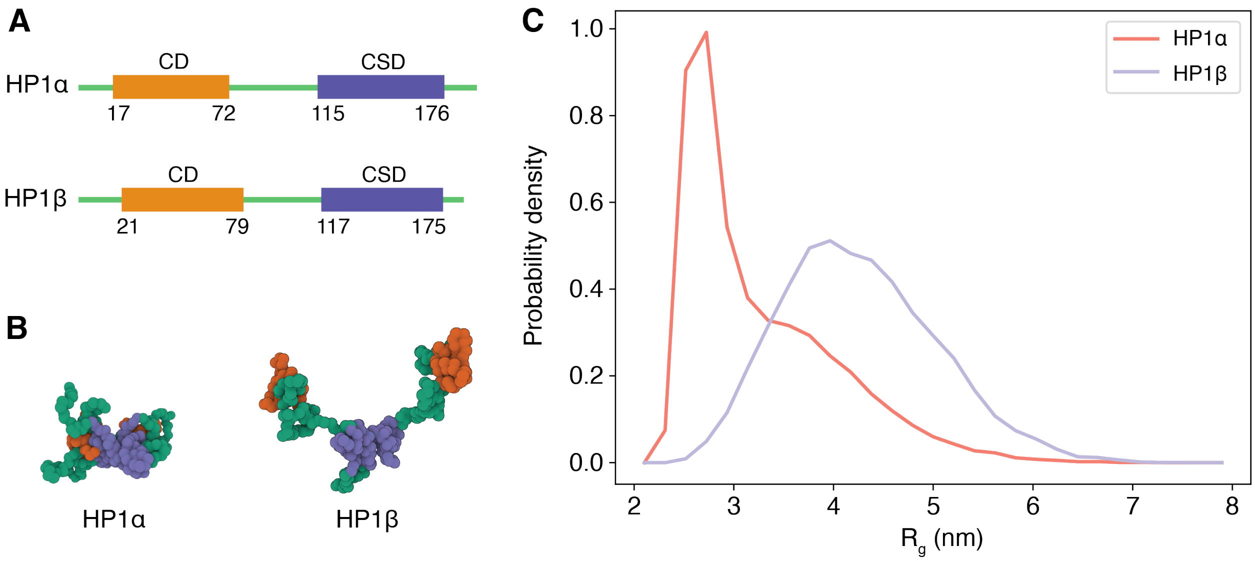


Figure 3.4: **OpenABC produces consistent results with a previous studying, resolving the structural differences between two HP1 homologs.** (A) Secondary structures of HP1α and HP1β along sequences. (B) Representative structures for HP1α and HP1β dimer rendered with Mol* Viewer [156]. The radii of gyration (R_g) for the two structures are 2.77 and 4.44 nm, respectively. We colored the chromodomain (CD) in orange, the chromoshadow domain (CSD) in blue, and the rest in green. (C) Probability density distributions of R_g for HP1α (red) and HP1β dimer (blue).

Using the MRG-DNA model, we computed the persistence length of a 200-bp-long DNA segment. The estimated value at a monovalent salt concentration of 100 mM, 48.83 ± 2.71 nm (see Figure B.1), is consistent with that reported in a previous study using simulations of the same model but with GROMACS [116]. Additional simulation details for estimating the persistence length are provided in Appendix B *Estimating the persistence length of MRG-DNA*

3.3.5 Application: Coarse-grained simulation of protein condensates

As additional evaluations of force field implementation and to demonstrate the usefulness of OpenABC, we performed slab simulations to determine the phase diagram of four proteins, which are known to form various biomolecular condensates inside the cell. For example, HP1

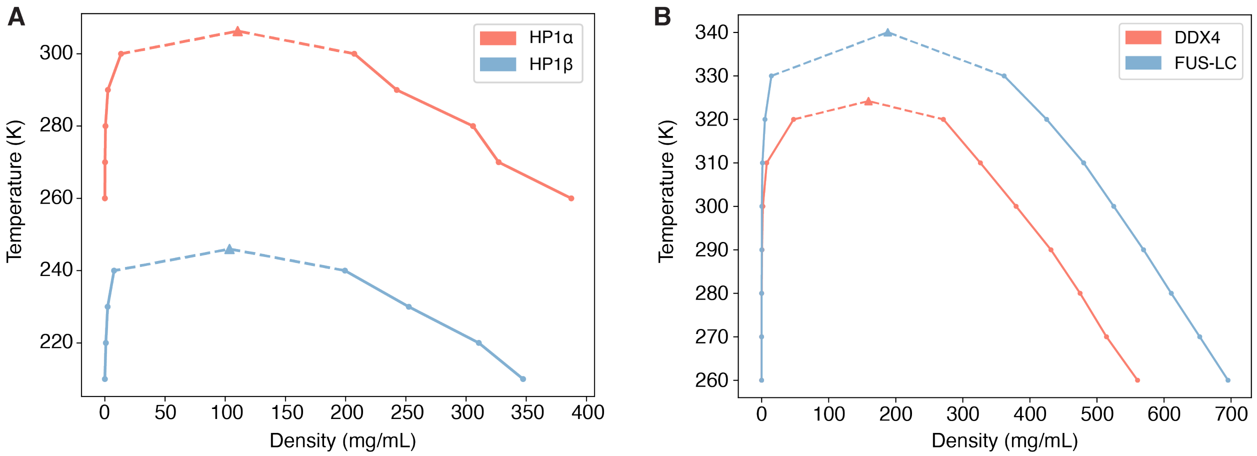


Figure 3.5: **OpenABC produces phase diagrams that match previous results.** (A) Phase diagrams for HP1α (red) and HP1β (blue) dimer condensates computed with MOFF. (B) Phase diagrams for DDX4 (red) and FUS LC (blue) computed with the HPS model parameterized using the Urry hydrophobicity scale. The dots in both plots denote the density values determined from slab simulations, and the triangles represent the critical point obtained from numerical fitting.

dimers are involved in chromatin compaction and regulation [157], while DDX4 and FUS are a primary constituent of nuage or germ granules [158] and cytoplasmic RNP granules [159], respectively. The simulations for HP1α and HP1β were performed with MOFF, while those for FUS LC and DDX4 were modeled with the HPS model using the shifted Urry hydrophobicity scale [57].

The resulting phase diagrams are shown in Fig 3.5, with the concentrations listed in Tables B.14-16. The density profiles at different temperatures and the representative snapshots at the lowest temperatures are shown in Figures B.2-3. We fitted the computed phase diagrams with an analytical expression to determine the critical temperature T_c (see Methods). The critical temperatures are 306.30 K for HP1α and 245.99 K for HP1β, consistent with previous results obtained with GROMACS simulations [55]. Similarly, the critical temperatures for DDX4 and FUS LC are 324.21 K and 340.04 K, respectively, matching values reported in a previous study that used the software HOOMD-Blue for simulations [57]. Thus, OpenABC produces statistically indistinguishable results on the phase behavior of protein condensates as in previous studies.

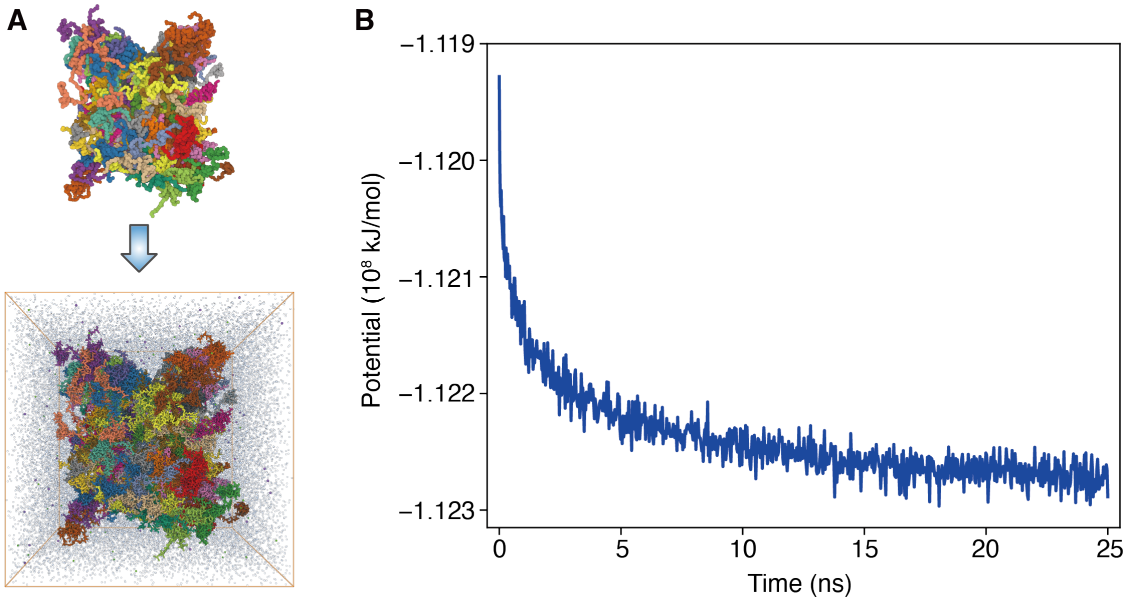


Figure 3.6: **OpenABC facilitates all-atom simulations by producing equilibrated initial atomistic configurations.** (A) Illustrations of the conversion from a coarse-grained configuration (top) to a fully atomistic model with explicit solvent molecules (bottom). Only 2% of water molecules and counter ions of the atomistic model are shown for clarity. The system consists of 100 HP1α dimers, and different molecules are shown in one of 25 colors. Both figures are rendered with Mol* Viewer [156]. (B) The atomistic potential energy evaluated using the CHARMM force field is shown as a function of simulation time.

3.3.6 Application: Atomistic simulation of protein condensates

While residue-level CG models are helpful for long timescale simulations, their limited resolution may prove insufficient to characterize specific properties of condensates, including the solvation environment [117], counter-ion distributions [118], and protein-ligand interactions [119]. Therefore, we implemented functionalities in OpenABC to convert equilibrated CG configurations to atomistic structures. Starting from these structures, well-established tools, such as CHARMM-GUI [123], GROMACS [122], [125], and AMBER [124], can be easily applied to set up explicit solvent simulations with diverse force fields. Furthermore, for explicit solvent simulations, the advantage of OpenMM over other MD packages is less evident. Therefore, we terminate the OpenABC workflow at producing atomistic condensate structures and leave the users with flexibility to choose MD packages and force fields for further studies.

As proof of principle, we converted the final snapshot from the slab simulation of HP1α dimer at 260 K to an atomistic configuration (Fig 3.6). This conversion leverages the software REMO [152] to build atomistic details starting from the Cα positions of each amino acid. We solvated the atomistic HP1α dimer condensates with water molecules and counter ions. After energy minimization, we carried out an all-atom MD simulation using GROMACS with the CHARMM36m force field [160] and the CHARMM-modified TIP3P water model [161]. More details on simulation preparation can be found in Appendix B *Building and relaxing atomistic structures from coarse-grained configurations*. As shown in Fig 3.6, the system relaxes with a continuously decreasing potential energy in the first 20 ns and remains stable afterward.

3.4 Conclusion

We introduced a software package, OpenABC, to facilitate coarse-grained and all-atom simulations of biomolecular condensates. The package implements several of the leading coarse-grained force fields for protein and DNA molecules into OpenMM, enabling GPU-accelerated simulations with performances rivaling GROMACS simulations with hundreds of CPUs. New force fields can be quickly introduced within the framework, and we plan to incorporate RNA models into the package as the next step. Comprehensive tutorials are provided to familiarize the users with the various functionalities offered by OpenABC. We anticipate the intuitive Python interface of OpenABC to reduce entry barriers and promote coarse-grained modeling for its adoption by a broader community.

Chapter 4

Chromatin fiber breaks into clutches under tension and crowding

4.1 Introduction

Eukaryotic genomes are packaged into nucleosomes by wrapping DNA around histone proteins. While the structure of a single nucleosome has been extensively characterized [3]–[5], the organization for a string of nucleosomes, i.e., chromatin, remains debatable [162]–[164]. Regular, fibril configurations are commonly observed in experiments that study chromatin materials extracted from the nucleus [165]–[168]. The invention of in vitro reconstituted nucleosome arrays with strong-positioning DNA sequences [169] helped to remove sample heterogeneity in nucleosome spacing and made possible the determination of high-resolution structures [9], [170]–[173]. However, despite the large amount of evidence supporting their formation in vitro, fibril structures are rarely detected by in vivo experiments that have managed to characterize chromatin at a fine resolution [10], [12], [13], [174]. Therefore, their biological relevance has been questioned, and chromatin organization inside the nucleus remains controversial.

It is worth noting that the nuclear environment is rather complex. In addition to interactions among nucleosomes, many factors, including tension and crowding, can impact chromatin organization. Chromatin is known to associate with various force-generating protein molecules involved in transcription and nucleosome remodeling [175]–[179]. Furthermore, chromatin is often attached to the nuclear envelope and other liquid droplet-like nuclear bodies [96], [180]–[182]. Dynamical fluctuations in these nuclear landmarks could exert forces on chromatin as well [183], [184]. Finally, local nucleosome density can be quite high, especially in heterochromatin regions [60]. Such a crowded environment could lead to cross-chain contacts that might compete with interactions stabilizing single-chain conformations [10]. Therefore, both tension and crowding could destabilize the most stable configuration for isolated chromatin, driving chromatin unfolding and the formation of irregular structures.

Chromatin unfolding has indeed been studied extensively with various techniques [185], [186]. Single-molecule force spectroscopy is a powerful tool for characterizing chromatin organization under tension [50], [187]–[189]. Force-extension curves at low-force regimes are particularly informative regarding inter-nucleosomal interactions [190]. Single-molecule Förster resonance energy transfer is another popular technique for probing nucleosome contacts and chromatin conformational dynamics [24], [191]–[193]. Mesoscopic modeling has also been frequently used to interpret experimental data with structural details [194]–[201]. However, because of the experimental techniques' low resolution and assumptions on nucleosome-nucleosome interactions introduced in computational models, the exact conformations of unfolded chromatin have not reached a consensus and necessitates further investigations.

We perform computer simulations of a 12-nucleosome-long chromatin segment (12mer) to investigate chromatin unfolding under tension and crowding. Residue-level coarse-grained representations are adopted for protein and DNA molecules to capture their interactions with physical chemistry potentials at high resolution. Using a combination of enhanced sampling techniques and machine learning, we show that the computed force-extension curve agrees well with results from single-molecule force spectroscopy experiments [202]. Our simulations support chromatin unfolding under tension proceeds through intermediate structures with nucleosome clutches, i.e., configurations that have been directly observed via super-resolution imaging of cell nucleus [203]. These structures sacrifice nucleosomal DNA by unwrapping to preserve close contacts among neighboring nucleosomes. In addition, the presence of another 12mer promotes inter-chain interactions to stabilize extended chromatin configurations as well. Together, our results suggest that in vivo chromatin configurations can arise from the unfolding of fibril configurations as a result of tension and crowding.

4.2 Methods

4.2.1 Coarse-grained modeling of chromatin organization

We applied a coarse-grained model to study a chromatin segment with twelve nucleosomes. The structure-based model [133], [204] was used to represent protein molecules with one bead per amino acid and stabilize the tertiary structure of the histone octamer while maintaining the conformational flexibility of disordered tail regions. Secondary structure motifs in the disordered regions of histone proteins do not impact nucleosome stability and protein-DNA interactions (Figure C.1) and were not explicitly accounted for in the model. Protein molecules from different nucleosomes interact through both an electrostatic and amino acid-specific potential [205]. We represented the DNA molecule with three beads per nucleotide using the 3SPN.2C model [206]. Protein-DNA interactions were described with the screened Debye-Hückel potential at a salt concentration of 150 mM and the Lennard-Jones potential for excluded volume effect. We ignored electrostatics interactions for particles that are farther than four times the screening length (3.14 nm), at which point the Debye-Hückel potential is expected to become negligible. Further increasing the cutoff length does not quantitatively impact simulation results (Figure C.2). The coarse-grained model has been used extensively in prior studies with great success to investigate protein-protein/protein-DNA interactions [55], [102], the energetics of single nucleosome unwinding [148], [150], nucleosome-nucleosome interactions [207], and the folding pathways of a tetra-nucleosome [46]. More details on the model setup and validation and force field parameters can be found in the Appendix C.

The software package LAMMPS [208] was used to perform molecular dynamics simulations with periodic boundary conditions and a time step of 10 fs. The length of the cubic simulation box was set as 2000 nm, which is much larger than the maximum chromatin extension length to prevent interactions between periodic images. We used the Nosé-Hoover style algorithm [209] to maintain the temperature at 300 K with a damping constant of 1 ps. The globular domains of histone proteins and the inner layer of nucleosomal DNA were rigidified. Positions and velocities of all the atoms within each rigid body were updated together such that the body moves and rotates as a single entity. Disordered histone tails, outer nucleosomal DNA, and linker DNA remained flexible, and no restrictions were applied to their conformational dynamics. Our partition of the rigid and flexible parts was motivated by prior studies showing that unwinding the inner layer of nucleosomal DNA does not occur at forces below 5 pN [190], [202]. Furthermore, the histone core remains relatively stable during DNA unwinding [210], [211]. As shown in Figure S11 of [46], this treatment does not impact the accuracy in sampling inter-nucleosome interactions but significantly reduces the computational cost. Rigidizing inner layer DNA with the globular domains of histone proteins does not affect the energetics of outer layer DNA unwrapping either (Figure C.3).

4.2.2 Force extension curves from enhanced sampling

To characterize chromatin structures under tension and compute force-extension curves, we introduced two collective variables that monitor the important degrees of freedom for chromatin unfolding. The first variable, d_{stack}, measures the average geometric center distance between the i-th and $(i+2)$-th nucleosomes. For small values of d_{stack}, nucleosomes are stacked on top of each other as in the zigzag conformation [8], [9]. The second variable, q_{wrap}, quantifies the average degree of nucleosome unwrapping. The two variables can better differentiate the various chromatin conformations and capture the energetic cost of extension than the DNA end-to-end distance. Mathematical expressions for the two variables are provided in the Appendix C.

For extension forces below 3 pN, we carried out a set of two-dimensional umbrella sampling based on q_{wrap} and d_{stack}. q_{wrap} was restricted to centers from 0.45 to 0.90 with a spacing of 0.15 and a spring constant of 50.0 kcal/mol. d_{stack} was limited to centers from 10.0 nm to 30.0 nm with a spacing of 5 nm and a spring constant of 0.05 $\text{kcal}/(\text{mol} \cdot \text{nm}^2)$. Additional simulations were added to improve the overlap between umbrella windows and the convergence of free energy calculations. When extension forces are larger than 3 pN, chromatin can adopt fully unstacked structures with large end-to-end distances. Covering the entire accessible phase space with two-dimensional umbrella simulations becomes too costly computationally. Therefore, we restricted to one-dimensional free energy calculations using d_{stack} as the collective variable.

Most simulations were initialized from the most probable configurations predicted by a neural network model for chromatin stability under the same umbrella biases (see below and Appendix C for details). They lasted for at least 10 million steps. Details of the umbrella centers and spring constants used in simulations and exact trajectory lengths are provided in Table C.1. We computed the error bars by dividing the data into three equal-length, non-overlapping blocks and calculated the respective quantities using data from each block. The standard deviations of the three estimations were used to measure the errors of the mean.

4.2.3 Facilitating conformational sampling with a neural network model for chromatin

Conformational sampling for the 12mer is challenging because of the many possible degenerate configurations. For example, both unstacking and unwrapping can extend chromatin, and different combinations of the two from various nucleosomes can result in many structures that share similar end-to-end distances. Conformational transitions are slow due to considerable energetic barriers arising from non-specific electrostatic interactions.

To alleviate the sampling problem, we introduced a neural network model for the 12mer. As detailed in the Appendix C, the model quantifies the stability and the free energy of chromatin structures using inter-nucleosome distances (Figure C.4). It was parameterized using mean forces estimated with coarse-grained simulations for 10,000 independent tetra-nucleosome configurations [46]. The neural network model is computationally efficient and allows exhaustive Monte Carlo sampling to determine the most likely chromatin structures at a given setup. These structures were provided to initialize coarse-grained simulations and free energy calculations.

The neural network model is imperfect due to approximations introduced when building the free energy surface with tetra-nucleosome calculations. However, it does reproduce the force-extension curve reasonably well at the lower force regime (Figure C.5). We only used the neural network model for conformational exploration, and all quantitative results presented in the manuscript were obtained with coarse-grained simulations.

4.2.4 Exploring the impact of crowding on chromatin extension

To study the effect of crowding on chromatin organization, we computed the free energy profile as a function of two collective variables that measure intra- and inter-chain contacts. Umbrella sampling was used to enhance conformational exploration, and details on the restraining centers and constants are provided in Table C.2.

Umbrella simulations were initialized from configurations in which the two chains were separated far apart with zero contacts. For simulations biased toward small values of $\bar{d}_{\text{stack}} < 10$ nm, we prepared each chromatin with a two-helix zigzag configuration that resembles the cryo-EM structure [9]. The rest of the simulations were initialized with extended chromatin configurations predicted by the neural network model. More simulation details can be found in the Appendix C.

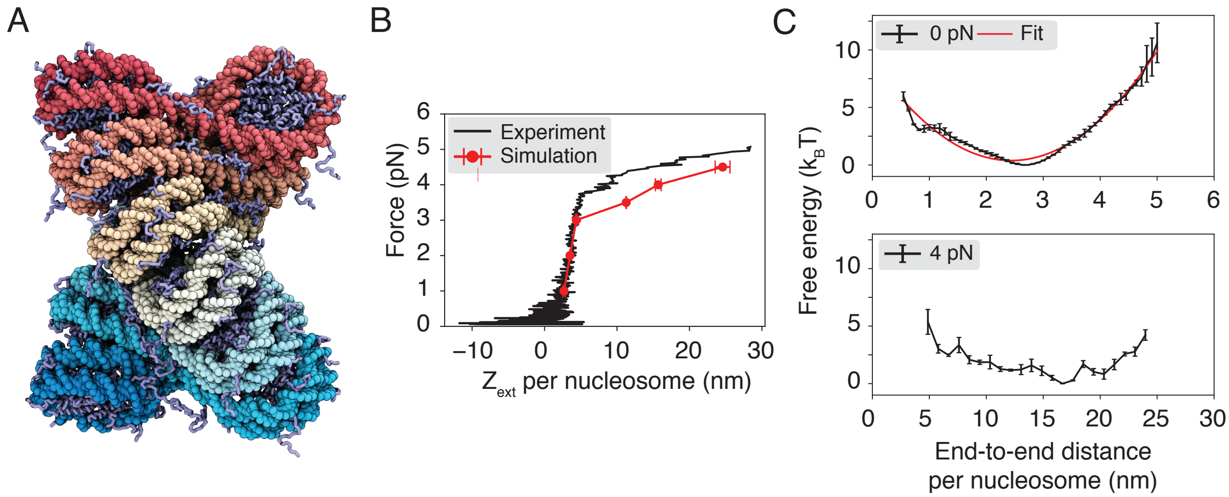


Figure 4.1: **Coarse-grained modeling reproduces the force-extension curve for chromatin.** (A) Illustration of the two-helix fibril chromatin structure with a linker length of 20 bp. The DNA molecule varies from red to cyan across the two ends, and histone proteins are drawn in ice blue. (B) Comparison between the simulated (red) and experimental [202] (black) force-extension curve. (C) Free energy profiles as a function of the DNA end-to-end distance computed with the presence of 0 pN (top) and 4 pN (bottom) extension force. A harmonic fit to the 0 pN simulation result is shown in red. Error bars correspond to the standard deviation of the mean estimated via block averaging by dividing simulation trajectories into three independent blocks of equal length.

4.3 Results

4.3.1 Coarse-grained modeling reproduces force-extension curve

We applied a residue-level coarse-grained model to characterize the unfolding of a 12mer chromatin with the 601 nucleosome positioning sequence [169] and a linker length of 20 bp (Figure 4.1A). One bead per amino acid and three sites per nucleotide were employed to describe protein and DNA molecules, leading to a system of 23590 coarse-grained beads in size. Interactions among the coarse-grained beads were parameterized by accounting for solvent effect implicitly with physically motivated potentials (see Methods for model details). Similar approaches have been extensively used to characterize single nucleosomes [51], [150], [207] and nucleosome oligomers [46], [212] with great success.

We computed the average chromatin extension length under various pulling forces along the z-axis for a direct comparison with results from single-molecule pulling experiments [202]. Comprehensive sampling of chromatin conformations can be rather challenging because of the non-specific and strong electrostatic interactions between nucleosomes that give rise to slow dynamics. To alleviate the sampling difficulty, we carried out umbrella simulations [75] on two collective variables that quantify the degree of nucleosomal DNA unwrapping (q_{wrap}) and nucleosome unstacking (d_{stack}) (Figure C.6). The simulations were initialized from the most probable configurations at respective umbrella centers obtained from an exhaustive sampling of a neural network model that approximates the free energy landscape of the 12mer in terms of inter-nucleosome distances (see *Methods*). This initialization protocol attempts to prepare umbrella simulations with equilibrium configurations to avoid traps of local minima.

As shown in Figure 4.1B, the simulation results match well with the experimental force-extension curve measured by Kaczmarczyk et al. [202]. In particular, we observe a linear extension regime at low forces ($\leq$ 3 pN). The sharp increase in extension at large forces deviates from the linear behavior, resulting in a plateau regime. We emphasize that there are no tuning parameters in the model, and we do not make assumptions regarding stacking energies.

The free energy profiles as a function of the DNA end-to-end distance are consistent with the linear and plateau regimes seen in force-extension curves (Figure 4.1C). In particular, at 0 pN force, the free energy curve can be well approximated with a harmonic potential, which naturally produces a linear relationship between the force and extension. Consistent with a harmonic behavior near the minimum, theoretical predictions based on the free energy profile at 0 pN match well with simulation results at 1-3 pN (Figure C.7). However, the free energy profile at 4 pN is strongly anharmonic. The bottom panel shows that the curve is relatively flat over a wide range of end-to-end distances. Because of the lack of energetic penalty, a slight change in pulling force can produce significant variations in the extension length, giving rise to the observed plateau regime.

We note that several factors could contribute to the discrepancy between simulated and experimental curves at large forces. For example, the relatively flat landscape over a wide range of chromatin conformations makes it challenging to predict the free energy minimum and the average end-to-end distance. Minor errors in conformational sampling and free energy calculations could be amplified into significant changes in the chromatin extension. In addition, our implicit treatment of counter ions introduces approximations to protein-DNA interactions. It may be insufficient to mimic the exact experimental setting with both monovalent and divalent ions [202]. Consistent with this hypothesis, varying the salt concentration in simulations improved the agreement in the average extension length with the experimental value (Figure C.8).

4.3.2 Intermediate states support nucleosome-clutch formation

The nucleosome arrangement in extended, unfolded chromatin has been the subject of numerous studies [24], [188], [189], [196]. The residue-level coarse-grained simulations offer a unique opportunity to produce high-resolution structures with minimal assumptions. Their success in reproducing experimental observations shown in Figures 4.1B and A.8 supports the biological relevance of the predicted structures.

We determined representative structures at various forces to better characterize chromatin unfolding under tension (Figure 4.2). These structures share end-to-end distances close to the mean force-dependent extension lengths. They correspond to the central configurations of the most populated clusters identified by the single-linkage algorithm [125] using root mean squared distance (RMSD) as the distance between structures. At small forces (≤ 3 pN), though chromatin extends linearly, we do not observe a uniform extension of nucleosomes along the principal fiber axis (Figures 4.2 and C.9). The conformational change mainly occurred in the plane perpendicular to the fiber axis via a shearing motion, causing the formation of irregular, compact structures. Such structures are more kinetically accessible as they avoid complete unstacking, which could cause a significant energetic

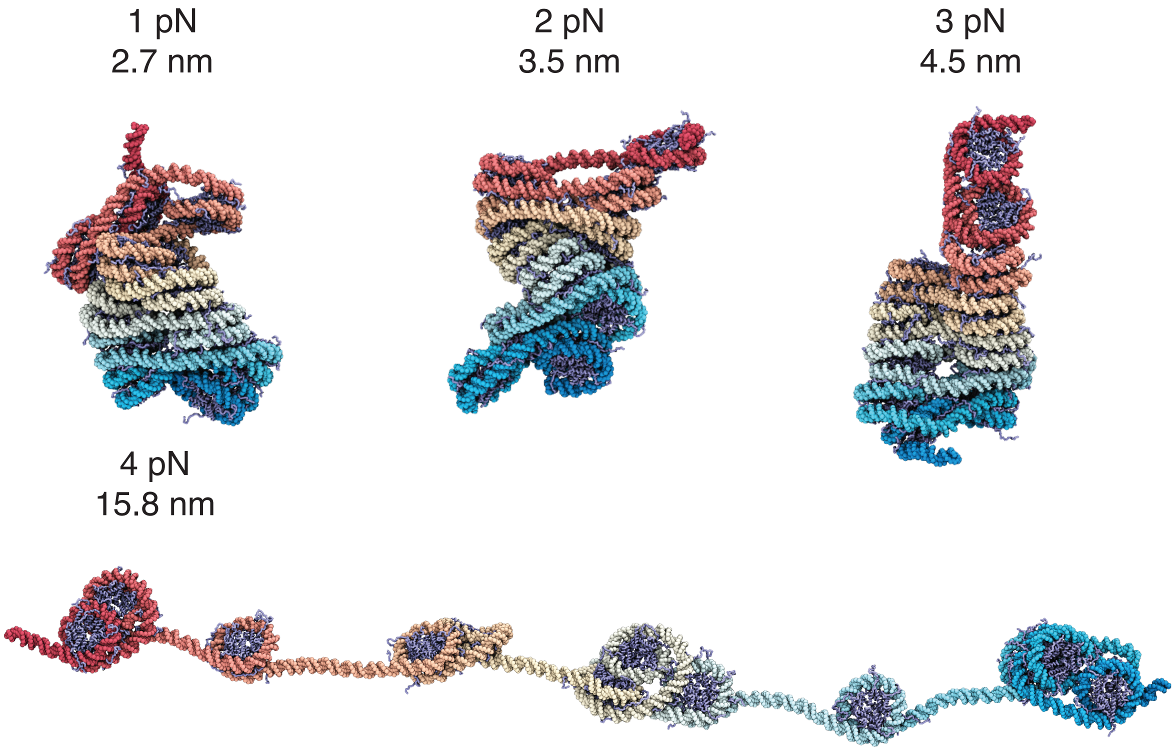


Figure 4.2: **Representative chromatin structures from simulations performed under various extension forces (also see Figure C.9)**. The values for the extension force and the end-to-end distance are provided next to the structures. The same coloring scheme as in Figure 4.1A is adopted here.

penalty as shown by [207]. We note that an ensemble of chromatin configurations exists at a given end-to-end distance, and only example ones are shown in Figure 4.2. Averaging the entire ensemble produces more symmetric structures and nucleosome contact patterns (Figure C.10).

The preference of shearing over complete unstacking can be readily seen in Figure 4.3. There, we decomposed the distance between two nucleosomes into motions that are within or perpendicular to the nucleosomal plane (Figure C.11). We further computed the free energy profile for the two decomposed distances under no extension force. It is evident that the energetic penalty for chromatin unfolding along the shearing direction is much smaller. Shearing can better preserve inter-nucleosome contacts as nucleosomes move away from each other, lowering the energetic penalty.

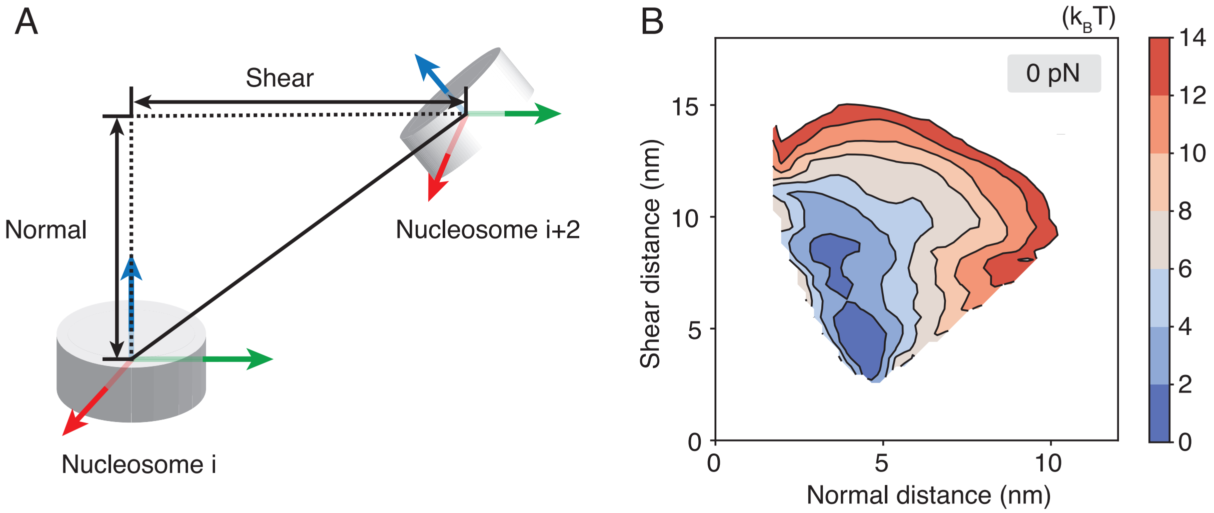


Figure 4.3: **Chromatin extension favors shearing motion within the nucleosomal plane over the normal motion perpendicular to the plane.** (A) Illustration of the nucleosome coordinate system and the decomposition of the inter-nucleosome distance into shearing and normal components. (B) The free energy profile as a function of the two different modes of breaking inter-nucleosome distances shown in part A.

The representative structure from 3 pN to 4 pN undergoes a dramatic transformation from a compact configuration to one with many nucleosomes losing stacking interactions. Notably, the unfolded structures fall into small clusters of nucleosomes. These structures often feature one or two nucleosomes with a highly unwrapped outer layer. Unwrapping the outer layer DNA only incurs modest energetic cost [148], [150], [213] and serves as an economic strategy to extend chromatin under force. Nucleosome clutch formation is not specific to a particular end-to-end distance and can be readily seen in structures with smaller distances as well (Figure 4.4 and C.12). We note that the nucleosomes that remain in contact are not perfectly stacked as in the crystal structure of a tetranucleosome [8], but are somewhat irregular as configurations observed in prior simulations [46], [199] and in vivo experiments [13], [25], [214], [215]. Further stretching the chromatin eventually leads to configurations with most of the outer nucleosomal DNA unwrapped.

Our results suggest chromatin unfolding does not proceed via a uniform nucleosome unstacking. On the contrary, nucleosomes prefer to stay in close contact as much as possible by

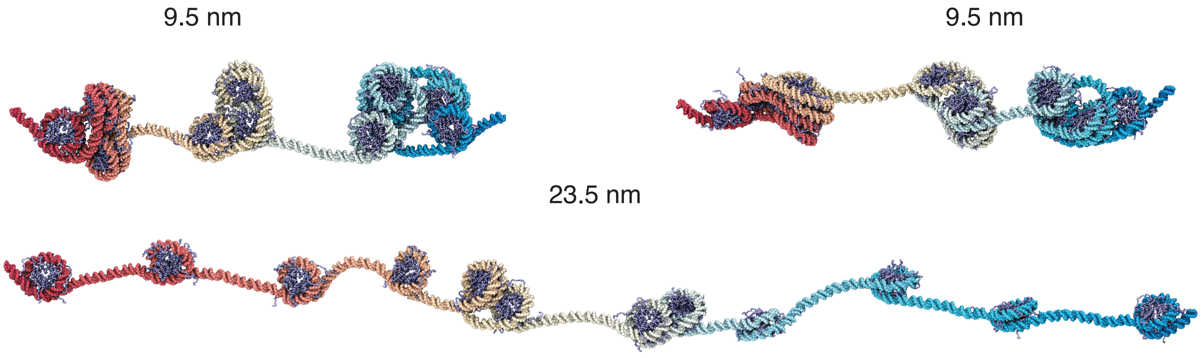


Figure 4.4: **Representative chromatin structures at smaller and larger distances than the average extension at 4 pN force (see also Figure C.12)**. These structures again support the formation of nucleosome clutches, which do not break into individual nucleosomes until at very large per-nucleosome end-to-end distances around 23.5 nm.

forming clusters separated by unwrapped DNA. To ensure that the formation of nucleosome clutches is not a result of biases from initial configurations prepared by the neural network model, we carried out an additional set of simulations starting from uniformly extended chromatin structures (Figure C.13A). More simulation details are provided in the Appendix C and Tables C.3 and C.4. As shown in Figure C.14A, these new simulations produced a free energy profile as a function of the end-to-end distance that matches well with the one presented in Figure 4.1C, supporting the statistical convergence of our simulations. To resolve the degree of clutch formation in chromatin configurations, we introduced a new collective variable, α, that quantifies the ratio of the maximum and minimum distance between 1-3 nucleosomes, i.e., $\alpha = d^{\max}_{i,i+2}/d^{\min}_{i,i+2}$. For clutched configurations, the distance between two nucleosome clusters is expected to be much larger than the distance between nucleosomes within the same cluster, and α will be much larger than one. On the other hand, for more uniformly extended configurations, α will approach one. The free energy profile as a function of α exhibits a global minimum at values much larger than one (Figure C.14B), supporting the stability of clutched configurations. Example configurations at various end-to-end distances adopt large α values (Figure C.13B) and resemble those presented in Figure 4.2. Therefore, the formation of nucleosome clutches under tension is an inherent property of chromatin and robust to simulation protocols.

We further confirmed that the unwrapping of the outer nucleosomal DNA is essential for clutch formation. In a new set of umbrella simulations, we removed DNA unwrapping by rigidifying the entire 147 bp nucleosomal DNA with the histone core. These simulations were performed with the presence of 4 pN force and were initialized from the fibril structure (see the Appendix C for additional simulation details). Figure C.15 shows that, when the DNA was prohibited from unwrapping, chromatin favors more uniform configurations when extended. The free energy profile as a function of α computed with the new simulations reaches the minimum value at around 3 ((Figure C.15B). On the other hand, much larger values for α are favored when unwrapping is allowed. DNA unwrapping helps chromatin preserve inter-nucleosome contacts when stretched, leading to energetically more favorable clutched configurations (Figure C.16). Without unwrapping, inter-nucleosome contacts must be broken to satisfy geometric constraints to reach a given extension, resulting in more uniform chromatin structures (Figure C.15C).

4.3.3 Inter-chain contacts stabilize unfolded chromatin

The pulling simulations suggest that in vivo configurations can arise from the unfolding of chromatin fiber under tension. Inside the nucleus, chromatin is not in isolation but surrounded by other chromatin segments in a crowded environment [12], [203]. The more exposed nucleosomes in the intermediate configurations could facilitate inter-chain interactions, further stabilizing the unfolded structures.

To evaluate the impact of crowding on chromatin stability, we computed a two-dimensional free energy profile using simulations with two 12mers. The first collective variable quantifies the inter-chain contacts as the number of nucleosome pairs within a distance of 15 nm. Only pairs with one nucleosome from each chromatin segment were included to define the contacts. The other dimension measures chromatin extension using the average unstacking of the two chains, $\bar{d}_{\text{stack}}$. Figure 4.5A shows that configurations with close contacts between the two

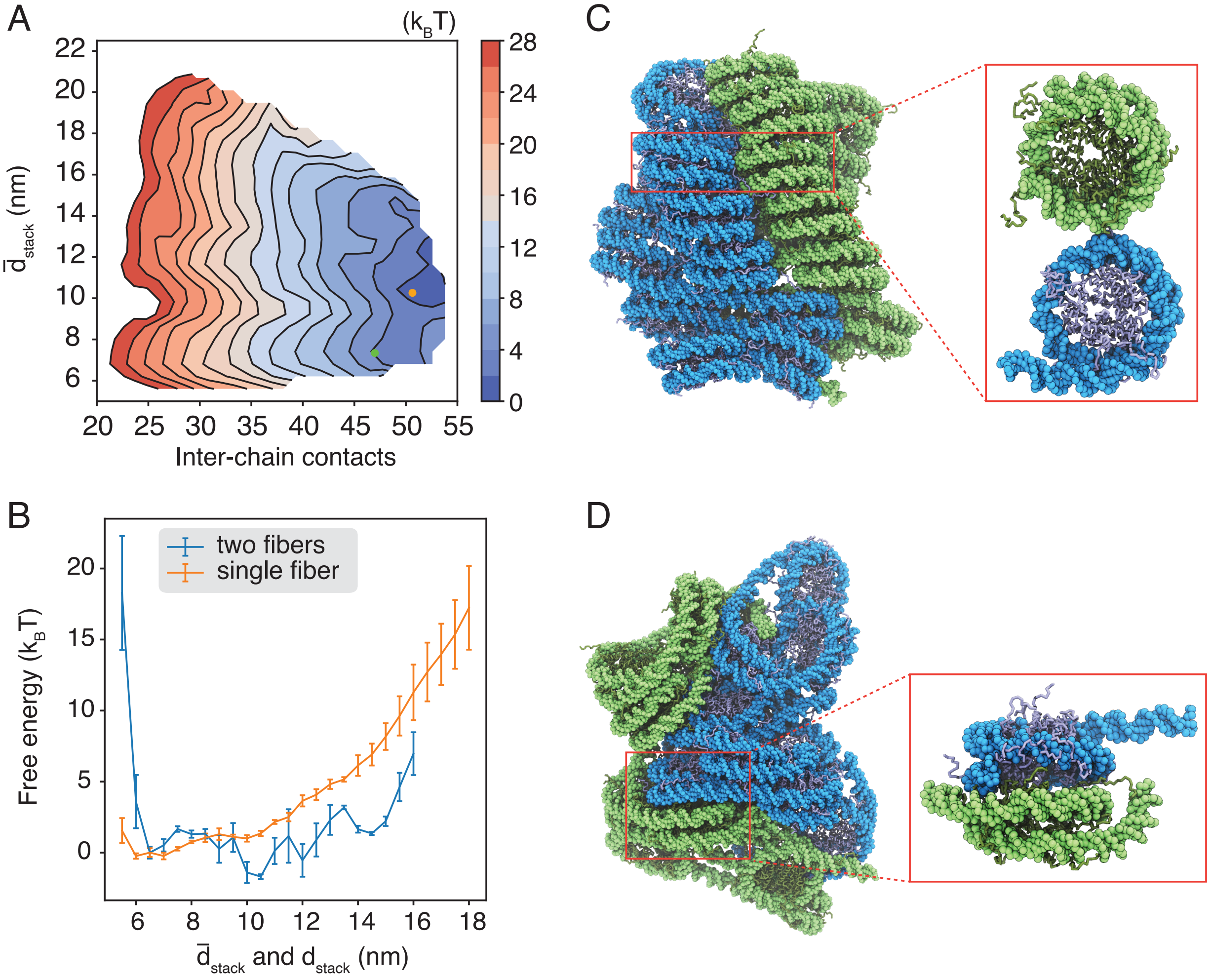


Figure 4.5: **Crowding and inter-chain contacts stabilize extended chromatin configurations.** (A) The free energy surface as a function of the inter-chain contacts and the average extension of the two 12mers. (B) Free energy profiles of chromatin unstacking with (blue) and without (orange) the presence of an additional 12mer. Chromatin unstacking is quantified with d_{stack} and $\bar{d}_{\text{stack}}$ for single and two fiber simulations, respectively. (C) Representative structure for two contacting chromatin segments that maintain fibril configurations, with the corresponding collective variables indicated as the green dot in part A. The inset highlights the side-side contacts between inter-chain nucleosomes. (D) Representative structure of the free energy minimum, with the corresponding collective variables indicated as the orange dot in part A. The inset highlights the stacking interactions between inter-chain nucleosomes.

chromatin segments are more favorable. A representative structure for two contacting fibril chromatin identified by the single-linkage clustering algorithm is provided in Figure 4.5C. The contacts are mediated mainly by histone tail-DNA interactions, as can be seen in the inset that provides a zoomed-in view of the interface. Favorable interactions for compact chromatin are consistent with previous simulation studies that support the liquid chromatin state [48].

Notably, the global minimum of the free energy profile resides at larger values for $\bar{d}_{\text{stack}}$ corresponding to more extended chromatin configurations. While extending chromatin is unfavorable (Figure 4.5B), such structures promote close contacts between nucleosomes from different chains (Figure 4.5D). In particular, trans-nucleosomes can now engage in stacking interactions (Figure 4.5D), which are more favorable energetically compared to side-side contacts [207]. The emergence of a new binding mode, unavailable when chromatin is constrained into fibril configurations, compensates for the energetic penalty of breaking cis-chain contacts. Further extending the chromatin leads to more intertwined structures at a rather modest energetic cost (Figure C.17).

Similar to the single-chain simulations, extending chromatin again led to irregular configurations with nucleosome clutches. As shown in Figures C.18 and A.19, the degree of irregularity increases monotonically with $\bar{d}_{\text{stack}}$ and for intermediate values of inter-chain contacts. DNA unwrapping in irregular chromatin configurations relieves the torsional constraints on nucleosomes to sample a much wider range of relative nucleosome-nucleosome orientations and distances. As a result, nucleosomes can now engage in many simultaneous energetically-favorable interactions, both with nucleosomes from the same chain and different chains.

To further evaluate the contribution of DNA unwrapping to inter-chain contacts, we carried out a new set of simulations with fully rigidified nucleosomes. As before, the core nucleosomes move as rigid bodies, and only linker DNA and histone tails were kept flexible. The setup of umbrella centers and restraining constants are similar to simulations that allow DNA unwrapping (Table C.5), and more details are provided in the Appendix C. We found that the two chromatin forms fewer contacts in the new simulations. The free energy minimum for inter-chain contacts is located around 42 (Figure C.20), a value that is much smaller than that shown in Figure 4.5. Chromatin is less extended when DNA unwrapping is prohibited, reducing the free energy minimum for $\bar{d}_{\text{stack}}$ from 10 to 8 nm. When chromatin does extend, the configurations are also more uniform with less irregularity (Figure C.20), hindering the formation of interdigitated structures.

4.4 Conclusions and Discussion

We characterized the impact of tension and crowding on chromatin organization with computational modeling using a coarse-grained model. The compact fibril configuration with nucleosomes following a zigzag path was most stable for a 12mer chromatin segment in isolation. Consistent with previous studies [195], [196], [200], [216], we observed both unwrapping of nucleosomal DNA and unstacking between nucleosomes as chromatin unfolds from the fibril configuration due to the presence of tension. However, these changes are non-uniform and are initially localized to a small set of nucleosomes, leading to the formation of nucleosome clutches separated by unwrapped nucleosomal DNA. Such intermediate structures emerge as a result of balancing intra- and inter-nucleosome interactions. The clutched configurations sacrifice nucleosomal DNA by unwrapping to extend chromatin and preserve the energetically more favorable inter-nucleosome contacts.

Notably, the simulated intermediate structures resemble in vivo chromatin configurations. For example, super-resolution imaging of the core histone protein H2B in interphase human fibroblast nuclei has revealed the formation of nucleosome clutches of varying size [203]. High-resolution electron tomography studies further support the prevalence of trimers in the clutches [13], [215]. Cross-linking-based experiments that detect nucleosome contacts in situ support nucleosome clutches with tri- or tetranucleosome as well [214], [217]. Our results generalize the findings from a previous study on tetra-nucleosomes [46]. They support that certain in vivo chromatin structures may form as a result of unfolding from the fibril configuration. Since chromatin inside the nucleus can experience forces from various active processes [175], [176], [218]–[221], even for DNA sequences and linker lengths that strongly favor the fibril structure, chromatin may adopt irregular configurations because of tension. The clutched configurations are mostly seen at forces below 4 pN, a value that is indeed within the range expected from molecular motors [175], [219]–[221].

We further showed that unfolded chromatin could promote inter-chain contacts, leading to the formation of interdigitated structures. Such structures present an alternative binding mode compared to the close contacts between two fibril configurations. Interdigitation is indeed consistent with electron microscopy images of two chromatin segments that are in close contact [222]–[228]. These images revealed structures with diameters less than twice the 30 nm fiber, supporting an overlap between the two chromatin. In addition to supporting chromatin unfolding in a crowded environment, the interdigitated structures suggest that chromatin may, in fact, form gels at high density inside the nucleus. Gelation can form due to the stacking interactions between exposed nucleosomes from different chains, which are stronger than side-side interactions that are only accessible for nucleosomes in closely stacked fibers. The emergence of strong interactions could arrest the coarsening dynamics of small clusters to drive the percolation transition [229]. Furthermore, interdigitation could give rise to topological entanglements among chromatin chains, further producing slow kinetics and gelation. Therefore, the two binding modes could help understand the observation of both liquid and gel state of chromatin mixtures [17], [20], [25], [48], [230], [231].

We studied idealized chromatin with uniform DNA linker length and strong positioning sequence. Nucleosomes from natural chromatin are more heterogeneous with variations in histone modifications [232]–[236], linker DNA lengths and DNA sequences [170], [212], [237]–[241], and linker histone binding [9], [203], [242]–[244]. Such heterogeneity could also contribute to the formation of irregular chromatin structures and clutches, as shown recently by the Schlick group [236], [240], [245]. Our findings complement these studies and point to additional intrinsic factors that affect the stability of chromatin fibers. They might be particularly relevant for interpreting chromatin organization in heterochromatic regions and mitotic chromosomes. Due to its low transcriptional activity, chromatin in these systems is expected to be more uniform in histone modifications and linker DNA length, and its irregular organization may indeed arise from tension and crowding effects.

Chapter 5

Non-Markovian dynamics model reveals chromatin fiber destabilization in nucleosme condensates

5.1 Introduction

The genome organization is vital for diverse genetic functions. The hierarchical genome organization is dominated by different rules at different scales [1], [2], [47], [246], [247]. At the near-atomistic level, the genome is organized as nucleosomes connected by linker DNA, similar to beads on a string. Such DNA string of nucleosomes folds as chromatin. Although the crystal structure of short chromatin with uniform $10n$-bp linkers are well-knwon as 30-nm fibers [8], [9], the dynamics and structure of chromatin in vivo remain elusive and controversial [10], [248]. Many in vivo experimental techniques including cryo-electron microscopy, Micro-C, ChromEMT, show the lack of ordered fibril-like structures, while the 10-nm disordered arrays with some prevaling local oligomer motifs, such as trimers, α-tetrahedron and β-rhombus tetramers [10], [13], [174], [214], [249], [250].

Recently, condensate has been uncovered as a rule that pervades and governs diverse biological systems [14], [88], [251]. In vitro experiments show that chromatin can also form liquid-like condensates under diverse conditions [17], [18]. The condensate property is affected by many factors, such as linker lengths, post-translational modifications (PTMs), salt conditions, and involvement of other biomolecules [17], [18], [20], [21], [25], [252], [253]. Although the situation is very complex in vivo as other molecules interact with chromatin and cause loop extrusion, transcription, and topologically associating domains (TAD) [254]–[257], recent studies provide compelling evidence that inter-nucleosomal interactions alone can reproduce many chromatin organization features in vivo: with interaction energy defined based on the condensability score, which measures the ability of various types of nucleosomes to condense [258], molecular dynamics (MD) simulations recover the nucleosomal contact map and compartments [2]. This supports the idea that chromatin can be viewed as block copolymers and microphase separation drives chromatin organizations. Beyond the cause of phase separation, the exact physical properties of chromatin condensates in cells remain obscure. They demonstrate both liquid- and solid-like properties [19], [20]. This duality can be attributed to the intrinsic viscoelasticity of the polymers, or the heterogeneity of various chromatin condensates in cells [2], [21]. Therefore, understanding the phase separation behavior of nucleosomes is essential for understanding genome organization.

Computational models play a pivtol role in understanding genome organizations. At the nucleosome level, coarse-grained (CG) modeling is the appropriate model as it balances chemical details and efficiency. Numerous studies have proven the effectiveness of combining a three-bead-per-nucleotide DNA model called 3SPN2 [146] with a residue-level CG protein model, such as AWSEM [31] or SMOG [204] or AICG [259]. These CG models have been applied to study nucleosome unwrapping [150], [211], transcription factor (TF) binding [51], and chromatin fibers [46], [49]. Ding et al. applied 3SPN2 and SMOG to model the tetra-nucleosome. They further utilized enhanced sampling methods, neural network free energy landscape, and the string method to explore the folding mechanism of the tetra-nucleosome [46]. This study demonstrates the strength of exhaustive parallel sampling, while the exact flux of pathways and the transition dynamics remain to be solved.

The integration of the CG model with molecular dynamics (MD) simulations provides a powerful approach to investigate the intricate conformational changes of chromatin under different circumstances. A seamless approach for elucidating the folding dynamics and pathway flux from MD trajectories involves integrating Markov state models (MSMs) with transition path theory (TPT) [260]–[267]. This approach has been successfully applied to investigate a range of conformational changes in chemical and biological processes, such as protein folding [268], [269], protein-ligand recognition [270], [271] and the self-assembly of soft materials [272], [273]. Chromatin has a high degree of freedom even under CG representations (about 2000 CG atoms per nucleosome), and normally biologically relevant phenomena are very slow, so it takes long time for MD simulations to sample a continuous long trajectory to directly observe the phenomena. MSMs could coarse-grain MD trajectories into conformational states and bridge the timescale gap by integrating multiple short MD trajectories and modeling dynamics as a series of Markovian jumps between conformational states under given lag times. Additionally, the transition path theory (TPT) can be naturally applied to MSMs to compute the committor functiosn and identify the kinetic pathways [260], [261], [274]. Constructing MSMs with parallel short trajectories requires employing a large number of states to ensure their transitions are Markovian, which greatly complicates intuitive interpretation. A new approach called integrative generalized master equation (IGME), which builds non-Markovian dynamics models by explicitly evaluating memory effects [63], promotes human comprehension. The non-Markovian dynamics model built by IGME with much fewer metastable states provides an intuitive picture of the slowest processes, and non-Markovian dynamics can be encoded in the memory kernels. Such a method is promising for deciphering various slow processes in biological systems with abundant short trajectories.

In this study, we combined CG modeling, MD simulations, and transition path analysis with non-Markovian dynamics models to quantitatively determine the folding dynamics of the tetra-nucleosome. Specifically, we study two biologically related factors: the high local nucleosome concentration and the linker lengths. We try to understand how these two factors affect the folding dynamics, including folding pathways, intermediate states, and transition rates.

5.2 Methods

5.2.1 Coarse-grained modeling and molecular dynamics simulations of chromatin organization

We applied a structure-based model (SBM) called SMOG for histones [133], and a 3-site per nucleotide model called 3SPN for DNA [146], [147], [206]. Note that the histone core and the inner layer 73 bp of the core DNA were always rigid during all simulations. This setting stabilized the nucleosome cores and was proved valid in previous studies [46], [49]. Starting from the 10,000 diverse configurations of tetra-nucleosome with 20 bp linkers explored with enhanced sampling methods reported in a previous study [46], we picked all the 4643 configurations with $d_{13} \geq d_{24}$ (d_{ij} is the distance between the i-th and j-th nucleosomes) as our starting configurations. We measured the d_{ij} $(i, j = 1, 2, 3, 4)$ of 4643 configurations and enforced the tetra-nucleosome structure with 20 or 25 or 30 bp linkers towards the target d_{ij} values with restrained MD. All the single tetra-nucleosome systems are shown in Figures 5.1A-C. Then we removed all the restraints and performed 10-ns unbiased NVT simulations for 4643 structures (Figure 5.2A). The 10-ns NVT simulation trajectories were utilized for post-analysis, which will be introduced below.

We further performed simulations for the NRL = 167 system in a sea of nucleosomes to mimic the nucleosome condensate environment. Since such simulations were expensive, we reduced the number of initial configurations by selecting 530 representative structures from the Markov State Model (MSM) results of the individual NRL = 167 tetra-nucleosome. The initial tetra-nucleosome configurations were placed in a cubic box of length 55 nm. Then we added single nucleosomes to the box so that the total nucleosome concentration was 0.3 mM. The overall system is composed of a tetra-nucleosome and 26 single nucleosomes. We began with fixing the whole tetra-nucleosome as a rigid body and relaxing the single nucleosomes with NVT simulations for 2 ns. Next, we released the global rigid body restraints on the tetra-nucleosome (the histone core and core DNA of tetra-nucleosome and single nucleosomes were still rigid) and ran NVT simulations as the production run. Each production NVT trajectory lasts at least 70 ns. Exemplary snapshots are shown in Figure 5.1D, which shows the tetra-nucleosome unfolding process within the condensate.

All the electrostatic interactions were computed under temperature 300 K and 150 mM ionic strength with Debye-Hückel potential. All simulations were performed under 300 K with a time step of 10 fs. All the single tetra-nucleosome simulations were performed with LAMMPS on CPUs [127], and all the condensate simulations were performed with OpenMM and OpenABC packages on GPUs [61], [62]. More details of the simulations are provided in SI.

5.2.2 Markov State Modeling

We constructed MSMs for four systems: single nucleosomes with NRL = 167 individual, 172, and 177, as well as the NRL = 167 system in a sea of nucleosomes (named as NRL = 167 condensate system, or condensate system for brevity). These MSMs were based on six pairwise distances d_{ij} between the nucleosomes. Although these six distances effectively illustrate the global topology of the tetra-nucleosome as proved in the previous study [46], some of them exhibit high correlations and redundant information. Therefore, we further performed tICA [275]–[278] to recombine the six distances to obtain independent collective variables (CVs). The number of CVs and tICA relaxation time were determined through cross-validation using GMRQ scores [279]. Upon projecting all the samples onto the selected CVs, the resulting samples were clustered into hundreds of microstates by the K-Means algorithm (Figure 5.2B). Furthermore, microstate MSMs were constructed, and their Markovian properties were validated using the implied time scale (ITS) and the Chapman-Kolmogorov (CK) test [263]–[266].

Utilizing the microstate MSM in conjunction with transition path theory (TPT), we analyzed kinetic transition pathways and folding dynamics [272]. By designating the source and sink microstates as extended and folded conformational states, respectively, we identified numerous pathways that traverse multiple microstates. To simplify analyses, we further grouped these pathways into three channels based on their distribution in the tICA space using the latent space path clustering algorithm [267] (Figure 5.2C).

To construct representative and interpretable models, we aggregated hundreds of microstates into six macrostates for all four systems. Since relaxation within macrostates requires a longer lag time than for microstates, potentially exceeding the trajectory length, we considered non-Markovian effects to determine macrostate model properties [280]. Utilizing the recently developed integrative generalized master equation methods [63], we effectively captured long-term dynamical behavior by incorporating historical memory (Figure 5.2D). This enabled us to accurately calculate the associated stationary populations and mean first passage times (MFPTs).

5.3 Results

5.3.1 Computing transition pathways and rates for folding tetra-nucleosomes

We used coarse-grained (CG) models and molecular dynamics (MD) simulations to study the folding dynamics of tetra-nucleosomes. Specifically, we used a one-bead-per-amino-acid and three-bead-per-nucleotide model to capture the physical chemistry interactions. Such models have been successfully applied to chromatin systems and match the experimental results

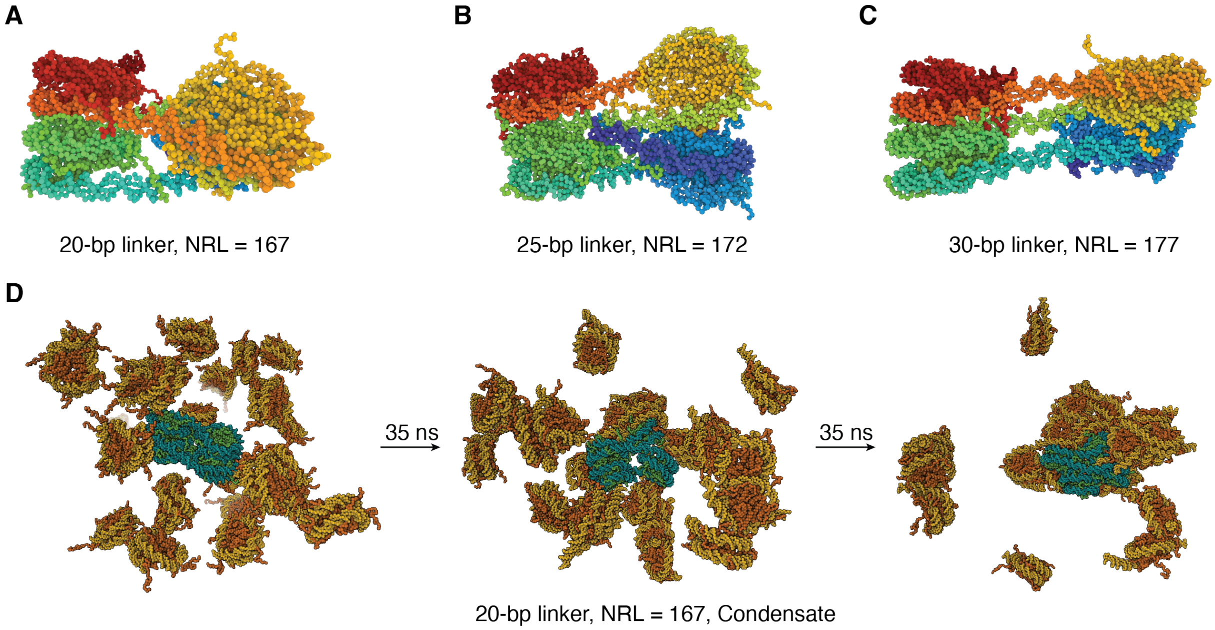


Figure 5.1: **The chromatin systems simulated in this study.** (A) 20-bp linker tetra-nucleosome (NRL = 167). (B) 25-bp linker tetra-nucleosome (NRL = 172). (C) 30-bp linker tetra-nucleosome (NRL = 177). (D) Example snapshots from one production trajectory of the condensate simulation with tetra-nucleosome unfolding. The tetra-nucleosome is shown in green and cyan, and the single nucleosomes are shown in orange and yellow.

accurately [46], [49], [50]. To study the effects of linker lengths, which is the length of DNA connecting two neighboring nucleosomes, we built a series of tetra-nucleosome structures with 20, 25, and 30 bp linkers, correspoding to nucleosomal repeat length (NRL) equal to 167, 172, and 177 (Figure 5.1A-C). To study the more biologically relevant condition, we placed the 20-bp linker tetra-nucleosome in the sea of single nucleosomes, leading to an overall nucleosomal concentration of 0.3 mM (Figure 5.1D), which is on the same scale as the interphase nucleosome concentration (about 0.1 mM) [60], and close to the condensate concentration in vitro (about 0.35-0.5 mM) [17], [25]. More details about building the systems can be found in Appendix D *Simulation Details.*

We performed exhaustive unbiased simulations at 300 K to explore the conformational changes of the tetra-nucleosome. For the individual tetra-nucleosome systems with 20, 25, and 30 bp linkers, we adopted 4643 independent simulations for each system starting from diverse configurations obtained from unified free energy sampling [281]. For the NRL =

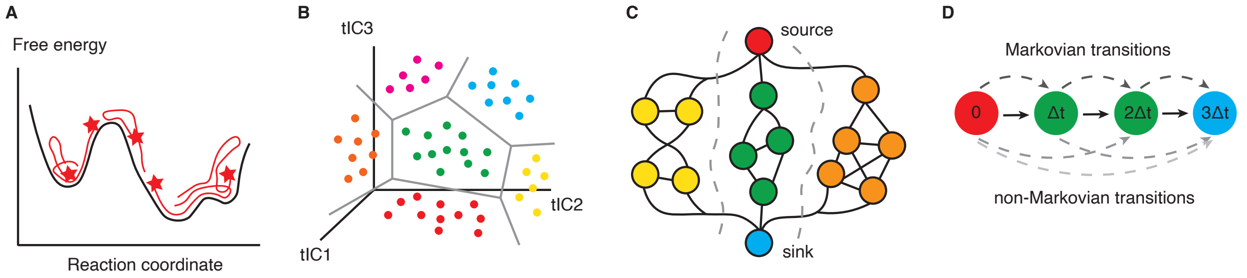


Figure 5.2: **The cartoon overview of the simulation and post-analysis protocol.** (A) Explore and sample the free energy landscape with exhaustive unbiased sampling starting from diverse configurations. (B) Project MD conformations onto the collective variables identified by tICA, and further cluster them into microstates. (C) Employ TPT to identify ensemble of kinetic pathways connecting source and sink based on microstate MSM. The pathways are further categorized into metastable path channels by LPC algorithm. The yellow, green, and orange color represent the assignments of three path channels. (D) Construct non-Markovian dynamics model by IGME method. The microstates are further lumped to few interpretable macrostates, and the transition dynamics are modeled by incorporating the memory kernel through the generalized master equation.

167 tetra-nucleosome condensate system, we adopted 530 independent long-time trajectories (at least 70 ns for each trajectory) as the production run. The simulations were long enough to fully relax the system and reach equilibrium, and the condensate simulation box is large enough to prevent tetra-nucleosome self-contact across box boundaries (Figure D.1-2). According to the evolution of the example trajectory, the nucleosomes aggregate as the simulation proceeds (Figure 5.1D). These unbiased simulations extensively explore the free energy landscape and facilitate subsequent analysis (Figure 5.2A).

We performed the folding pathway analysis based on the microstate Markov state models (MSMs), and further constructed the macrostate non-Markovian dynamics models to understand the folding mechanism. First, we doubled the trajectories based on the symmetry that nucleosome indices (1, 2, 3, 4) are equivalent to (4, 3, 2, 1), which was applied in a previous study as well [46]. For the construction of microstate MSMs and the application of TPT, we began by featurizing the MD conformations using six inter-nucleosome distances, denoted as $\mathbf{d} = (d_{12}, d_{13}, d_{14}, d_{23}, d_{24}, d_{34})$, which has been demonstrated to effectively differentiate tetra-nucleosome configurations and capture the folding dynamics [46]. Subsequently, we utilized time-independent component analysis (tICA) with the kinetic mapping algorithm to identify three independent collective variables [275]–[278]. After projecting MD conformations onto these three collective variables, we employed the K-means algorithm to group them into hundreds of microstates (Figure 5.2B). Then we applied TPT to discern all the kinetic pathways connecting the unfolded and folded states [260], [261], [274]. Furthermore, we used LPC algorithms to categorize thousands of identified pathways into three metastable path channels for each system to improve understanding of mechanisms [267], [272], [282] (Figure 5.2C).

To comprehend the dynamics of conformational changes in tetra-nucleosomes and improve understanding, we employed IGME to construct non-Markovian dynamics models [63], encompassing six macrostates through kinetic lumping from hundreds of microstates (Figure 5.2D). IGME outperforms conventional MSMs by considering non-Markovian dynamics with the generalized master equation, thereby providing more accurate predictions for long-term dynamics between macrostates. More details about the MSM analysis and non-Markovian dynamics models can be found in Appendix D *Markov State Model and non-Markovian Dynamics Model Construction*.

5.3.2 Downhill folding channels for tetra-nucleosome

We initiated the investigation into the folding dynamics of the NRL = 167 tetra-nucleosome. Utilizing an ensemble of unbiased MD trajectories, we estimated the free energy landscape of the NRL = 167 tetra-nucleosome based on d_{13} and d_{24} coordinates (Figure 5.3A). The downhill landscape, displaying an approximate 8 kcal/mol disparity between the unfolded and folded states, signifies the stability of the native tetra-nucleosome structure and aligns with previous findings from deep learning fitting and umbrella sampling [46]. By constructing a 530-state MSM and conducting TPT analysis, we found over 25,000 pathways linking unfolded and folded states, with the dominant pathway representing just 0.09% of total flux. This contrasts with typical protein folding; for example, while the top 10 pathways account

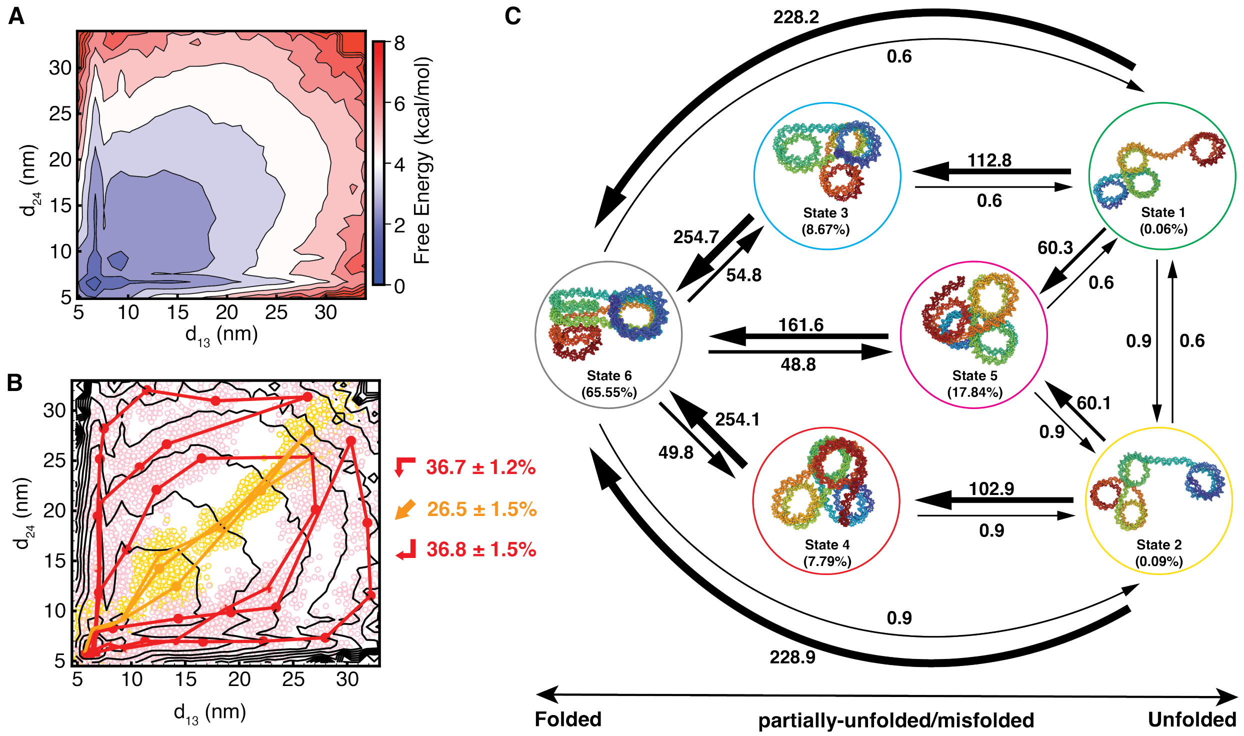


Figure 5.3: **Analysis results of the NRL = 167 single tetra-nucleosome system.** (A) The free energy profile along d_{13} and d_{24}. (B) Top three transition pathways of each type of path channels. The red pathways are sequential pathways, and the yellow ones are concerted pathways. The dots are the samples along the top pathways. The overall flux of each path channel are labeled as percentage. (C) Macrostate non-Markovian dynamics model with inverse MFPT labeled in unit (10 μs)$^{-1}$. Histones are hidden for clarity.

for 25% of NTL9 folding flux [269], approximately 600 pathways are required to achieve the same level of flux in tetra-nucleosome folding. The downhill landscape and identification of parallel pathways with comparable fluxes suggest that tetra-nucleosome behavior is more similar to heterogeneous aggregation and self-assembly systems instead of the proteins [272].

Due to the abundance of parallel pathways with similar fluxes, we further grouped the pathways into three path channels to facilitate understanding folding mechanism. We got three path channels: two sequential channels (up and down sequential channels), and one concerted channel (Figure 5.3B and B.17). Sequential channels indicate that one pair of nucleosomes stacks before the other, while the concerted channel exhibits simultaneous stacking motions. This is consistent with the existing study, which characterized the tetra-nucleosome folding pathways with the string method [46]. Beyond qualitatively identifying path channels, we quantify the flux of each path channel with slightly higher flux for sequential ones, possibly due to stable intermediates. Meanwhile, we visualized transition states with committor Q close to 0.5 and the states with highest fluxes (Figures D.23-24). Due to the downhill nature of the free energy landscape, the $Q \approx 0.5$ states displayed more extended and unfolded conformations, while those structures with larger Q show structures similar to the β-rhombus structures and oligomers observed in situ [13], [214].

We further lumped hundreds of microstates into six metastable macrostates and built the non-Markov dynamics model by considering the memory effects with IGME to promote the comprehension of the folding mechanism. The transition network with the inverse mean first passage time (MFPT) as shown in Figure 5.3C. The inverse MFPT values, which represents reaction rates, and the stationary populations, align with the strong folding tendency. The partially-unfolded state 3 and state 4 resembled metastable intermediates in sequential channels. State 5 is recognized as the misfolded state since no top 3,000 pathways going through the state. The β-rhombus structures, through which the concerted channel proceeds, are assigned to folded state 6, suggesting their low metastability and fast transition to folded structures. Together, these analyses build a quantitative and intuitional picture of the NRL = 167 tetra-nucleosome folding process.

5.3.3 Crowding environment promotes and accelerates chromatin unfolding

Inspired by recent studies, which show that nucleosomes tend to form liquid- or solid-like condensates [17], [18], [20], [21], [25], we placed the NRL = 167 tetra-nucleosome into a sea of single nucleosomes with overall nucleosomal concentration as 0.3 mM to investigate how the condensate environment affects the folding of tetra-nucleosomes. Initiating the simulations with uniformly distributed single nucleosomes, we observed that they undergo a self-assembly process, either interacting and stacking with tetra-nucleosomes or aggregating

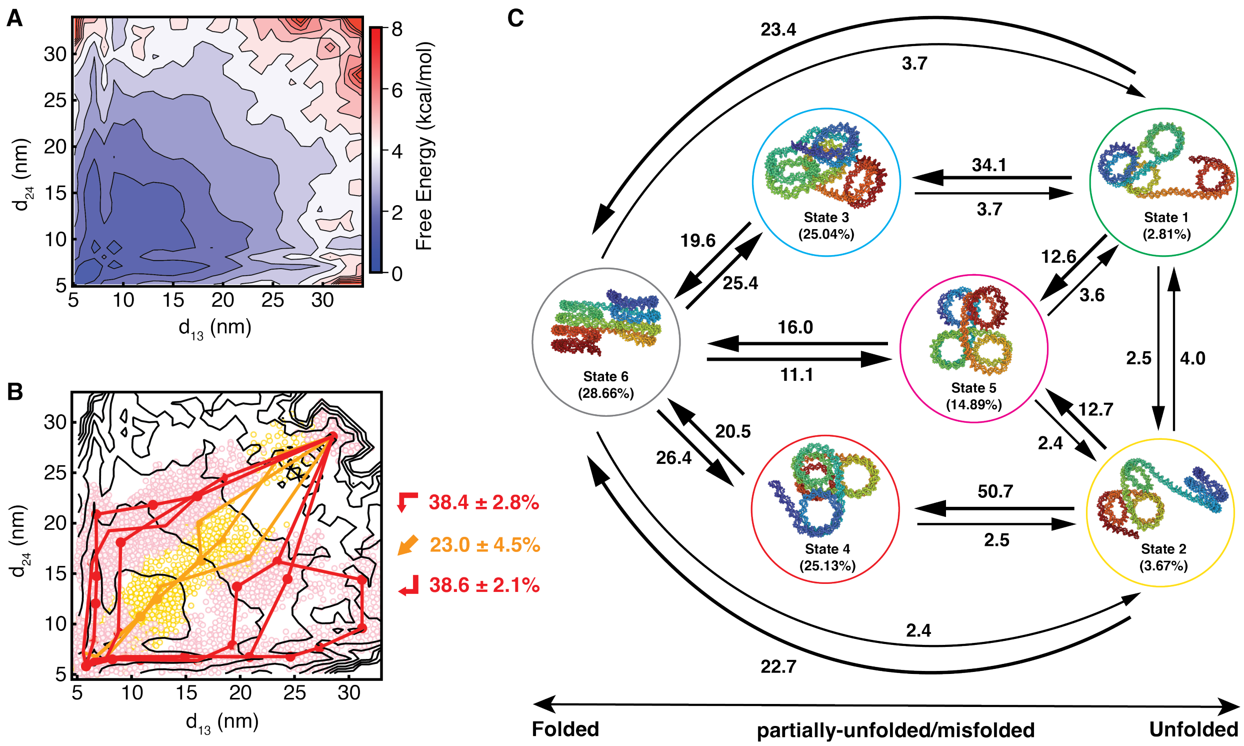


Figure 5.4: **Analysis results of the NRL = 167 tetra-nucleosome condensate system.** (A) The free energy profile along d_{13} and d_{24}. (B) Top three transition pathways of each type of path channels. The red pathways are sequential pathways, and the yellow ones are concerted pathways. The dots are the samples along the top pathways. The overall flux of each path channel are labeled as percentage. (C) Macrostate non-Markovian dynamics model with inverse MFPT labeled in unit $(10\ \mu\text{s})^{-1}$. Histones are hidden for clarity.

into clusters independently. We estimated the free energy landscape of the tetra-nucleosome within condensate based on the simulations. As illustrated in Figure 5.4A, although the difference between the unfolded and folded states remains around 8 kcal/mol, a broader range of configurations around the native structure exhibit lower free energy. This occurs because interactions between tetra-nucleosome and single nucleosomes resemble intra-chain contacts and stabilize partially unfolded conformations (Figure 5.1D).

By further identifying the folding pathways and clustering them into path channels, we discovered numerous pathways with comparable fluxes, and the pathways can be attributed to similar sequential and concerted channels. The distribution of fluxes among these channels remained largely unchanged compared to the single NRL = 167 tetra-nucleosome (Figure

5.4B and D.17). However, the transition rates change dramatically, as demonstrated by the six-macrostate non-Markovian dynamics model (Figure 5.4C). It is evident that the nucleosome condensate influences both the thermodynamics and kinetics of the tetra-nucleosome. The populations of partially unfolded states (state 3 and 4) have significantly increased, approaching that of the folded state, and the unfolding rates increases dramatically relative to folding ones. In general, the nucleosome condensate promoted and accelerated all unfolding processes relative to the folding process. This rapid unfolding is consistent with the liquid-like properties of wild-type chromatin condensate in vitro [17], [18] and perhaps related to chromatin organizations and functions in vivo. The misfolded state 5 persists, with its population remaining relatively unchanged. In all, we observed that the nucleosome condensate did not introduce new conformations or folding modes of the tetra-nucleosome, however, it modulates the free energy landscape and favors unfolding dynamics.

5.3.4 The role of DNA linker length on chromatin folding

Linker length is another critical factor that influences chromatin organization and phase separation behaviors [17]. With a DNA twist periodicity of about 10 base pairs, a linker length of $10n$ facilitates well-aligned stacking between i and $i+2$ nucleosomes. However, inserting additional 5-bp into linker DNA causes an additional half-turn twist, thus hindering the well-aligned nucleosomal stacking and destabilizing the native state of the tetra-nucleosome. Supporting evidence is the lack of crystal structures of $10n+5$ linker chromatin, despite the abundance of solved chromatin structures with $10n$ linkers [8], [9], [244], [283]. $10n+5$ linkers are important as there are plentiful $10n+5$ linkers in certain mammalian cells [284]. Here, we further investigate how longer linkers, particularly the $10n+5$ linkers, impact the folding of the tetra-nucleosome. We performed the similar workflow to tetra-nucleosomes with 25-bp linkers (NRL $=$ 172) and 30-bp linkers (NRL $=$ 177), respectively.

In the NRL $=$ 172 system, we observed that the half-turn linker DNA significantly perturbs the free energy landscape of the tetra-nucleosome, with a broader global minimum and a smaller free energy gap between the folded and unfolded states (Figure 5.5A). Again, we constructed the six-macrostate non-Markovian dynamics model (Figure 5.5B) and characterized representative structures. State 1 is the extended state. Although its population is low, the structure does not suffer from unfavorable DNA twists or curves. It demonstrates that when DNA is fully relaxed, consecutive nucleosomes will reside on different sides of the linker due to $10n + 5$-bp linker lengths, while $10n$-bp linkers will position two nucleosomes on the same side. This is the main topological difference between $10n$ and $10n + 5$ linkers. State 2, 3, and 4 are also of low populations. They feature one pair of nucleosomes in contacts. They are unstable as DNA is over-curved and wrapped, while the inter-nucleosomal interactions are not strong enough to compensate. For example, in 5.5B, the DNA wrapped more on the green nucleosome in state 2 structure, the orange nucleosome in state 4 structure. Similarly, DNA is overwrapped in the orange and green nucleosomes in Figure 5.5C. State 5 and 6 are the most populated macrostates. State 5 adopts α-shaped conformations without stackings (Figure 5.5D). Such conformations are favored because DNA curves and inter-nucleosomal interactions are balanced. State 6 has more contacts between i and $i + 2$ nucleosomes, but perfectly aligned stackings are still topologically disfavored due to the additional 5 bp.

In the NRL = 177 system, with the longer linker of $10n$-bp pattern, the folded minimum basin clearly reappears, and the free energy landscape becomes similar to that of the NRL = 167 system, though slightly more extended. The six-state non-Markovian dynamics model reveals similar patterns to NRL = 167 single tetra-nucleosome system (Figure D.22). For example, state 2 and 4 in NRL = 177 system correspond to state 3 and 4 in NRL = 167 system, respectively. Due to the slightly increased flexibility and electrostatic repulsion caused by additional 10-bp, the native state (state 6) is slightly more destabilized, while some partially-unfolded states thrive with less well-aligned stackings, consistent with in vitro experiments [285]. This increased flexibility also decrease the folding rate towards the native state compared to the NRL = 167 system. This suggests that $10n$ linkers lead to similar topological restraints, but longer linkers (i.e. larger integer n) enhance flexibility and slightly

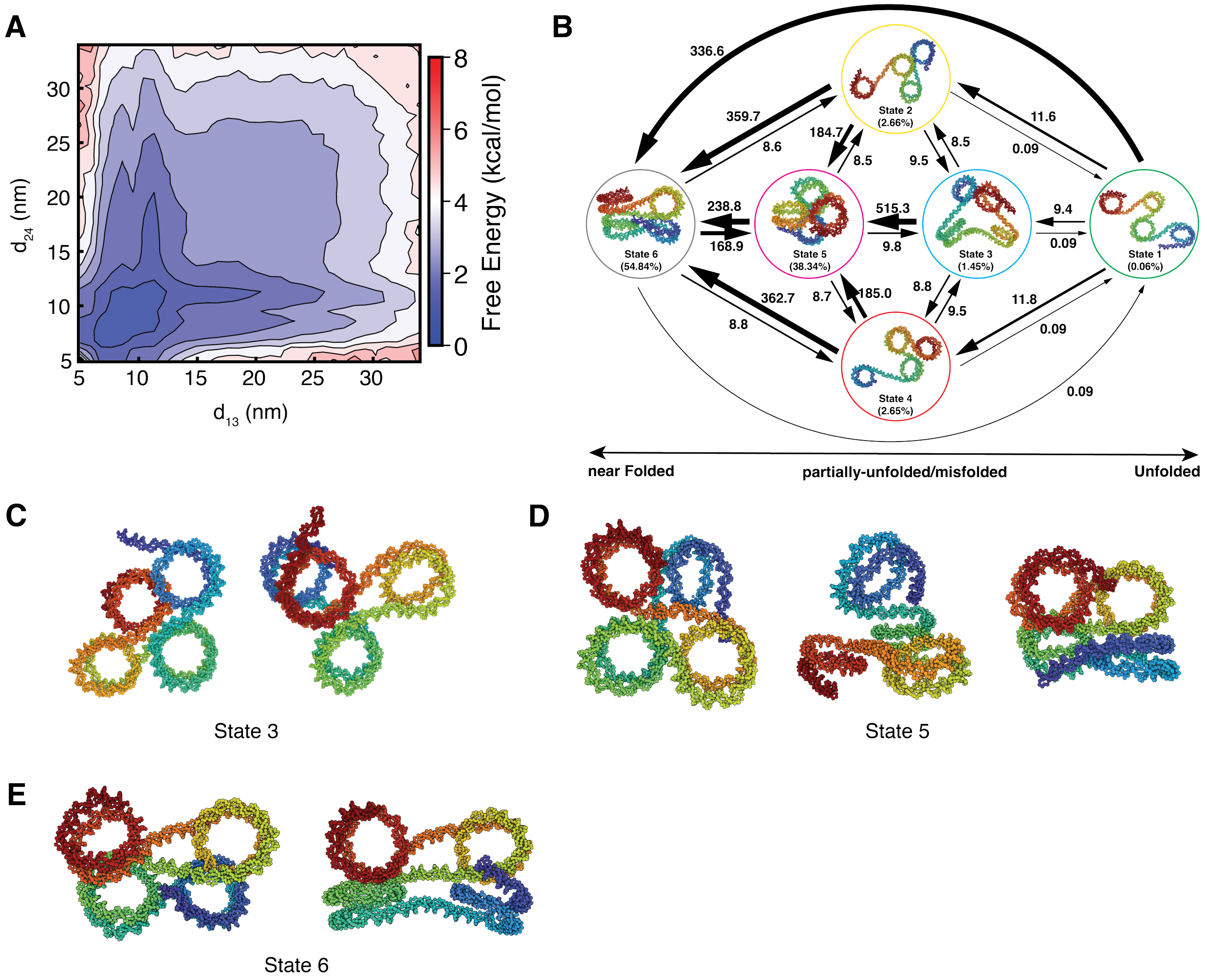


Figure 5.5: **Analysis results of the NRL = 172 tetra-nucleosome system.** (A) The free energy profile along d_{13} and d_{24}. (B) Macrostate non-Markovian dynamics model with inverse MFPT labeled in unit $(10\ \mu s)^{-1}$. (C)-(E) More representative structures from most populated microstates of macrostate 3, 5, and 6, respectively. Histones are hidden for clarity.

extend. Interestingly, although increasing flexibility seems to be correlated with more active genome regimes, longer DNA may lead to more silent genes [284], possibly due to more binding space for histone H1 [286]. So that we can see that linker lengths also affect chromatin structure by mediating protein binding with chromatin.

5.4 Discussion

We combined CG modeling, long-time MD simulations, and transition path analysis and non-Markovian dynamic modeling techniques to explore the folding dynamics of tetra-nucleosome with different linker lengths and within a biologically relevant concentration. Our condensate simulation shows configurations similar to observations by ChromEMT and cryo-ET [13], [25]. Specifically, the chromatin is distributed heterogeneously with diverse contact patterns, but lacks well-stacked fibril structures, consistent with the findings that ordered 30-nm fiber lacks in vivo. By comparing the dynamics of the 20-bp individual tetra-nucleosome (similar to traditional in vitro study condition with low concentration) and the same tetra-nucleosome in biologically relevant condensate environment, we found the unfolding dynamics of tetra-nucleosome, which is the basic unit of genome organization, is significantly accelerated relative to the folding dynamics. This quantitatively reveals the liquid-like dynamic nature of chromatin in condensate. This canonical liquid-like property of chromatin condensate has been validated by some in vitro condensate studies [17], [18], and such intrinsic physical chemistry interactions drive chromatin organizations [2], [256], [258]. The liquid-like property at nucleosomal oligomer level can be related to diverse processes and functions, such as chromatin remodeling, loop extrusion, and transcription, since biomolecules can diffuse rapidly within the liquid-like condensate, and fast local unfolding facilitates the binding of other molecules and nucleosomal disassembly [7], [255], [287], [288]. The relatively rapid local dynamics of biomolecular condensates is actually general as validated by the experiments, although some large-scale motions are slower in condensates [289]. Meanwhile, linker lengths, especially $10n$-bp and $10n + 5$-bp, exhibit strikingly different patterns of chromatin structures due to the linker twist. Such topological effects caused by an additional 5 bp are most significant for short linkers, as longer linkers can enhance flexibility and extend chromatin [285]. The polymeric and ampholytic nature of chromatin can lead to viscoelasticity and gel-like properties [2], [290]. Therefore, chromatin may show more solid-like properties at larger scales [19], [20].

It is biologically significant to understand the inter-nucleosomal interactions and the polymeric nature of the chromatin. In vitro yeast chromatin reconstitution has shown that specific linker patterns controlled by remodelers are sufficient to form chromatin domains and boundaries, while loop extrusion and transcription are not required [64], [291]. This highlights that inter-nucleosomal interactions and polymeric effects play the most fundamental role in genome organization, as they are sufficient to reproduce many genome organization features. PTM can further perturb and diversify inter-nucleosomal interactions, thus acting as an additional layer of control factors [2], [258]. Meanwhile, linker DNA plays the role more than spatial joints. Regulated linker lengths directly control the nucleosomal orientations and binding positions of other molecules [286], thus further governing the genome states and functions [284]. Meanwhile, other molecules such as linker histones, chromatin remodelers, transcription factors, and multivalent cations can also affect chromatin organizations [7], [11], [255], [287], [288], [292]–[294]. These molecules can not only mediate inter-nucleosomal interactions and linker patterns, but also actively change the chromatin organizations. All these factors collectively organize the genome in a complicated mechanism.

Appendix A

Supplementary information for chapter 2

A.1 Training Methods

A.1.1 Proof of contrastive learning loss

Contrastive learning can be formulated as logistic regression [44], [45]. Given N_0 samples from distribution p_0 and N_1 samples from distribution p_1, there are $N = N_0 + N_1$ samples in all. If we mix these N samples together and randomly pick a sample x_i from them, the probability that x_i comes from p_0 is

$$\begin{aligned} P(x_i \in X_0|x_i) =& \frac{P(x_i|x_i \in X_0)P(x_i \in X_0)}{P(x_i|x_i \in X_0)P(x_i \in X_0) + P(x_i|x_i \in X_1)P(x_i \in X_1)} \\ =& \frac{p_0(x_i)N_0/N}{p_0(x_i)N_0/N + p_1(x_i)N_1/N} \\ =& \frac{1}{1 + \nu^{-1}p_1(x_i)/p_0(x_i)} \end{aligned} \tag{A.1}$$

here X_0 and X_1 means the collection of samples from p_0 and p_1, and $\nu = N_0/N_1$. Similarly we can get

$$P(x_i \in X_1|x_i) = \frac{1}{1 + \nu p_0(x_i)/p_1(x_i)} \tag{A.2}$$

In the context of contrastive learning, p_0 is the noise distribution, while p_1 is the data distribution. p_1 is unknown and we intend to learn the reduced energy $u_1(x;\theta)$ such that $p_1(x;\theta) \propto e^{-u_1(x;\theta)}$ (θ represents parameters). On the other hand, $p_0 \propto e^{-u_0(x)}$ is the noise distribution with known reduced energy u_0. The log-likelihood of correctly distinguish data from noise

$$l = \frac{1}{N}\left[\sum_{i=1}^{N_0} \log P(x_i^{(0)} \in X_0 | x_i^{(0)}) + \sum_{i=1}^{N_1} \log P(x_i^{(1)} \in X_1 | x_i^{(1)})\right] \tag{A.3}$$

where samples from p_0 are $x_i^{(0)}$ and samples from p_1 are $x_i^{(1)}$. During training, we aim to maximize the log-likelihood.

In practice, the normalization factors of p_0 and p_1, which are the partition functions, cannot be solved. By defining the reduced free energy as $f_i = -\log \int e^{-u_i(x)} dx$ ($i = 0$ or 1), we have probability density $p_i(x) = e^{-u_i(x)+f_i}$. Thus we have

$$\begin{aligned} l = \frac{1}{N}\Bigg[\sum_{i=1}^{N_0} \log \frac{1}{1+\nu^{-1}\exp(-u_1(x_i^{(0)}) + u_0(x_i^{(0)}) + \Delta f)} \\ + \sum_{i=1}^{N_1} \log \frac{1}{1+\nu \exp(u_1(x_i^{(1)}) - u_0(x_i^{(1)}) - \Delta f)}\Bigg] \end{aligned} \tag{A.4}$$

where $\Delta f = f_1 - f_0$. This is the contrastive log-likelihood. In practice, u_0 is known, while u_1 is the potential including parameters θ to be optimized. The negative of the log-likelihood can be written as the binary cross-entropy (BCE) with logit loss $\mathcal{L}_{\text{BCEWithLogit}}$, which is the contrastive loss $\mathcal{L}$

$$\mathcal{L}(\theta, \Delta f) = \mathcal{L}_{\text{BCEWithLogit}}(\theta, \Delta f) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log \sigma(\alpha_i) + (1-y_i)\log(1-\sigma(\alpha_i))\right] \tag{A.5}$$

where y_i is the binary label as 0 or 1, indicating whether sample x_i comes from distribution p_0 or p_1. σ is the sigmoid function $\sigma(\alpha) = 1/(1+\exp(-\alpha))$, and $\alpha_i = -\log \nu + u_0(x_i) - u_1(x_i, \theta) + \Delta f$ (α_i is called logit). Thus maximizing log-likelihood in equation A.4 is equivalent to minimizing contrastive loss in equation A.5 over θ and Δf. Notably, when doing optimization, we view Δf as an independent parameter, though it depends on u_0 and u_1.

In the following subsection, we will show the feasibility of viewing Δf as an independent parameter.

A.1.2 Prove the feasibility of viewing $\boldsymbol{\Delta f}$ as an independent variable

Here we prove the feasibility of viewing Δf as an independent parameter to optimize, though it depends on u_0 and u_1. To show the feasibility, we need to prove the Δf achieved by maximizing the log-likelihood in equation A.4 is the same as Δf computed with MBAR equation [76], [295], which is the canonical method to compute Δf based on the sample weights. We first derive MBAR equation for the general case of multiple thermodynamic states, then show when there are only 2 thermodynamic states corresponding to the noise and data, Δf given by MBAR equation is the same as the one given by maximizing the log-likelihood in equation A.4.

Suppose we collect samples from M thermodynamic states, and each state is characterized with the reduced potential energy as u_i. For state i, we collect N_i samples. The collection of all the samples is defined as a generalized ensemble $G = \{x_{j=1,\dots,N_i}^{(i=1,\dots,M)}\}$ and we can view that the samples come from such distribution

$$P(\lambda = i, x) \propto \exp\left(-u_i(x) - v_i\right) \tag{A.6}$$

where λ is the state index, and v_i is a reweight factor to be determined that balances the relative weight of different states. Note this form ensures $P(x|\lambda = i) \propto \exp(-u_i(x))$. Given a point x, the probability that it comes from the i-th state is

$$\begin{aligned} P(\lambda = i|x) =& \frac{P(\lambda = i, x)}{\sum_{j=1}^{M} P(\lambda = j, x)} \\ =& \frac{\exp(-u_i(x) - v_i)}{\sum_{j=1}^{M} \exp(-u_j(x) - v_j)} \end{aligned} \tag{A.7}$$

If we randomly pick a sample x from G (i.e. $P(x) = 1/N$), the probability that the sample comes from the i-th state is $P(\lambda = i) = N_i/N$, thus we have

$$\begin{aligned} P(\lambda = i) =& \sum_{x\in G} P(\lambda = i|x)P(x) \\ \Rightarrow \frac{N_i}{N} =& \frac{1}{N}\sum_{x\in G}\frac{\exp(-u_i(x) - v_i)}{\sum_{j=1}^{M}\exp(-u_j(x) - v_j)} \\ \Rightarrow N_i =& \sum_{x\in G}\frac{\exp(-u_i(x) - v_i)}{\sum_{j=1}^{M}\exp(-u_j(x) - v_j)} \end{aligned} \tag{A.8}$$

This is the MBAR equation and v_i can be solved iteratively. Alternatively, solving equation A.8 is equivalent to minimizing loss $\mathcal{L}_{\text{MBAR}}(v_1, \ldots, v_M)$ defined as

$$\mathcal{L}_{\text{MBAR}}(v_1, \ldots, v_M) = \frac{1}{N}\sum_{x\in G}\log\left(\sum_{i=1}^{M}\exp(-u_i(x) - v_i)\right) + \sum_{i=1}^{M}\frac{N_i v_i}{N} \tag{A.9}$$

So far we have shown the canonical way to compute the relative weight of samples. Now consider the context of contrastive learning, where state 0 is the noise ensemble, and state 1 is the data. With MBAR equation we have

$$\begin{aligned} N_0 =& \sum_{x\in G}\frac{\exp(-u_0(x) - v_0)}{\sum_{j=0,1}\exp(-u_j(\mathbf{x}) - v_j)} \\ N_1 =& \sum_{x\in G}\frac{\exp(-u_1(x) - v_0)}{\sum_{j=0,1}\exp(-u_j(\mathbf{x}) - v_j)} \end{aligned} \tag{A.10}$$

Note these two equations are dependent on each other as the sum of them is an identical equation $N_0 + N_1 = N$. We can define $b(x) = \exp(u_1(x) + v_1 - u_0(x) - v_0)$, and we get

$$N_0 = \sum_{x\in G}\frac{b(x)}{1 + b(x)} \tag{A.11}$$

Now we want to compare the free energy of two states. Without adding factor v_i, the reduced free energy is f_i. By adding factor v_i to u_i, the reduced free energy is increased to $f_i' = f_i + v_i$.

f'_i reflects the relative population of the states within the generalized ensemble, and f'_i is

$$f'_i = -\log P(\lambda = i) = -\log \frac{N_i}{N} \tag{A.12}$$

So in the context of contrastive learning with only 2 states, $f'_0 = f_0 + v_0$ and $f'_1 = f_1 + v_1$, thus

$$\Delta f = f_1 - f_0 = f'_1 - f'_0 + v_0 - v_1 = \log \nu + v_0 - v_1 \tag{A.13}$$

recall that $\nu = N_0/N_1$. So that

$$b(x) = \exp(u_1(x) + v_1 - u_0(x) - v_0) = \nu \exp(u_1(x) - u_0(x) - \Delta f) \tag{A.14}$$

so we can rewrite the log-likelihood in equation A.4 as

$$l = \frac{1}{N}\left(\sum_{i=1}^{N_0} \log \frac{b(x_i^{(0)})}{1 + b(x_i^{(0)})} + \sum_{i=1}^{N_1} \log \frac{1}{1 + b(x_i^{(1)})}\right) \tag{A.15}$$

Since $\partial b(x)/\partial \Delta f = -b(x)$, to maximize log-likelihood, we have

$$\frac{\partial l}{\partial \Delta f} = \frac{1}{N}\left(-\sum_{i=1}^{N_0} \frac{1}{1 + b(x_i^{(0)})} + \sum_{i=1}^{N_1} \frac{b(x_i^{(1)})}{1 + b(x_i^{(1)})}\right) \tag{A.16}$$

when log-likelihood is maximized, $\partial l/\partial \Delta f$ is equal to 0, this leads to

$$\begin{aligned} \sum_{i=1}^{N_0} \frac{1}{1 + b(x_i^{(0)})} &= \sum_{i=1}^{N_1} \frac{b(x_i^{(1)})}{1 + b(x_i^{(1)})} \\ \Rightarrow \sum_{i=1}^{N_0} \frac{1}{1 + b(x_i^{(0)})} + \sum_{i=1}^{N_0} \frac{b(x_i^{(0)})}{1 + b(x_i^{(0)})} &= \sum_{i=1}^{N_1} \frac{b(x_i^{(1)})}{1 + b(x_i^{(1)})} + \sum_{i=1}^{N_0} \frac{b(x_i^{(0)})}{1 + b(x_i^{(0)})} \\ \Rightarrow N_0 &= \sum_{x \in G} \frac{b(x)}{1 + b(x)} \end{aligned} \tag{A.17}$$

Here sum over $x \in G$ means sum over all the samples in the data and noise ensembles. This result is consistent with the results achieved from MBAR equation (equation A.11), indicating the by viewing Δf as an independent parameter to be optimized, when log-likelihood is maximized (equivalent to minimizing contrastive loss), Δf obtained from the optimization is consistent with the MBAR equation, which rationalizes the workaround to view Δf as an independent parameter during training.

It is also worth mentioning that the conclusion remains valid even if we add a regularizer on θ. With the regularizer $\mathcal{R}(\theta)$, we aim to maximize $l - \mathcal{R}(\theta)$ or equivalently minimize $\mathcal{L} + \mathcal{R}(\theta)$, and all the derivations shown before are still valid.

A.1.3 A sufficient condition for the convexity of the contrastive loss

Theoretically, given a functional with sufficient capacity, infinite data and noise samples, and non-zero noise density where data density is non-zero, the learned potential is the exact data potential [45]. However, in practice, limited functional forms with finite parameters can possibly lead to local optimal solutions [44], [45]. Here we compute the second-order derivatives of the loss function to show the sufficient condition under which the contrastive loss is convex, thus leading to a unique solution.

We begin with the contrastive loss (equation A.5). In the context of contrastive learning, noise samples are produced with known potential energy u_0, while potential u_1 depends on parameters θ, which remain to be optimized. We define

$$g(x, y) = -y \log \sigma(\alpha) - (1 - y) \log(1 - \sigma(\alpha)) \tag{A.18}$$

where $\alpha = -\log \nu + u_0(x) - u_1(x; \theta) + \Delta f$. Thus contrastive loss $\mathcal{L} = (\sum_{i=1}^{N} g(x_i, y_i))/N$. The first order derivative of g to θ_i is

$$\frac{\partial g}{\partial \theta_i} = \begin{cases} (1 - \sigma(\alpha)) \dfrac{\partial u_1}{\partial \theta_i} & \text{if } y = 1 \\ -\sigma(\alpha) \dfrac{\partial u_1}{\partial \theta_i} & \text{if } y = 0 \end{cases} \tag{A.19}$$

and the second order derivative of g to θ_i is

$$\frac{\partial^2 g}{\partial\theta_i\partial\theta_j} = \begin{cases} (1-\sigma(\alpha))\left[\frac{\partial^2 u_1}{\partial\theta_i\partial\theta_j} + \sigma(\alpha)\left(\frac{\partial u_1}{\partial\theta_i}\right)\left(\frac{\partial u_1}{\partial\theta_j}\right)\right] & \text{if } y = 1 \\ \sigma(\alpha)\left[-\frac{\partial^2 u_1}{\partial\theta_i\partial\theta_j} + (1-\sigma(\alpha))\left(\frac{\partial u_1}{\partial\theta_i}\right)\left(\frac{\partial u_1}{\partial\theta_j}\right)\right] & \text{if } y = 0 \end{cases} \tag{A.20}$$

If $\frac{\partial^2 u_1}{\partial\theta_i\partial\theta_j} = 0$, which is the case when u_1 is linear function of θ_i, then

$$\frac{\partial^2 g}{\partial\theta_i\partial\theta_j} = \sigma(\alpha)(1-\sigma(\alpha))\left(\frac{\partial u_1}{\partial\theta_i}\right)\left(\frac{\partial u_1}{\partial\theta_j}\right) \qquad y = 0 \text{ or } 1 \tag{A.21}$$

Meanwhile, consider the derivatives involving Δf, we have first order derivative

$$\frac{\partial g}{\partial\Delta f} = \begin{cases} \sigma(\alpha) - 1 & \text{if } y = 1 \\ \sigma(\alpha) & \text{if } y = 0 \end{cases} \tag{A.22}$$

for second order derivatives,

$$\frac{\partial^2 g}{\partial\Delta f\partial\theta_i} = -\sigma(\alpha)(1-\sigma(\alpha))\frac{\partial u_1}{\partial\theta_i} \qquad y = 0 \text{ or } 1 \tag{A.23}$$

and

$$\frac{\partial^2 g}{\partial\Delta f^2} = \sigma(\alpha)(1-\sigma(\alpha)) \qquad y = 0 \text{ or } 1 \tag{A.24}$$

Now we can write out the Hessian matrix of g with respect to $\theta_1, \ldots, \theta_n, \Delta f$. We define the column vector

$$v = \left(\frac{\partial u_1}{\partial\theta_i}, \ldots, \frac{\partial u_N}{\partial\theta_N}, -1\right)^T \tag{A.25}$$

so the Hessian matrix is

$$H = \sigma(\alpha)(1-\sigma(\alpha))vv^T \tag{A.26}$$

since $0 < \sigma(\alpha) < 1$, it is obvious that H is positive semi-definite. Thus u_1 being a linear function of θ is a sufficient condition for $g(x, y)$ being a convex function relative to θ and

Δf, and the contrastive loss $\mathcal{L}$, which is the mean of $g(x_i, y_i)$, is also convex. Minimizing contrastive loss should lead to a unique optimal solution. Furthermore, if the regularizer $\mathcal{R}(\theta)$ is also convex to θ, then $\mathcal{L} + \mathcal{R}$ is also convex with respect to θ and Δf, and the conclusion remains.

A.1.4 Generate the mixed noise ensemble

In practice, we ran multiple independent umbrella simulations with known reduced energy to generate noise, so the noise samples cover diverse configurations. To correctly reweight the noise samples and view them together as coming from a noise potential u_0, we need to use the idea of generalized ensemble and MBAR equation introduced before. In this subsection, we describe how to correctly compute u_0, and the sampling procedures are explained in section *Details of noise simulations.*

Suppose that we performed M independent noise samplings, each with reduced potential $u_i^{\text{noise}}(x)$ $(i = 1, \ldots, M)$. In our case, we applied umbrella sampling [75] to collect noise samples, so $u_i^{\text{noise}}(x) = u_{\text{unbiased}}^{\text{noise}}(x) + b_i(x)$, where $u_{\text{unbiased}}^{\text{noise}}(x)$ is the reduced energy of the unbiased system, and $b_i(x)$ is the i-th reduced umbrella bias. Other enhanced sampling methods, such as replica exchange [128], can also be applied to enhance noise sampling. N_i^{noise} noise samples are collected in the i-th sampling. All the noise samples are mixed together as a generalized noise ensemble $G^{\text{noise}} = \{x_{j=1,\ldots,N_i^{\text{noise}}}^{\text{noise},i=1,\ldots,M}\}$. The samples can be viewed as sampled from distribution

$$P(\lambda = i, x) \propto \exp(-u_i^{\text{noise}}(x) - v_i) \tag{A.27}$$

and v_i can be solved by minimizing MBAR loss as shown in equation A.9. The samples in the generalized noise ensemble can be viewed as coming from probability

$$p_0(x) = \sum_{i=1}^{M} P(\lambda = i, x) \propto \sum_{i=1}^{M} \exp(-u_i^{\text{noise}}(x) - v_i) \tag{A.28}$$

so the reduced potential of the noise ensemble should be

$$u_0(x) = -\log p_0(x) = -\log\left(\sum_{i=1}^{M} \exp(-u_i^{\text{noise}}(x) - v_i)\right) \tag{A.29}$$

A.1.5 Validate and filter hyperparameters with reweighting scheme

In practice, since we only have finite data and noise samples, and the model has finite capacity with finite parameters, the trained model deviates from the ground truth [45]. We applied regularization term $\mathcal{R}$ over parameters θ to control the scale of interactions strengths, and the regularizer strength is controlled by hyperparameters ζ

$$\mathcal{R}(\theta) = \frac{\zeta}{2}\text{mean}(\theta^2) \tag{A.30}$$

In contrast to maximum log-likelihood, which performs sampling during training, contrastive learning does not have direct control over the samples generated with the trained model during training, so that we need to run simulations with the trained model to validate and choose the optimal hyperparameters, which is computationally expensive. A practical workaround is to use the reweighting scheme. By reweighting the noise ensemble with the trained model, we can rapidly estimate the average radius of gyration (R_g) produced with the trained model to select the best hyperparameters. The workflow is essentially the same as estimating the free energy along R_g with the trained model and the noise ensemble with the free energy perturbation method [296] and can be executed with FastMBAR [76]. Suppose u_0 is the noise ensemble reduced energy, and $u_1(x;\theta^*)$ is the potential with optimized parameters θ^*, the reweighting factor should be

$$w(x) = \exp(-u_1(x;\theta^*) + u_0(x)) + \text{const.} \tag{A.31}$$

which has an arbitrary constant factor that does not affect the result but in practice affects numerical accuracy. By dividing the R_g axis as discrete bins, the reduced free energy along R_g in the i-th bin under potential $u_1(x;\theta^*)$ is

$$f_i = -\log\left(\frac{1}{N_0}\sum_{j=1}^{N_0} w(x_j^{(0)}) I_i(\hat{R}_g(x_j^{(0)}))\right) \tag{A.32}$$

where $I_i(R_g)$ is the indicator function and equals 1 if R_g is within the i-th bin, otherwise the indicator function equals 0. $x_j^{(0)}$ is the j-th sample of the noise ensemble. $\hat{R}_g$ represents the function that computes R_g as the function of the sample configuration including n atoms

$$\hat{R}_g(x) = \sqrt{\frac{\sum_{i=1}^{n} m_i |\vec{r}_i - \vec{r}_{\mathrm{COM}}|^2}{\sum_{i=1}^{n} m_i}} \tag{A.33}$$

where $\vec{r}_i$ is the coordinate of the i-th atom, and $\vec{r}_{\mathrm{COM}}$ is the center-of-mass coordinate

$$\vec{r}_{\mathrm{COM}} = \frac{\sum_{i=1}^{n} m_i \vec{r}_i}{\sum_{i=1}^{n} m_i} \tag{A.34}$$

The average R_g is

$$\bar{R}_g = \frac{\sum_{i}^{N_{\mathrm{bins}}} R_g^{(i)} \exp(-f_i)}{\sum_{i}^{N_{\mathrm{bins}}} \exp(-f_i)} = \frac{\sum_{i}^{N_{\mathrm{bins}}} \sum_{j=1}^{N_0} R_g^{(i)} w(x_j^{(0)}) I_i(\hat{R}_g(x_j^{(0)}))}{\sum_{i}^{N_{\mathrm{bins}}} \sum_{j=1}^{N_0} w(x_j^{(0)}) I_i(\hat{R}_g(x_j^{(0)}))} \tag{A.35}$$

where $R_g^{(i)}$ is the center value of the i-th bin, and there are N_{bins} in all. We chose the model with the estimated $\bar{R}_g$ that best matches the benchmark. We ended up with MD simulations to confirm that the $\bar{R}_g$ values computed with MD simulations match the training data or experiments. Figure A.2 shows the average R_g estimated by the reweighting scheme is close to the one achieved by MD simulation, supporting the effectiveness of the reweighting scheme. The key to successful reweighting lies in the decent overlap between the noise ensemble and the ensemble produced by the trained model, which is almost guaranteed in our case, since we performed multiple umbrella simulations to explore diverse noise configurations, and the noise potential is similar to the trained potential. In particular, this reweighting scheme enables comparison of any collective variable and is compatible with any training method that avoids sampling during training to accelerate validation.

Finally, we underscore that the reweighting scheme is only for rapidly filtering hyperparameters, and all the results displayed are computed with MD simulations unless otherwise specified.

A.2 Force Field Definitions

In this section, we elaborate the force field potential functional forms used as our trained model. The noise potential is produced with existing models and is explained in section *Details of noise simulations*.

A.2.1 Force field of intrinsically disordered proteins

For the force field for intrinsically disordered proteins (IDPs), we used the HPS model [56], which is a coarse-grained (CG) potential of Cα representation (i.e. one CG bead per amino acid at Cα). We summarize the model below.

The force field for IDPs is represented as

$$U_{\mathrm{IDP}} = U_{\mathrm{bond}} + U_{\mathrm{AH}} + U_{\mathrm{elec}} \tag{A.36}$$

which is the same functional form as the HPS model. The bonds between two residues are harmonic bonds

$$U_{\mathrm{bond}} = \sum_i \frac{k_{\mathrm{bond}}}{2}(r_{i,i+1} - r_0) \tag{A.37}$$

where $r_{i,i+1}$ is the distance between two connected CG beads, and r_0 is the bond length. Based on the mean bond lengths measured from the all-atom trajectories, we chose $r_0 = 0.386$ nm. We set $k_{\mathrm{bond}} = 8000$ kJ/mol/nm^2 so that the simulation timestep can be 10 fs.

The nonbonded contact between two CG beads is the Ashbaugh-Hatch functional form [74]

$$U_{\mathrm{AH}}(r) = \sum_{i<j} \begin{cases} U_{\mathrm{LJ}}(r_{ij}) + (1 - \lambda_{ij})\epsilon_{\mathrm{LJ}} & \text{if } r_{ij} \leq 2^{1/6}\sigma_{ij} \\ \lambda_{ij} U_{\mathrm{LJ}}(r_{ij}) & \text{otherwise} \end{cases} \tag{A.38}$$

where U_{LJ} is the Lennard-Jones (LJ) potential

$$U_{\mathrm{LJ}}(r_{ij}) = 4\epsilon_{\mathrm{LJ}} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \tag{A.39}$$

Here r_{ij} is the distance between the two Cα atoms. λ_{ij} is the hydrophobic scale parameter that captures the interaction strength between two CG atoms and depends on the types of amino acids. Interaction strength parameter $\epsilon_{\mathrm{LJ}} = 0.2$ kcal/mol. σ_{ij} is the size parameter. Both λ_{ij} and σ_{ij} depend on atom types. The cutoff distance is $4\sigma_{ij}$. When computing the nonbonded contact, bonded atom pairs (i.e. i and $i+1$ pairs on the same chain) are excluded. We directly applied the σ_{ij} values used in existing HPS series models, while λ_{ij} were optimized with contrastive learning. Notably, U_{AH} is a linear function of λ_{ij}, which ensures the contrastive loss is convex.

The electrostatic interaction is defined as

$$U_{\mathrm{elec}} = \sum_{i<j} \frac{q_i q_j}{4\pi\epsilon_0 \epsilon_{\mathrm{water}} r_{ij}} \exp(-r_{ij}/\lambda_D) \tag{A.40}$$

where q_i and q_j are charges of two CG atoms, ϵ_0 is the vacuum permittivity, $\epsilon_{\mathrm{water}} = 80.0$ is the dielectric, and λ_D is the Debye length

$$\lambda_D = \sqrt{\frac{k_B T \epsilon_0 \epsilon}{2 N_A I e_c^2}} \tag{A.41}$$

where k_B is the Boltzmann constant, T is the temperature, N_A is the Avogadro constant, I is the ionic strength, and e_c is the elementary charge. Charges of residues on the N- and C-terminal ends of the chains are added and subtracted by 1 charge, respectively. The cutoff distance is $5\lambda_D$, which ensures the cutoff error is small enough. Compared to using fixed cutoff length such as 3.5 nm or 4.0 nm, this setting significantly decreases the cutoff error for some systems with ionic strength much lower than the physiological condition (about 150 mM). Bonded atom pairs (i.e. i and $i+1$ pairs on the same chain) are excluded from the electrostatic interactions.

A.2.2 Force field of ordered proteins

For the force field of ordered proteins, we generalized the HPS model by adding structure-based terms, thus capturing the folding tendency of ordered proteins (OPs). The overall potential is

$$U_{\text{OP}} = U_{\text{bond}} + U_{\text{angle}} + U_{\text{dihedral}} + U_{\text{native pair}} + U_{\text{AH}} + U_{\text{elec}} \tag{A.42}$$

The bond potential U_{bond} is the same as the one used for IDPs (equation A.37) with the same parameters $r_0 = 0.386$ nm and $k_{\text{bond}} = 8000$ kJ/mol/nm^2.

The angle, dihedral, and native pair potentials are parameterized with the given reference structures, which should be the native structure, thus stabilizing the ordered protein. The method to select the reference structure from all-atom trajectories is elaborated in *Section: Details of noise simulations for ordered proteins*.

The angle potential is

$$U_{\text{angle}} = k_{\text{angle}} M^2 \left[1 - \cos\left(\frac{\theta - \theta_0}{M}\right)\right] \tag{A.43}$$

where $k_{\text{angle}} = 120$ kJ/mol/rad^2 and $M = 5$. When $\theta - \theta_0 \in [-\pi, \pi]$, such potential is similar to a harmonic potential $k_{\text{angle}}(\theta - \theta_0)^2/2$, but the cosine function facilitates more stable simulations. θ_0 is the angle read from the reference structure.

The dihedral potential is a periodic torsion force

$$U_{\text{dihedral}} = \sum_{n=1,3} k_{\text{dihedral},n}[1 + \cos(n(\phi - \phi_0 - \pi))] \tag{A.44}$$

with periodicity $n = 1$ or 3. $k_{\text{dihedral},1} = 3.0$ kJ/mol and $k_{\text{dihedral},1} = 1.5$ kJ/mol. ϕ is the dihedral, and ϕ_0 is the reference value read from the reference structure.

The native pair potential stabilize continuous ordered α-helices or β-sheets. The native pairs are applied between Cα atoms which satisfy: (1) both Cα atoms are within the same continuous α-helix or β-sheet (this implicitly requires them to be on the same chain), with secondary structures of the reference structure recognized by DSSP [297] implemented in MDTraj [298]; (2) the two Cα atoms are not involved in bond, angle, or dihedral potentials (i.e. the indices of two Cα atoms i and j satisfy $|i-j|>3$ if on the same chain); (3) the two residues of the Cα atoms have contacts in the reference all-atom structure identified by the shadow algorithm [299], which is implemented in OpenABC [61]. The native pair potential is

$$U_{\text{native pair}} = \sum_{\substack{\text{native pairs } (i,j) \\ i<j}} \epsilon_{\text{native pair}} \left[5\left(\frac{\mu_{ij}}{r_{ij}}\right)^{12} - 6\left(\frac{\mu_{ij}}{r_{ij}}\right)^{10} \right] \tag{A.45}$$

where $\epsilon_{\text{native pair}} = 3$ kJ/mol. μ_{ij} is the distance between two Cα atoms in the reference structure.

The functional forms of non-bonded contact U_{AH} and electrostatic interaction U_{elec} are the same as the ones used in IDP force field (equations A.38 and A.40), the parameters are also the same. The only difference is now the atom pairs involved in bond, angle, dihedral, and native pairs (i.e. pairs i and $i+1/2/3$ on the same chain, and native pairs) are all excluded from both non-bonded interactions.

A.3 Training and validation dataset

To train the coarse-grained force field by contrastive learning, we need high-quality molecular configurations as data samples. We used the Cα representations of structures collected from all-atom explicit solvent simulations with known temperatures and ionic strengths. Our noise samples are produced with existing Cα models implemented in OpenABC [55], [57], [61]. Here we provide more details of the data and noise samples.

A.3.1 Details of all-atom simulations

All the training data samples come from long-time unbiased all-atom simulations. For IDPs, all the training data samples come from all-atom simulations with the a99SB-*disp* force field [72], which is the state-of-art all-atom explicit ion force field for IDPs. We include 7 IDPs reported by D. E. Shaw Research [72] and 34 additional IDPs simulated in house. The all-atom simulations of ordered proteins were simulated with either a99SB-*disp*, DES-Amber, or CHARMM22*/TIP3P, which have been reported by D. E. Shaw Research [71]–[73]. Some detailed information of the all-atom simulations are provided in Table A.1 and A.2. Here, we provide details of in-house all-atom simulations of 34 additional IDPs (named as Evo dataset).

The sequences of the 34 IDPs simulated in house were selected from the *S. cerevisiae* proteome based on the previous research by Zarin et al [78]. They identified over 5000 intrinsically disordered regions in the *S. cerevisiae* proteome and clustered them into 53 clusters based on their evolutionary signatures using a hierarchical clustering algorithm. To ensure a broad representation of IDPs, we selected sequences from each cluster for subsequent simulation, following the process outlined below:

1. Initially, we filtered sequences to a length of 40 amino acids, which is short enough for long time scale all-atom molecular dynamics simulations.

2. To identify sequences with a significant degree of intrinsic disorder, we employed

ColabFold [300] to predict the per-residue pLDDt scores of each sequence. According to previous study, the pLDDt score ≤ 72 can be used to measure whether a sequence is disordered [301]. Sequences were considered sufficiently disordered if over 90% of their residues exhibited a pLDDt score at or below this threshold. This criterion resulted in 47 clusters each containing at least one qualifying sequence.

3. The next step of selection focused on the composition of the amino acids. We observed a deficiency in cysteine (C) and tryptophan (W) within the candidate sequences. To ensure a balanced amino acid combination, which is crucial for training force field, we ranked sequences within each cluster based on the number of C and W. In cases of ties, we employed the number of amino acid types present in the sequence as a secondary ranking criterion. The top-ranked sequence from each cluster was selected, yielding 47 sequences in total.

4. We then conducted 10.5 μs all-atom molecular dynamics simulation for each selected sequence. The structures predicted from the ColabFold were used as the initial structures for the all-atom simulations. We parameterized the a99SB-*disp* force field [72] with the a99SB-*disp* water model using the GROMACS [125] software. Each protein was placed in a dodecahedron box with a length of 8.0 nm. NaCl at a concentration of 150 mM was added to the simulation box. The simulation was performed using the OpenMM [62] package. Following energy minimization, a constant pressure, and temperature (NPT) simulation was carried out for 10.5 μs at a temperature of 300 K and a pressure of 1 Bar. The hydrogen mass repartitioning [302] method was used to enable a timestep of 4 fs. The Langevin middle integrator [130] was used to produce more accurate configurational sampling. Configurations were saved every 100 ps during the simulation, resulting in 105,000 configurations for each sequence. Based on these simulations, sequences exhibiting narrow ranges of the radius of gyration (R_g) were excluded to ensure a diverse set of expanded protein conformations. Sequences that frequently exhibited configurations where the smallest distance between two periodic boundary condition images was shorter than the non-bonded interaction cutoff (0.9 nm) were also excluded to ensure the accuracy of the conformation distributions. Following

the above criteria, we finalized a set including 34 IDP sequences.

A.3.2 Details of noise simulations

Since noise samples have to be generated with known potentials, we produced noise samples with existing Cα CG force fields with small modifications and additional bias. Notably, small modifications of the parameters of the existing force fields are only applied to produce noise samples. When we compare the performance of the trained model and existing force fields, we keep parameters of the existing force fields unchanged unless otherwise specified.

Details of noise simulations for intrinsically disordered proteins

For the noise of IDPs, we produced noise samples with HPS-Urry model [57] with slight modifications and umbrella bias. The overall potential of the i-th umbrella sampling is

$$U^{(i)}_{\text{noise, IDP}} = U_{\text{bond}} + U_{\text{AH}} + U_{\text{elec}} + U^{(i)}_{\text{bias}} \tag{A.46}$$

The bond potential is the harmonic bond potential (equation A.37). The parameters are also $k_{\text{bond}} = 8000$ kJ/mol/nm^2 and $r_0 = 0.386$ nm. We used these values as such parameters were used in our final trained model.

The nonbonded contact potential is the Ashbaugh-Hatch potential in equation A.38 with the ϵ_{LJ} σ_{ij} and λ_{ij} values reported in a previous study [57]. Note the λ_{ij} values are shifted by -0.08 to achieve the optimal results reported before (all the HPS-Urry simulations use such shifted Urry scale) [57]. Atom pairs with bonds were excluded from the nonbonded contact.

The electrostatic interaction is the Debye-Hückel potential in equation A.40. Dielectric constant $\epsilon_{\text{water}} = 80.0$, electrostatic interaction cutoff is 4.0 nm. Charges of Cα atoms on the N-terminal and C-terminal ends have +1 and -1 additional charge, respectively. The Debye length was computed with the ionic strength and temperature same as the ones used in the corresponding all-atom simulations. Similar to the contact potential, atom pairs connected with bonds were excluded.

We applied umbrella bias along R_g to enhance sampling. The umbrella bias is

$$U_{\text{bias}}^{(i)} = \frac{\kappa_i}{2}(\hat{R}_g(\vec{r}) - R_g^{(i)})^2 \tag{A.47}$$

where κ_i and $R_g^{(i)}$ are the force constant and umbrella center of the i-th umbrella sampling. $\hat{R}_g(\vec{r})$ is the function that computes radius of gyration of the configuration defined as

$$\hat{R}_g(\vec{r}) = \sqrt{\frac{\sum_{i=1}^{N} m_i(\vec{r}_i - \vec{r}_{\text{COM}})^2}{\sum_{i=1}^{N} m_i}} \tag{A.48}$$

where $\vec{r}_i$ is the position of the i-th atom, m_i is the mass of the i-th atom, and r_{COM} is the center-of-mass position of the molecule, with all the mass of each amino acid on the Cα atom

$$\vec{r}_{\text{COM}} = \frac{\sum_{i=1}^{N} m_i \vec{r}_i}{\sum_{i=1}^{N} m_i} \tag{A.49}$$

The umbrella bias helps explore both compact and extended configurations, thus facilitate sampling a wide distribution of noise and ideally has a decent overlap with the data sample distribution.

Details of noise simulations for ordered proteins

For the noise of ordered proteins, to keep some overlap between the noise distributions and data distributions, we kept using the AH potential as the nonbonded potential, with structure-based terms to stabilize the structure with respect to the reference native structure. Additionally, umbrella bias on the root-mean square deviation (RMSD) relative to the reference structure is applied to enhance noise sampling.

First, we explain how to select reference structures from the all-atom trajectories of ordered proteins. Reference structures were required to define structure-based potentials (angles, dihedrals, and native pairs) and RMSD umbrella bias. Considering the funneled landscape of folding proteins, we used hierarchical clustering to locate the folded structure with SciPy [303]. Given the all-atom trajectory, we converted the all-atom trajectory to the CA trajectory and computed the RMSD matrix, where the (i, j) element of the matrix is the RMSD between the i-th and j-th snapshots. Hierarchical clustering is then performed on the RMSD matrix using the metric that the distance between two clusters is the average of distances between any two points from each cluster separately. Hierarchical clustering is further converted to flat clustering with the maximal number of clusters as 50. The folded configurations should be the largest flat cluster, and we chose the configuration from the largest cluster that minimizes the mean RMSD with all other members in the largest cluster as the reference structure.

The overall potential for the i-th umbrella sampling is defined as

$$U^{(i)}_{\text{noise, OP}} = U_{\text{bond}} + U_{\text{angle}} + U_{\text{dihedral}} + U_{\text{native pair}} + U_{\text{AH}} + U_{\text{elec}} + U^{(i)}_{\text{bias}} \tag{A.50}$$

The bond potential U_{bond} is still the harmonic bond potential (equation A.37) with the same parameters $k_{\text{bond}} = 8000$ kJ/mol/nm^2 and $r_0 = 0.386$ nm. The angle potential U_{angle} is the cosine function in equation A.43 with the same parameters $k_{\text{angle}} = 120$ kJ/mol/rad^2 and $M = 5$. θ is the angle, and θ_0 is read from the reference structure. The dihedral potential U_{dihedral} is the periodic torsion force in equation A.44 with the same parameters $n = 1$ or 3, $k_{\text{dihedral},1} = 3.0$ kJ/mol and $k_{\text{dihedral},1} = 1.5$ kJ/mol. ϕ is the dihedral, and ϕ_0 is the reference value read from the reference structure. The native pair potential $U_{\text{native pair}}$ is the same as equation A.45, and native pairs were searched in the same way as explained in *Section: Force field of ordered proteins*. The nonbonded contact, electrostatic interaction, and umbrella bias are the same as the ones for producing noise of IDPs, which have been introduced in *Details of noise simulations for intrinsically disordered proteins*.

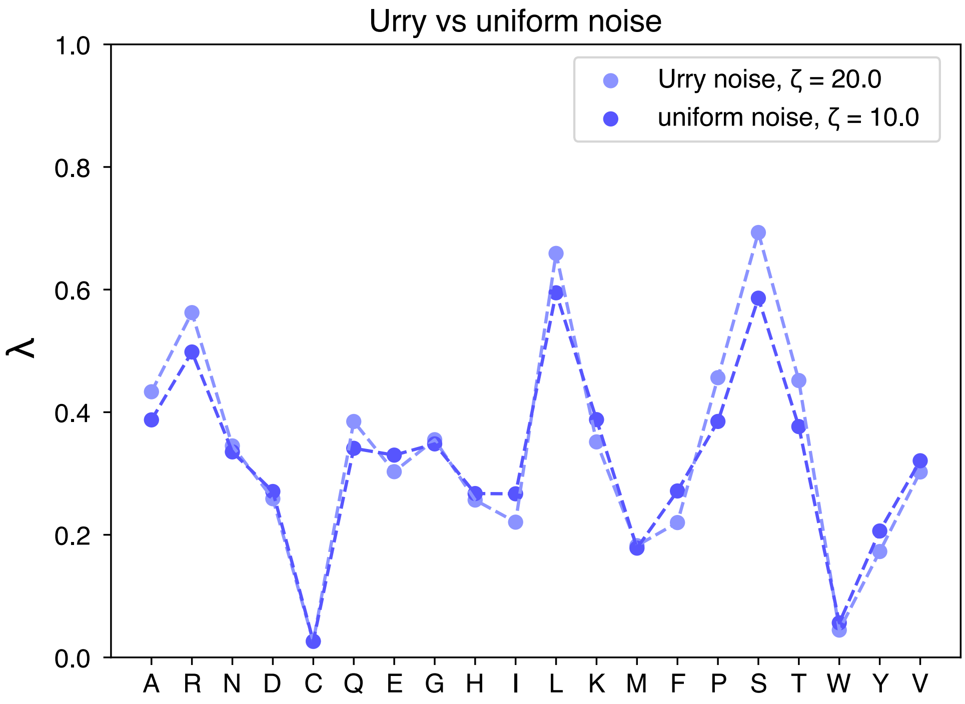


Figure A.1: 20 hydrophobic scale parameters trained on 7 IDPs with noise samples generated with different models. One set of noise samples were generated with the HPS-Urry model with regularization parameter $\zeta = 20.0$, and the other set of noise samples were generated with the same non-bonded potential functional form, but all the λ values are uniformly set as 0.3, with electrostatic cutoff as $5\lambda_D$ (λ_D is the Debye length), and $\zeta = 10.0$. The λ values in the trained model are not sensitive to the exact parameters of the potential that produced the noise.

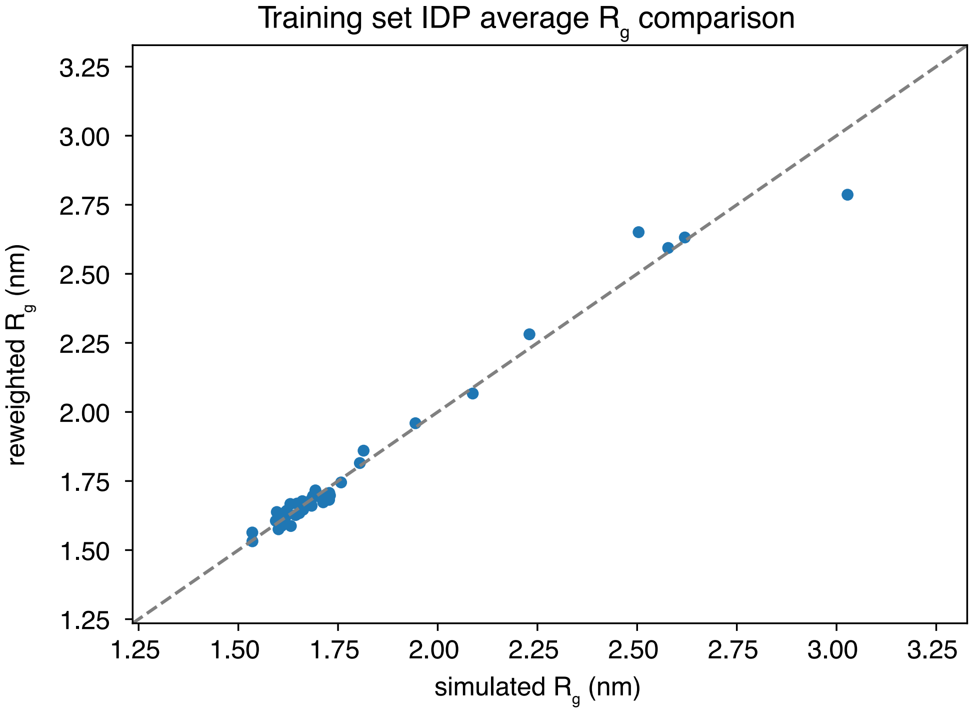


Figure A.2: The comparison of average R_g between the simulation and reweighting on the training set with the model trained on 41 IDPs and 20 hydropathy parameters, which is the same model as shown in main text Figure 2.

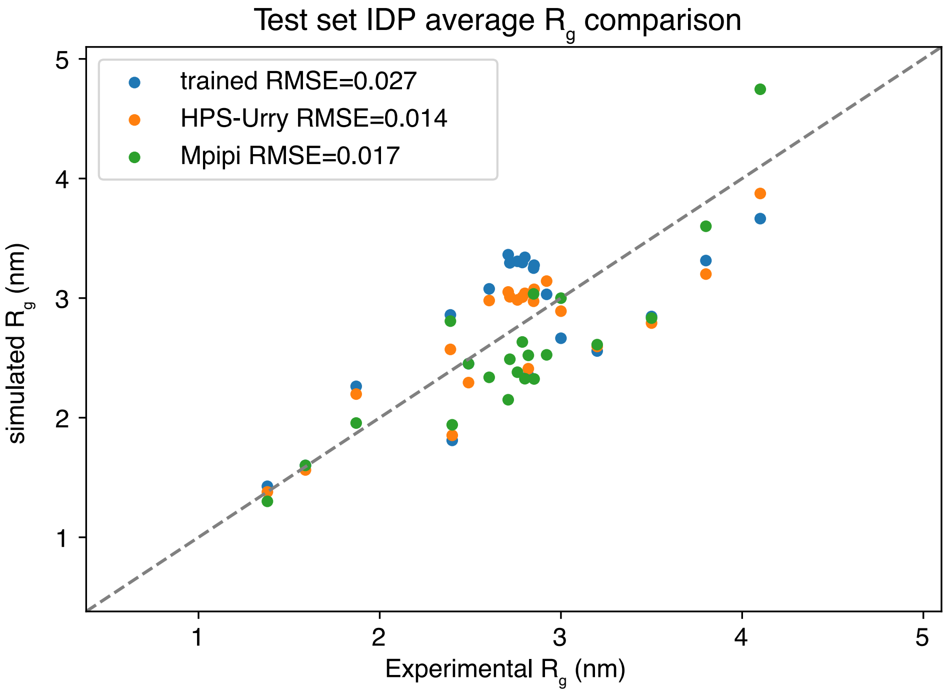


Figure A.3: The comparison of average R_g between several CG models and experimental values on test set IDPs. Here the trained model is trained with 41 IDPs and 20 hydropathy parameters, which is the same as the model shown in main text Figure 2. Root mean square errors (RMSEs) are provided.

Protein	T (K)	I (mM)	t (μs)	N	Force field
ACTR	300	150	30	71	a99SB-*disp*
Aβ40	300	50	30	40	a99SB-*disp*
Ash1	300	150	30	83	a99SB-*disp*
$\mathrm{N_{tail}}$	300	100	30	132	a99SB-*disp*
α-synuclein	300	100	30	140	a99SB-*disp*
drkN SH3	300	50	30	59	a99SB-*disp*
p15PAF	300	50	30	110	a99SB-*disp*
sic1	300	150	30	92	a99SB-*disp*
34 Evo proteins	300	150	10.5	40	a99SB-*disp*

Table A.1: All-atom simulation data of IDPs used for training. “34 Evo proteins" represents 34 IDPs clustered with evolutionary features [78], and each IDP has its own individual trajectory. T is the simulation temperature, I is the ionic strength, t is the trajectory duration, N is the sequence length. All the 34 Evo protein systems share the same T, I, t and N parameters.

Protein	T (K)	I (mM)	t (μs)	N	Force field
α3D	TRE	3.11			DES-Amber
BBA	TRE	33.04	1000		DES-Amber
BBL	TRE	199.58	1000		DES-Amber
engrailed	TRE	280.46	1000		DES-Amber
gpw	TRE	36.54	1000		DES-Amber
λ-repressor	TRE	51.04	1000		DES-Amber
NTL9	TRE	121.46	1000		DES-Amber
Protein B	TRE	51.44	1000		DES-Amber
BPTI	300	37.02			DES-Amber
calmodulin	300	249.98			DES-Amber
Ubiquitin	300	150			a99SB-*disp*
GB3	300	9.98			a99SB-*disp*
HEWL	300	26.63			a99SB-*disp*
Chignolin	340	25.9	106		CHARMM22*/TIP3P
Trp-cage	290	65	208		CHARMM22*/TIP3P
Villin	360	40	120		CHARMM22*/TIP3P
WW domain	360	7.1	1137		CHARMM22*/TIP3P
Homeodomain	360	45	327		CHARMM22*/TIP3P
Protein G	350	100	1154		CHARMM22*/TIP3P

Table A.2: All-atom simulation data of ordered proteins used for training. T is the simulation temperature, I is the ionic strength, t is the trajectory duration, N is the sequence length. TRE means temperature replica exchange simulations.

Appendix B

Supplementary information for chapter 3

B.1 Force Field Definitions

B.1.1 MOFF protein force field

MOFF is a transferable protein force field optimized for both ordered and disordered proteins utilizing the maximum entropy principle and ordered protein folding landscape [55]. Each amino acid is represented with one coarse-grained (CG) bead whose position is defined using the C_α atom from the atomistic structures. MOFF energy function is defined as

$$U_{\text{MOFF}} = U_{\text{bond}} + U_{\text{angle}} + U_{\text{dihedral}} + U_{\text{memory}} + U_{\text{contact}} + U_{\text{electrostatics}} \tag{B.1}$$

The bonded term, U_{bond}, consists of harmonic potentials for distances $r_{i,i+1}$ between bonded nearest neighbor beads

$$U_{\text{bond}} = \sum_i \frac{1}{2} k_{\text{bond}} (r_{i,i+1} - r_0)^2, \tag{B.2}$$

where the equilibrium length $r_0 = 0.38$ nm, and force constant $k_{\text{bond}} = 1000$ kJ/mol/nm^2.

The angular term, U_{angle}, consists of harmonic potentials for angles between nearest neighbor bonds

$$U_{\text{angle}} = \sum_i \frac{1}{2} k_{\text{angle}} (\theta_i - \theta_{i,0})^2, \tag{B.3}$$

where θ_i is the i-th angle. The equilibrium angles, $\theta_{i,0}$, are measured from the input atomistic structure and force constant $k_{\text{angle}} = 120$ kJ/mol/rad^2.

The dihedral term, U_{dihedral}, consists of periodic torsion potentials with periodicity $n = 1$ or 3

$$U_{\text{dihedral}} = \sum_i \sum_{n=1,3} k_{\text{dihedral},n} [1 + \cos(n(\theta_i - \theta_{i,0} - \pi))] \tag{B.4}$$

where θ_i is the i-th dihedral. The equilibrium values, $\theta_{i,0}$, are measured from the input atomistic structure, $k_{\text{dihedral},1} = 3.0$ kJ/mol/rad^2, and $k_{\text{dihedral},3} = 1.5$ kJ/mol/rad^2.

The memory potential, U_{memory}, also known as native pair potential, stabilizes the ordered secondary and tertiary structures of folded protein domains. It is limited to native pairs identified from the initial input native structure using the Shadow Algorithm detailed below. It is defined as

$$U_{\text{memory}} = \sum_{\substack{\text{native pairs } (i,j) \\ i<j}} \epsilon \left[5 \left(\frac{\mu_{ij}}{r_{ij}} \right)^{12} - 6 \left(\frac{\mu_{ij}}{r_{ij}} \right)^{10} \right]. \tag{B.5}$$

μ_{ij} is the distance between Cα atoms from residue i and j measured from the input native structure, and by default $\epsilon = 3$ kJ/mol.

The contact potential, U_{contact}, measures nonbonded interactions between pairs of amino acids. It involves a repulsion term for excluded volume effect and a specific term that measures the energetic cost of bringing a pair of amino acids into contact. The expression is

$$U_{\text{contact}} = \sum_{i<j} \left\{ \frac{\alpha_{ij}}{r^{12}} - \frac{\epsilon_{ij}}{2} [1 + \tanh(\eta(r_0 - r))] \right\} \tag{B.6}$$

$\eta = 7$ nm^{-1} and $r_0 = 0.8$ nm. α_{ij} and ϵ_{ij} are parameters depending on amino acid type i and j. $\alpha_{ij} = \sigma_{ij}^{12} |\epsilon_{ij}|$. $\sigma_{ij} = (\sigma_i + \sigma_j)/2$ where σ_i is the size of amino acid i. The amino acid sizes (i.e. σ_i) are listed in B.1, and parameters ϵ_{ij} are listed in B.2. We used a cutoff distance $r_c = 2$ nm, and the potential is shifted to zero and remains continuous at $r = r_c$ (such shift is not explicitly shown in equation B.6). 1-2, 1-3, 1-4 atom pairs, and native pairs are excluded from the sum.

The electrostatic potential, $U_{\text{electrostatics}}$, is defined with the Debye-Hückel potential with a distance-dependent relative permittivity (dielectric) $\epsilon(r)$ as

$$U_{\text{electrostatics}} = \sum_{i<j} \frac{S(r_{ij}) Q_i Q_j}{4\pi\epsilon_0 \epsilon(r_{ij}) r_{ij}} \exp(-r_{ij}/\lambda_D(r_{ij})), \tag{B.7}$$

with

$$\epsilon(r) = A + \frac{B}{1 + \kappa \exp(-\zeta B r)}, \tag{B.8}$$

and

$$\lambda_D(r) = \sqrt{\frac{k_B T \epsilon_0 \epsilon(r)}{2 N_A c e^2}}. \tag{B.9}$$

Q_i and Q_j denote the charges of residues i and j with values provided in B.1, and ϵ_0 is the vacuum permittivity. 1-2, 1-3, and 1-4 pairs are excluded from the sum, but native pairs are included. The distance-dependent dielectric, $\epsilon(r)$, switches continuously between the value in water (ϵ_{water}) to the one in bulk protein, with $A = -8.5525$, $B = \epsilon_{\text{water}} - A$, $\kappa = 7.7839$, and $\zeta = 0.03627$ nm^{-1}. $\epsilon_{\text{water}} = 78.4$ is the relative permittivity (dielectric) constant of water. λ_D is the Debye length, which also depends on distance. The variables in equation B.9 correspond to the Boltzmann constant (k_B), temperature (T), the Avogadro constant (N_A), the monovalent salt concentration (c), and the elementary charge (i.e. proton charge) (e), respectively.

The switch function gradually turns off the electrostatic interaction within regime $r_1 \leq$

$r \leq r_2$

$$S(r) = \begin{cases} 1 & (r < r_1) \\ \displaystyle\sum_{i=0}^{5} a_n \left(\frac{r - r_1}{r_2 - r_1} \right)^n & (r_1 \leq r < r_2) \;, \\ 0 & (r \geq r_2) \end{cases} \tag{B.10}$$

where $a_0, a_1, \cdots, a_5$ equal 1, 0, 0, -10, 15, and -6, respectively. Thus the switch function equals 1 at $r_1 = 1.2$ nm while 0 at $r_2 = 1.5$ nm.

B.1.2 HPS protein force field

Hydropathy scale (HPS) models [56], [57] are designed to simulate the phase behaviors of intrinsically disordered proteins (IDP). The energy function of these models is defined as

$$U_{\text{HPS}} = U_{\text{bond}} + U_{\text{electrostatics}} + U_{\text{AH}}. \tag{B.11}$$

The bonded term, U_{bond}, consists of harmonic potentials for distances $r_{i,i+1}$ between nearest neighbor beads

$$U_{\text{bond}} = \sum_i \frac{1}{2} k_{\text{bond}} (r_{i,i+1} - r_0)^2, \tag{B.12}$$

where the equilibrium length $r_0 = 0.38$ nm, and force constant $k_{\text{bond}} = 8368$ kJ/mol/nm^2.

The electrostatic potential, $U_{\text{electrostatics}}$, is defined with the Debye-Hückel potential as

$$U_{\text{electrostatics}} = \sum_{i<j} \frac{Q_i Q_j}{4\pi\epsilon_0 \epsilon r} \exp(-r_{ij}/\lambda_D), \tag{B.13}$$

where ϵ_0 is the vaccum permittivity, $\epsilon = 80$ is the relative permittivity of water, and $\lambda_D = 1$ nm is Debye length at the monovalent salt concentration of 100 mM. Q_i and Q_j denote the charges of residues i and j with values provided in B.1.

The AH potential describes amino acid-specific interactions using Ashbaugh and Hatch's functional form [74] as

$$U_{\mathrm{AH}} = \sum_{i<j} \begin{cases} [\phi_{ij}^{\mathrm{LJ}}(r_{ij}) + (1-\lambda_{ij})\epsilon] & r_{ij} \le 2^{1/6}\sigma_{ij} \\ \lambda_{ij}\phi_{ij}^{\mathrm{LJ}}(r_{ij}) & r_{ij} > 2^{1/6}\sigma_{ij}. \end{cases} \tag{B.14}$$

$\lambda_{ij} = \mu\lambda_{ij}^0 - \Delta$ and $\lambda_{ij}^0 = (\lambda_i + \lambda_j)/2$, where λ_i is the normalized hydropathy scale of amino acid i. μ and Δ are the scaling and drift factors, respectively. λ_i represents normalized hydrophobicity scales. By default, the scale by Kapcha and Rossky (KR scale) [304] is applied with $\mu_{\mathrm{KR}} = 1$, $\Delta_{\mathrm{KR}} = 0.0$ and the scale by Urry et al. (Urry scale) [131] is applied with $\mu_{\mathrm{Urry}} = 1$, $\Delta_{\mathrm{Urry}} = 0.08$. These are the optimal scaling and drift factors for the two hydropathy scales [57].

ϕ_{ij}^{LJ} in equation B.14 corresponds to the normal Lennard-Jones potential defined as

$$\phi_{ij}^{\mathrm{LJ}}(r) = 4\epsilon\left[\left(\frac{\sigma_{ij}}{r}\right)^{12} - \left(\frac{\sigma_{ij}}{r}\right)^{6}\right], \tag{B.15}$$

with $\epsilon = 0.8368$ kJ/mol. $\sigma_{ij} = (\sigma_i + \sigma_j)/2$, where σ_i is the size of amino acid i with values provided in B.1.

We use a cutoff distance of 3.5 nm for electrostatic interactions and $4\sigma_{ij}$ for the amino-acid-specific AH potential. Both potentials were shifted by the values at the cutoff to ensure continuity (the shifts are not explicitly shown in equation B.13 and B.14). Furthermore, nearest neighbors (i.e. bonded atom pairs) are excluded from the electrostatic and AH potentials.

B.1.3 MRG-CG DNA force field

MRG-CG DNA force field was originally developed by Savelyev et al for simulating Watson-Crick (WC) paired dsDNA with the explicit presence of counter ions [145]. Each nucleotide is represented with one CG bead of mass 325 Da. The energy function excluding ions is defined as

$$U_{\text{DNA}} = U_{\text{bond}} + U_{\text{angle}} + U_{\text{fan-bond}} + U_{\text{contact}} + U_{\text{electrostatics}} \tag{B.16}$$

Bonded potentials are applied to every two neighboring CG atoms on each ssDNA chain, and

$$U_{\text{bond}} = \sum_i \sum_{n=2}^{4} k_{\text{bond},n}(r_{i,i+1} - r_0)^n. \tag{B.17}$$

Values for the spring constants, $k_{\text{bond},n}$, and the equilibrium distance, r_0, are listed in B.3.

Angular potentials are applied to every three neighboring CG atoms on each ssDNA chain, and

$$U_{\text{angle}} = \sum_i \sum_{n=2}^{4} k_{\text{angle},n}(\theta_i - \theta_0)^n. \tag{B.18}$$

θ_i is the i-th angle. Values for the spring constants, $k_{\text{angle},n}$, and the equilibrium distance, θ_0, are listed in B.3.

The fan bonds are introduced to capture base-pairing, cross-stacking, and other interactions between two ssDNA chains. They are applied between CG beads i and $j-5, j-4, \cdots, j+5$, where bead i and j form a WC pair.

$$U_{\text{fan-bond}} = \sum_{\substack{\text{WC pair }(i,j)\\ i<j}} \sum_{\Delta=-5}^{5} \sum_{n=2}^{4} k_{\text{fan bond},n}(r_{i,j+\Delta} - r_{\Delta,0})^n \tag{B.19}$$

Values for the spring constants, $k_{\text{fan bond},n}$, and the equilibrium distance, $r_{\Delta,0}$, are listed in B.5.

To adapt the force field to implicit-ion simulations, we scaled the bonded interactions by a factor of 0.9 so that the simulated persistence length for dsDNA matches the experimental value [55]. Note the parameter values listed in B.3, B.4, and B.5 are the original values reported in reference [145] before scaling. We further replaced the original electrostatic interaction with the same Debye-Hückel electrostatic interaction with distance-dependent dielectric defined in equation B.7. Each CG nucleotide atom possesses a $-e$ charge.

Finally, the nonbonded contact potential for excluded volume effect is defined as

$$U_{\text{contact}} = \sum_{i<j} \frac{\alpha_{\text{DNA-DNA}}}{r_{ij}^{12}}, \tag{B.20}$$

with $\alpha_{\text{DNA-DNA}}$ set as $1.678 \times 10^{-5}\ \text{nm}^{12} \cdot \text{kJ/mol}$.

Nearest neighbor CG beads with bonded and angular potentials are excluded from contact and electrostatic interactions. CG beads with fan bonds are from different DNA strands and are not excluded from such nonbonded interactions.

B.1.4 MOFF Protein-MRG DNA interactions

MOFF Protein-MRG DNA interactions include both contact and electrostatic potentials. The contact potentials are defined as

$$U_{\text{contact}} = \sum_{i<j} \frac{\alpha_{\text{protein-DNA}}}{r_{ij}^{12}}, \tag{B.21}$$

with $\alpha_{\text{protein-DNA}} = 1.6264 \times 10^{-3}\ \text{nm}^{12} \cdot \text{kJ/mol}$. Electrostatic interactions follow the same definition as in equation B.7. A previous study has shown that the above simple treatment of protein-DNA interactions successfully reproduces the binding free energy of several complexes [116].

B.1.5 Mpipi protein and RNA force field

The Mpipi force field is developed based on atomistic simulations and bioinformatic data [58]. Its energy function is defined as

$$U_{\text{Mpipi}} = U_{\text{bond}} + U_{\text{electrostatics}} + U_{\text{WF}} \tag{B.22}$$

The bond term U_{bond} is of the harmonic form

$$U_{\mathrm{bond}} = \sum_i \frac{1}{2} k_{\mathrm{bond}} (r_{i,i+1} - r_0)^2 \tag{B.23}$$

with $k_{\mathrm{bond}} = 1920$ kcal/mol/nm^2. r_0 is 0.381 nm for protein, and 0.5 nm for RNA molecules.

The electrostatic interaction is computed with the Debye-Hückel potential

$$U_{\mathrm{electrostatics}} = \sum_{i<j} \frac{Q_i Q_j}{4\pi\epsilon_0 \epsilon r} \exp(-r_{ij}/\lambda_D) \tag{B.24}$$

with $\epsilon = 80$ for water. The Debye length λ_D depends on the salt concentration of the system. Notably, the model uses scaled charges for amino acids and nucleotides, with $+0.75e$ for Arg and Lys, $-0.75e$ for Asp and Glu, $+0.375e$ for His, and $-0.75e$ for nucleotides.

Nonbonded interactions are captured by the Wang-Frenkel potential [305]

$$\begin{aligned} U_{\mathrm{WF}} =& \sum_{i<j} \epsilon_{ij} \alpha_{ij} \left[\left(\frac{\sigma_{ij}}{r} \right)^{2\mu_{ij}} - 1 \right] \left[\left(\frac{R_{ij}}{r} \right)^{2\mu_{ij}} - 1 \right]^{2\nu_{ij}} \\ \alpha_{ij} =& 2\nu_{ij} \left(\frac{R_{ij}}{\sigma_{ij}} \right)^{2\mu_{ij}} \left\{ \frac{2\nu_{ij} + 1}{2\nu_{ij} \left[\left(\frac{R_{ij}}{\sigma_{ij}} \right)^{2\mu_{ij}} - 1 \right]} \right\}^{2\nu_{ij}+1} \end{aligned} \tag{B.25}$$

with cutoff at $R_{ij} = 3\sigma_{ij}$. All the ϵ_{ij} and σ_{ij} values are listed in B.7 and B.8. μ_{ij} for most pairs is set as 2, except for two Ile as 11, for Val and Ile as 4, and for protein and RNA or RNA and RNA as 3. All the ν_{ij} values are equal to 1.

B.1.6 The generalized structure-based protein model

The structure-based model (SMOG) was originally introduced for modeling single folded proteins. It requires an input configuration file with the protein in the native state to define the stabilizing interaction potential. The SMOG energy function is defined as

$$U_{\text{SMOG}} = U_{\text{bond}} + U_{\text{angle}} + U_{\text{dihedral}} + U_{\text{memory}} + U_{\text{contact}} + U_{\text{electrostatics}} \tag{B.26}$$

The bonded term, U_{bond} is defined using harmonic potentials

$$U_{\text{bond}} = \sum_i \frac{1}{2} k_{\text{bond}} (r_{i,i+1} - r^o_{i,i+1})^2, \tag{B.27}$$

where $r_{i,i+1}$ is the distance between two Cα atoms. $r^o_{i,i+1}$ is the corresponding value measured from the input native structure. By default $k_{\text{bond}} = 50000$ kJ/mol/nm^2.

The angle term, U_{angle}, is defined as

$$U_{\text{angle}} = \sum_i \frac{1}{2} k_{\text{angle}} (\theta_i - \theta^o_i)^2, \tag{B.28}$$

where θ_i is the i-th angle, and θ^o_i is the native angle value measured from the input native structure. By default $k_{\text{angle}} = 100$ kJ/mol/rad^2.

U_{dihedral} is the dihedral potential defined as

$$U_{\text{dihedral}} = \sum_i \sum_{n=1,3} k_{\text{dihedral},n} [1 + \cos(n(\theta_i - \theta^o_i - \pi))] \tag{B.29}$$

where θ_i is the i-th dihedral, and θ^o_i is the dihedral measured from the input native structure. $k_{\text{dihedral},1} = 2.5$ kJ/mol, and $k_{\text{dihedral},3} = 1.25$ kJ/mol.

The memory potential U_{memory} includes Gaussian functions and stabilizes ordered domains. The native pairs are found by the shadow algorithm, and the potential is defined as

$$U_{\text{memory}} = \sum_{\substack{\text{native pairs } (i,j) \\ i<j}} -\epsilon_G \exp\left(-\frac{(r_{ij} - \mu_{ij})^2}{2\sigma_G^2}\right) + \frac{\alpha_G}{r_{ij}^{12}} \left[1 - \exp\left(-\frac{(r_{ij} - \mu_{ij})^2}{2\sigma_G^2}\right)\right], \tag{B.30}$$

where $\epsilon_G = 2.5$ kJ/mol, $\alpha_G = 4.194304 \times 10^{-5}$ kJ·nm^{12}/mol, and $\sigma_G = 0.05$ nm. μ_{ij} is the distance between the i-th and j-th Cα atoms measured from the native structure. The potential has cutoff distances as $\mu_{ij} + 6\sigma_G$. We note that when modeling proteins with disordered regions, native pairs should be limited only to amino acid pairs from the folded domains.

The contact potential is modeled with the Lennard-Jones (LJ) potential as

$$U_{\text{contact}} = \sum_{j>i+3} 4\epsilon_{ij} \left[\left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^{6} \right], \tag{B.31}$$

where $\sigma = 0.5$ nm and the cutoff is 1.25 nm. Values for ϵ_{ij} are defined using a scaled Miyazawa-Jernigan (MJ) statistical potential and provided in B.9. This potential is essential for describing the interactions between protein domains and across protein molecules.

Electrostatic interactions are described with the Debye-Hückel potential

$$U_{\text{electrostatics}} = \sum_{j>i+3} \frac{Q_i Q_j}{4\pi\epsilon_0 \epsilon r} \exp(-r/\lambda_D), \tag{B.32}$$

where Q_i and Q_j are the charges of residues i and j, respectively. ϵ_0 is the vaccum permittivity and $\epsilon = 78.0$ is the relative permittivity of water. λ_D is the Debye length computed as

$$\lambda_D = \sqrt{\frac{k_B T \epsilon_0 \epsilon}{2 N_A c e^2}} \tag{B.33}$$

where the variables are the Boltzmann constant (k_B), temperature (T), the Avogadro constant (N_A), the monovalent salt concentration (c), and the elementary charge (i.e. proton charge) (e). The arginine and lysine Cα atoms have $+1$ e charges, the aspartic acid and glutamic acid Cα atoms possess -1 e charges, and other amino acid Cα atoms have zero charges.

B.1.7 SMOG Protein-3SPN2 DNA interactions

SMOG and 3SPN2 are combined in OpenABC. Such implementations are suitable for studying protein-DNA complexes such as nucleosomes. Here we only elaborate the parameters for B-curved DNA.

The protein-DNA nonbonded contact interactions between Cα and any DNA CG atom also follow equation B.31, with $\epsilon_{ij} = 0.125$ kJ/mol, $\sigma = 0.57$ nm, and cutoff equal to 1.425 nm.

The protein-DNA electrostatic interactions also follow equation B.32. $+e$ charge is assigned to arginine and lysine Cα atoms, $-e$ charge is assigned to aspartic acid and glutamic acid Cα atoms, and phosphate. All the other CG atoms have zero charge.

B.2 Details for Preparing and Performing Simulations

B.2.1 Python implementation of the shadow algorithm for native pair identification

MOFF uses native pairs identified in the input atomistic configuration with the shadow algorithm [299] to stabilize secondary and tertiary structures. We provide a Python implementation of the shadow algorithm to avoid the installation of Perl, which is the programming language used in the original implementation [133].

The shadow algorithm searches contacts between residues based on atomistic configurations. Residues α and β are in contact if any heavy atom i from residue α contacts with any heavy atom j from residue β. To test if atom i contacts with atom j, imagine there are light sources at the center of the two atoms. If neither the light from i to j nor the light from j to i is blocked by a third heavy atom k, then atoms i and j are in contact. We further provide pseudocode below to clarify the implementation of the algorithm.

Note the shadow algorithm requires atomistic models as inputs, while outputs contacts between residues. The algorithm also ignores contacts between residues that are involved in bond, angle, or dihedral potentials in the CG force field (i.e. residues that are in 1-2, 1-3, or 1-4 relations).

Algorithm 1 Pseudocode for shadow algorithm implementation. r_c is the cutoff distance for searching close neighbors. We used the MDTraj package [298] to load and parse the initial input PDB structure, and the $O(N)$ complexity cell list spatial neighbor searching is achieved by MDAnalysis [306], [307]. By default $r_c = 0.6$ nm, $r = 0.1$ nm, and $r_b = 0.05$ nm.

```
Input: $\{\vec{x}_i\}$ $(i = 1, \cdots, N)$      # input atomistic coordinate
Input: $r_c, r, r_b$      # set cutoff $r_c$, atom radius $r$, and bonded atom radius $r_b$
Find all the spatially close heavy atom pairs NeighborHeavyAtomPairs = $\{(i, j)\}_{d_{ij} < r_c}$ and $i < j$      # use cell list to search spatially close pairs with $O(N)$ complexity
create empty list ResiduePairs = []
for $(i, j)$ in NeighborHeavyAtomPairs do
  if not ($i$ and $j$ are from same residue or $i$ and $j$ are from residues with 1-2, 1-3, or 1-4 relation or $d_{ij} < r$ or (Residue($i$), Residue($j$)) in ResiduePairs) then
    # Residue(i) means the residue that atom i is from
    # ensure i and j are from different residues without bonded interactions and $d_{ij} \geq r$
    Flag = True, $r_i = r$, $r_j = r$
    Find all heavy atoms $k$ ($k \neq i$ and $k \neq j$) that $d_{ik} < d_{ij}$ and $d_{jk} < d_{ij}$, the set including all such atoms $k$ is called HeavyAtomNeighbors($i$, $j$)      # find all the blocking candidates

    for $k$ in HeavyAtomNeighbors$(i, j)$ do
      if $k$ is connected to $i$ or $k$ is connected to $j$ then
        $r_k = r_b$
      else
        $r_k = r$
      end if
      # test if any atom k blocks the contact between i and j
      if $r_k > d_{ik}$ or $r_k > d_{jk}$ then
        Flag = False
        Break
      else
        # ensure $r_k \leq d_{ik}$ and $r_k \leq d_{jk}$, which is required by LightIsBlocked
        if LightIsBlocked$(i, j, k, r_j, r_k)$ or LightIsBlocked$(j, i, k, r_i, r_k)$ then
          # see the next algorithm for the definition of function LightIsBlocked
          Flag = False
          Break
        end if
      end if
    end for
    if flag then
      Add (Residue($i$), Residue($j$)) to ResiduePairs
    end if
  end if
end for
Output: ResiduePairs
```

Algorithm 2 Define function LightIsBlocked

```
Define LightIsBlocked($i$, $j$, $k$, $r_j$, $r_k$)    # check if the light from atom $i$ to $j$ is blocked by atom $k$
Require: $r_j \leq d_{ij}$ and $r_k \leq d_{ik}$    # ensure arcsin can be computed properly
$\theta_{jik}$ is the angle of $j$-$i$-$k$
if $\arcsin(r_j/d_{ij}) + \arcsin(r_k/d_{ik}) \geq \theta_{jik}$ then
    return True    # $k$ blocks the light from $i$ to $j$
else
    return False
end if
```

B.2.2 Implementation of the temperature replica exchange algorithm

We implemented the temperature replica exchange [128] simulation protocol in OpenABC. We used “torch.distributed" package to enable communications between replicas. The replica exchange is executed by exchanging coordinates and rescaled velocities while the replica temperatures are fixed. The exchange is accepted with probability determined by the Metropolis-Hastings algorithm. For example, for two replicas with indices 0 and 1, each one has temperature, coordinates, potential energy, and velocities as (T_0, x_0, E_0, v_0) and (T_1, x_1, E_1, v_1), respectively. If an exchange is attempted, the acceptance probability is $\min(1, \exp((\beta_1 - \beta_0)(E_1 - E_0)))$, where $\beta_i = 1/(k_B T_i)$ and k_B is the Boltzmann constant. If the exchange is accepted, the corresponding quantities will become $(T_0, x_1, E_1, v_1\sqrt{T_0/T_1})$ and $(T_1, x_0, E_0, v_0\sqrt{T_1/T_0})$. Rescaling the velocities ensures that the average kinetic energy is consistent with the new temperature. The Debye length is always computed at a temperature of 300 K, even if the replica thermostat has a different temperature, so the Hamiltonian of different replicas is the same.

B.2.3 Building and relaxing atomistic structures from coarse-grained configurations

We used “reconstruct atomic model from reduced representation (REMO)" to reconstruct atomic configurations of protein condensates [152]. Given a coarse-grained configuration with only coordinates for α carbons, REMO first removes steric clashes between these atoms. It then reconstructs atomistic representations by optimizing the hydrogen bond networks with backbone built from a backbone isomer library. We iteratively applied REMO to individual chains of the condensate system for improved computation efficiency (see Algorithm 3).

To perform atomistic simulations starting from the reconstructed structure, we solvated it with explicit water molecules and counter ions. The box size was chosen such that protein atoms are at least 2 nm from the boundary, resulting in a size of 39.887 × 44.420 × 45.281 nm^3. The concentration of monovalent ions was set as 82 mM NaCl.

From the solvated structure, we first carried out an energy minimization using a double-precision version of GROMACS. Subsequently, using mixed-precision GROMACS, we performed a 62.5-ps-long NVT simulation with a timestep of 0.5 fs, followed by a 1-ns-long NVT simulation with a timestep of 1 fs. Finally, we performed the production simulation for more than 20 ns with a timestep of 1.5 fs. The temperature was maintained at 260 K during simulations using the velocity rescaling [308] scheme with a coupling constant of 1 ps. 240 CPU cores were used to perform the atomistic simulation.

Algorithm 3 Reconstruct full atomic configurations for proteins

```
Input: $CGCoords = \{\vec{x}_i\}$ $(i = 1, \cdots, N)$   # input C$\alpha$ coordinates of the system
Input: $n$   # number of protein chains
Input: $NumRes = \{N_j\}(\sum_j^n N_j = N)$   # number of residues in each chain
Create empty atomic coordinates list $AtomicCoords = []$
if $n == 1$ then
    $AtomicCoords$ = REMO($CGCoords$)   # there is only one chain, so reconstruct using REMO model directly
else if $n > 1$ then
    $CGCoordsList = split(CGCoords, NumRes)$   # there are multiple protein chains, so split the C$\alpha$ coordinates of the whole system into $n$ coordinate lists $CGCoordsList = [CGCoords\_1, ..., CGCoords\_n]$.
    for $CGCoords\_i$ in $CGCoordsList$ do
        $AtomicCoords\_i = REMO(CGCoords\_i)$   # reconstruct i-th protein chain using REMO
        $AtomicCoords\_i = align(AtomicCoords\_i, CGCoords\_i)$   # align atomic coordinates of i-th chain with its CG cooridnates
    end for
    $AtomicCoords = combine([AtomicCoords\_1, ..., AtomicCoords\_n])$   # combine atomic coordinates of all the protein chains together
end if
Output: $AtomicCoords$
```

B.2.4 Setting up MOFF HP1 system

Setting up MD simulations with MOFF requires 3D protein native configurations as inputs. We followed the protocols established before [55], [116] to build HP1 dimer structures. We first built the monomeric atomistic structures with RaptorX [129]. Each monomer contains two ordered domains: chromodomain (CD) and chromoshadow domain (CSD). The CD and CSD of HP1α correspond to regions with residues 17-72 and residues 115-176, respectively, while the CD and CSD of HP1β correspond to regions with residues 21-79 and residues 117-175. To build dimer structures, we aligned the CSDs of two monomers to the bound configuration reported in PDB ID 3I3C for HP1α and PDB ID 3Q6S for HP1β.

From the constructed atomistic structures for HP1 dimers, we determined the equilibrium angles, dihedrals, and native pair distances with MDTraj [298] to define the corresponding terms in the force field. We only included native pairs between residues from the same CD, the same CSD, or between two CSDs. For HP1β, the initial dimer structure includes unphysical overlaps between the C-terminal tail and CSD from different monomers. To avoid the impact of these overlaps on native pair detection, we applied the shadow algorithm to a single monomer to determine intra-CD and CSD native pairs. For native pairs at the dimer interface between two CSDs, we applied the shadow algorithm to a structure containing only the two CSDs.

The interaction strength in the native pair potential was set as $\epsilon = 6$ kJ/mol (equation B.5). We used a concentration of 82 mM monovalent salt to compute the Debye length (equation B.9) and to closely mimic the experimental setup [252].

B.2.5 Benchmarking the performance of condensate simulations

We carried out simulations of condensate systems with N_1 HP1α dimers and N_2 200-bp-long dsDNA using both GROMACS and OpenMM. GROMACS version is 2018.4 and compiled with mixed precision. The precision style for OpenMM was set as “mixed", which means single precision for forces while double precision for integration.

For CPU simulations with GROMACS, we used the leap-frog stochastic dynamics integrator (sd integrator) with the coupling constant (tau-t) set as 1 ps. We used the repository that implements the MOFF and MRG-CG DNA force field available at https://github.com/ZhangGroup-MITChemistry/MOFF. For single-GPU simulations with OpenMM, the integrator was set as Langevin middle integrator with a friction coefficient of 1 ps^{-1}. All simulations were performed under temperature 300 K and lasted one million steps with a timestep of 10 fs.

B.2.6 Validating the force field implementation in OpenMM

To validate our implementation MOFF, we ran a 0.1-million-step NVT simulation for HP1α dimer with GROMACS at 300 K to collect ten configurations. We used sd integrator with time coupling constant as 1 ps and a timestep of 10 fs. We then evaluated the potential energies for the ten configurations using both OpenMM and GROMACS with a salt concentration of 82 mM and temperature at 300 K. The interaction strength for protein native pair contacts was set as $\epsilon = 6.0$ kJ/mol. The results are shown in B.10, and the energies computed from the two software agree well. The minor differences come from using tabulated functions for native pair, contact, and electrostatic potentials in GROMACS but analytical expressions in OpenMM. In addition, GROMACS does not shift nonbonded contact potentials to zero at cutoff distances, while OpenMM does.

To validate the MRG implementation and its integration with MOFF, we simulated a protein-DNA complex (HP1α dimer + 200-bp-long dsDNA) to collect ten configurations. An umbrella bias on the center of mass (COM) distance, r_{COM} defined as

$$U_{\mathrm{bias}} = \frac{k_{\mathrm{bias}}}{2} r_{\mathrm{COM}}^2 \tag{B.34}$$

was applied to promote protein-DNA contacts, with $k_{\mathrm{bias}} = 50$ kJ/mol/nm^2. The simulation was performed with GROMACS at 300 K and lasted 2 million steps, with configurations saved at every 0.2 million steps. The timestep is 10 fs and the time coupling constant is 1 ps. The umbrella bias was implemented using PLUMED [309]. We then evaluated the potential energies for the ten configurations using both OpenMM and GROMACS with a salt concentration of 82 mM and temperature at 300 K. The interaction strength for protein native pair contacts was set as $\epsilon = 6.0$ kJ/mol. The results are shown in B.11 and agree well. Again, the use of tabulated functions in GROMACS resulted in minor differences.

To verify our implementation of the HPS force field, we ran a 0.1-million-step simulation for DDX4 with OpenMM to collect ten configurations. We used Langevin middle integrator with friction coefficient as 1/ps and a timestep of 10 fs, with configurations saved at every 10,000 steps. We evaluated the potential energies for the ten configurations using both OpenMM and HOOMD-Blue. We tested both Urry and KR scales with the optimal parameter set ($\mu_{\text{Urry}}^{\text{opt}} = 1$, $\Delta_{\text{Urry}}^{\text{opt}} = 0.08$ and $\mu_{\text{KR}}^{\text{opt}} = 1$, $\Delta_{\text{KR}}^{\text{opt}} = 0$). Since HOOMD-Blue does not shift the potential to ensure continuity at cutoff distances, we did not offset nonbonded potentials in OpenMM for energy comparisons. The energies computed with OpenMM and HOOMD-Blue match exactly as shown in B.12.

To validate our implementation of the Mpipi force field, we ran a 1-million-step-long simulation for a polyR+polyK+polyU system with LAMMPS. The system consists of a chain of 10 arginines, a chain of 10 lysines, and 2 individual chains of 10 uracils. The simulation was performed with Langevin dynamics at 300 K, with a damping coefficient of 10.001 ps and a timestep of 10 fs. The configurations were saved every 0.1 million steps. We evaluated the interaction energies for the 10 configurations using both OpenMM and LAMMPS and the results as shown in B.13.

B.2.7 Computing DNA persistence length with the MRG-CG model

To compute the DNA persistence length, we simulated a 200-bp dsDNA chain with the MRG-CG model at 300 K with a monovalent salt concentration of 100 mM. The dsDNA was placed into a cubic box of length 500 nm. We performed three independent 5-billion-step NVT simulations using the Langevin middle integrator with a friction coefficient of 1/ps and a timestep of 10 fs. The configurations were saved every 0.1 million steps, and the first one billion steps were discarded as equilibration.

Using the simulated configurations, we computed the persistence length as follows. Since DNA is a double-strand helix with a periodicity of about 10 bp, we choose CG particles of index 66, 76, 86, 96, 106, 116, 126, 136 to define a pseudo chain. These particles are within the middle 70 bp of the DNA, thus avoiding boundary effects. By connecting neighboring particles on the pseudo chains, we can define a bond vector $\vec{b}_i$ as the normalized vector of pointing from the $i-1$-th to the i-th particle. The average correlation between two normalized pseudo bond vectors with gap n was computed as

$$C(n) = \left\langle \vec{b}_i \cdot \vec{b}_{i+n} \right\rangle \tag{B.35}$$

where the average was performed over pseudo bond index i and all the configurations. Assuming an exponential decay for $C(n) \approx \exp\left(-n\bar{l}_b/l_p\right)$, we determined the persistence length with numerical fitting of $\log C(n)$ and n as

$$\log C(n) = -\alpha n + \beta. \tag{B.36}$$

$\alpha = \bar{l}_b/l_p$ with $\bar{l}_b$ as the mean bond length along the pseudo chain. The mean bond length $\bar{l}_b$ was 3.37 nm, and the persistence length l_p was 48.83 ± 2.71 nm. The fitting results are shown in Figure B.1.

B.3 Sequences

The following sequences were used in simulations of respective proteins.

B.3.1 HP1α monomer

MGKKTKRTADSSSSEDEEEYVVEKVLDRRVVKGQVEYLLKWKGFSEEHNTWEPEKN
LDCPELISEFMKKYKKMKEGENNKPREKSESNKRKSNFSNSADDIKSKKKREQSND
IARGFERGLEPEKIIGATDSCGDLMFLMKWKDTDEADLVLAKEANVKCPQIVIAFY
EERLTWHAYPEDAENKEKETAKS

B.3.2 HP1β monomer

MGKKQNKKKVEEVLEEEEEEYVVEKVLDRRVVKGKVEYLLKWKGFSDEDNTWEPEE
NLDCPDLIAEFLQSQKTAHETDKSEGGKRKADSDSEDKGEESKPKKKKEESEKPRG
FARGLEPERIIGATDSSGELMFLMKWKNSDEADLVPAKEANVKCPQVVISFYEERL
TWHSYPSEDDDKKDDKN

B.3.3 FUS LC

MASNDYTQQATQSYGAYPTQPGQGYSQQSSQPYGQQSYSGYSQSTDTSGYGQSSYS
SYGQSQNTGYGTQSTPQGYGSTGGYGSSQSSQSSYGQQSSYPGYGQQPAPSSTSGS
YGSSSQSSSYGQPQSGSYSQQPSYGGQQQSYGQQQSYNPPQGYGQQNQYNS

B.3.4 DDX4

MGDEDWEAEINPHMSSYVPIFEKDRYSGENGDNFNRTPASSSEMDDGPSRRDHFMK
SGFASGRNFGNRDAGECNKRDNTSTMGGFGVGKSFGNRGFSNSRFEDGDSSGFWRE
SSNDCEDNPTRNRGFSKRGGYRDGNNSEASGPYRRGGRGSFRGCRGGFGLGSPNND
LDPDECMQRTGGLFGSRRPVLSGTGNGDTSQSRSGSGSERGGYKGLNEEVITGSGK
NSWKSEAEGGES

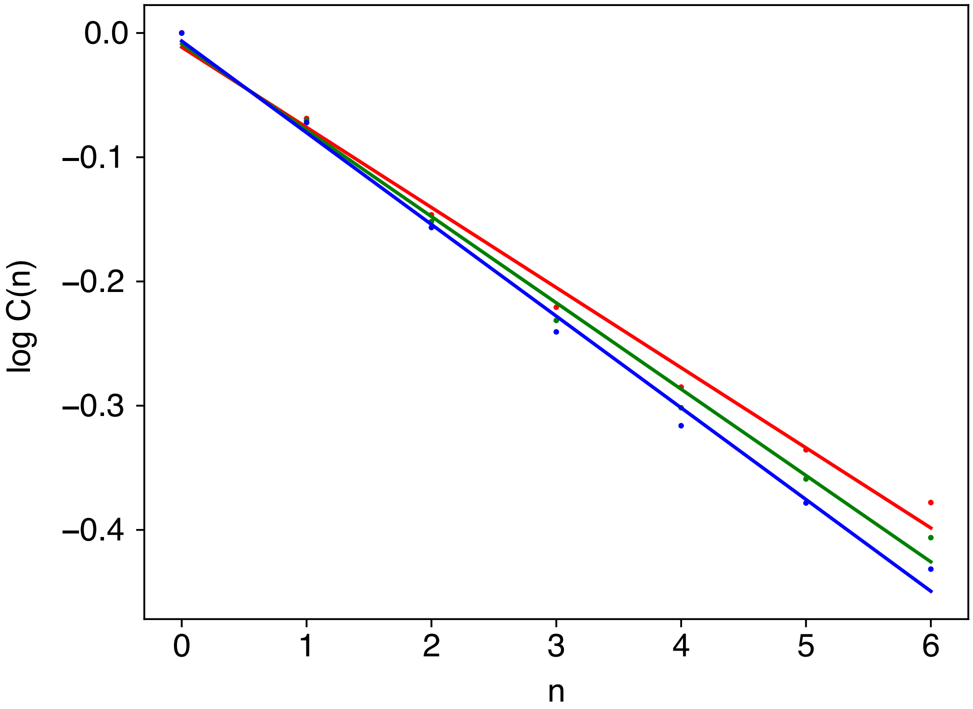


Figure B.1: The log of the bond vector correlation, $\log C(n)$, as a function of the bond separation n. The dots were obtained from MD simulations, with three colors indicate three independent simulations. The lines are numerical fits to the data. See text *Section: Computing DNA persistence length with the MRG-CG model* for simulation details and computing persistence length from the numerical fitting.

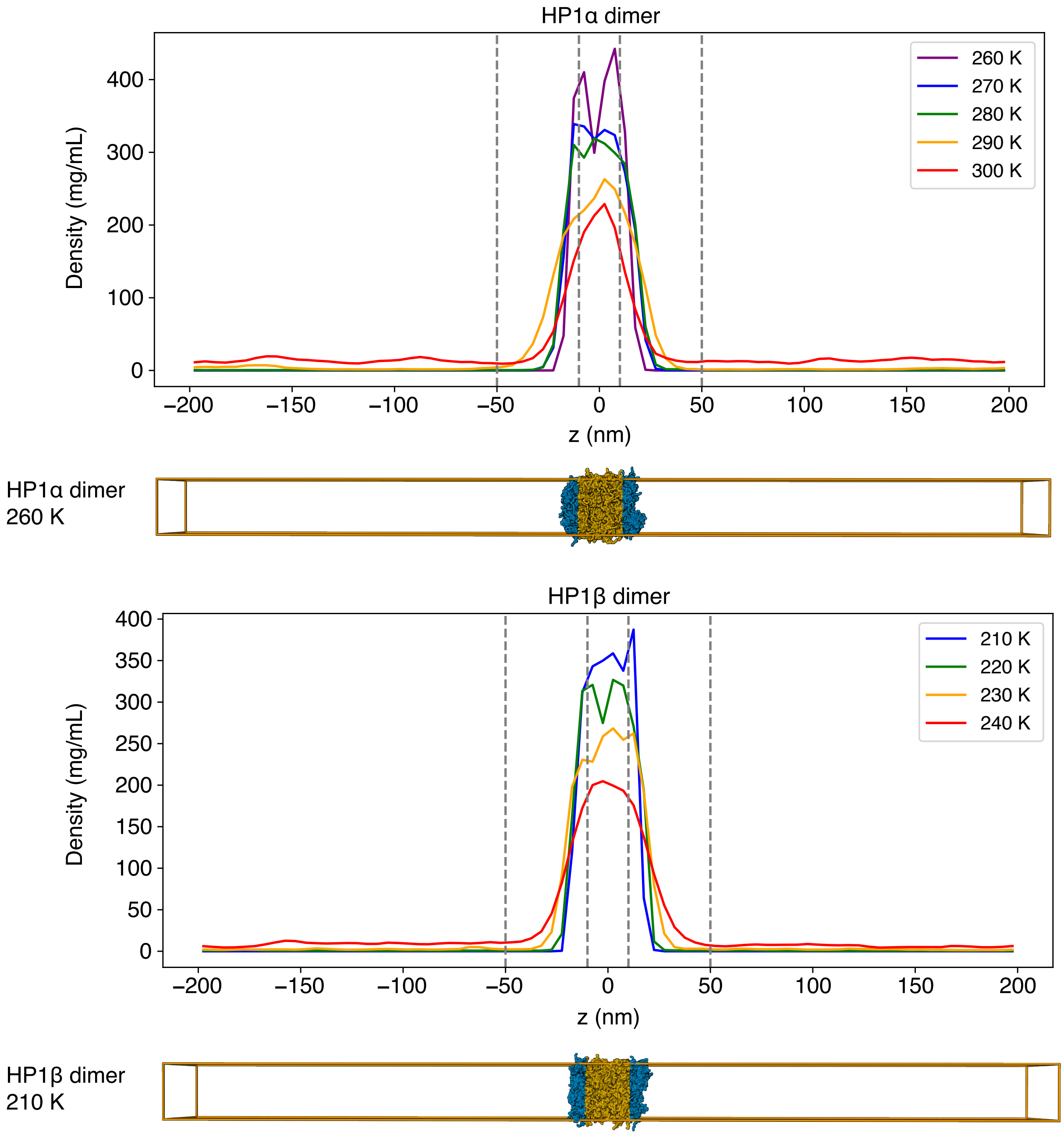


Figure B.2: Density profiles obtained from slab simulations of HP1α (left) and HP1β (right) dimers with the MOFF model. Vertical lines are set at $z = \pm 10$ and ± 50 nm. The final snapshots of the slab simulations at 260 K for HP1α and 210 K for HP1β are shown. CG atoms with $|z| < 10$ nm are colored in yellow, while the remaining are shown in blue.

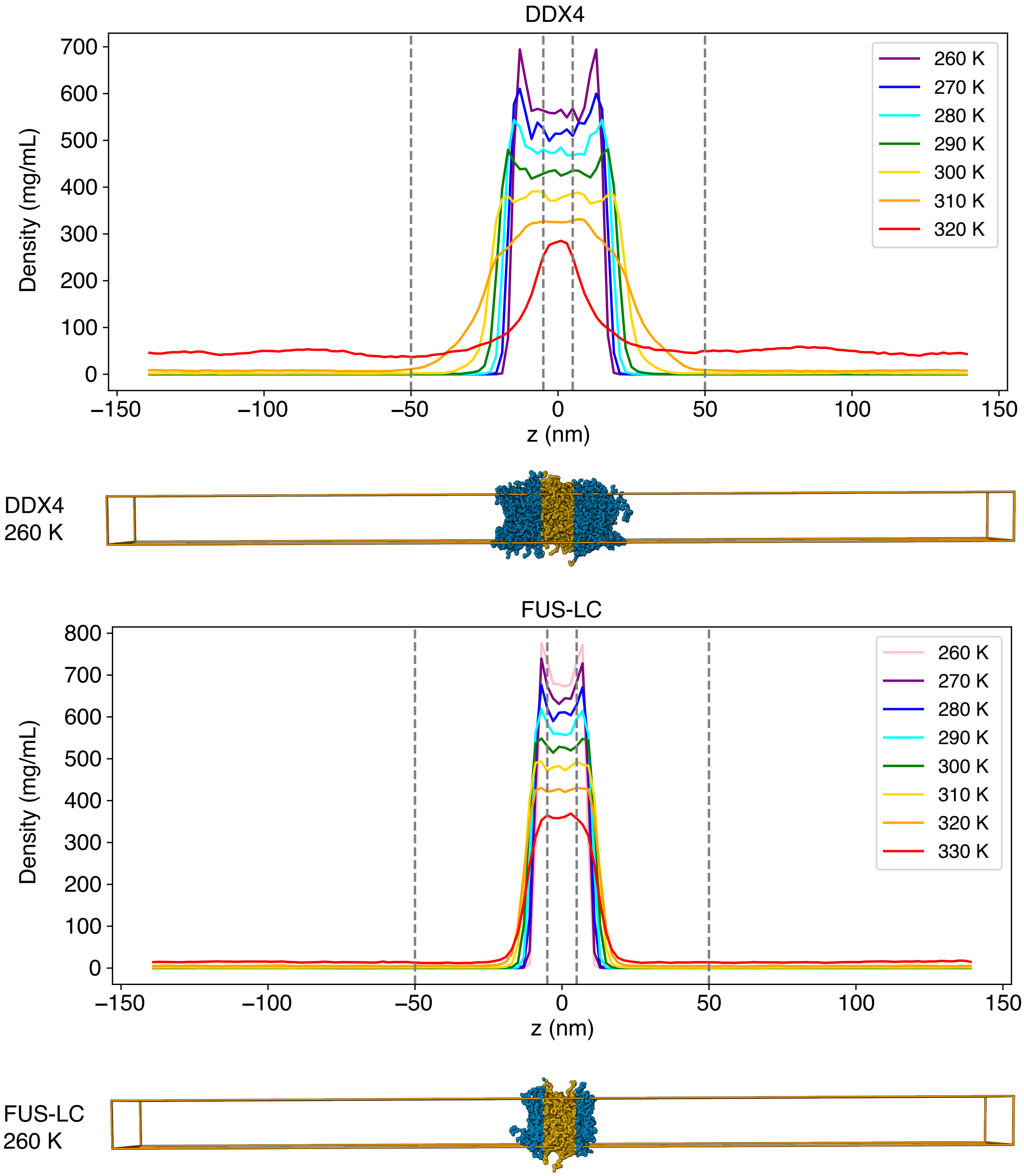


Figure B.3: Density profiles obtained from slab simulations of DDX4 and FUS LC with the HPS model using the Urry scale optimal parameter set ($\mu = \mu_{\text{Urry}}^{\text{opt}} = 1$ and $\Delta = \Delta_{\text{Urry}}^{\text{opt}} = 0.08$) at different temperatures. Vertical dashed lines are set at $z = \pm 5$ nm and ± 50 nm. The final snapshots of the slab simulations at 260 K are shown. CG atoms with $|z| < 5$ nm are colored in yellow, while the remaining are shown in blue. This figure shows that the $|z| < 5$ nm and $|z| > 50$ nm regimes can represent the concentrated and dilute phases, respectively.

Table B.1: The amino acid mass, sizes, and charges used by MOFF and HPS models. Both models share the same amino acid mass and sizes. The charge of HIS differs in the two models, while other amino acids share the same charge. Here e is the elementary charge.

Amino acid	Mass (Da)	Size (nm)	MOFF charge (e)	HPS charge (e)
ALA	71.08	0.504	0	0
ARG	156.20	0.656	1	1
ASN	114.10	0.568	0	0
ASP	115.10	0.558	-1	-1
CYS	103.10	0.548	0	0
GLN	128.10	0.602	0	0
GLU	129.10	0.592	-1	-1
GLY	57.05	0.450	0	0
HIS	137.10	0.608	0.25	0.5
ILE	113.20	0.618	0	0
LEU	113.20	0.618	0	0
LYS	128.20	0.636	1	1
MET	131.20	0.618	0	0
PHE	147.20	0.636	0	0
PRO	97.12	0.556	0	0
SER	87.08	0.518	0	0
THR	101.10	0.562	0	0
TRP	163.20	0.678	0	0
TYR	163.20	0.646	0	0
VAL	99.70	0.586	0	0

Table B.2: MOFF protein contact ϵ_{ij} values as defined in equation B.6. Due to limited space, the numbers are rounded to 3 decimal places. The values are in unit kJ/mol.

	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA	0.904	0.022	1.413	-0.436	1.376	0.025	-0.243	-0.439	-0.023	2.340	-0.111	-0.201	1.209	1.805	-0.178	-0.136	-0.210	3.887	3.467	1.320
ARG	-	5.746	-2.640	-0.566	0.295	1.583	-0.919	2.450	2.005	4.739	1.868	-1.751	-0.551	-0.432	-1.081	-1.697	-0.411	2.862	6.503	3.243
ASN	-	-	-0.429	1.105	4.326	0.400	-1.681	-0.168	-0.507	-0.763	-0.293	1.086	-1.218	-0.199	-2.204	-0.311	-0.156	3.380	-0.918	-0.383
ASP	-	-	-	0.961	1.460	-0.381	1.739	-1.073	0.396	-0.094	-0.398	-0.993	-1.507	-0.514	-0.802	-0.257	2.585	-0.523	-0.911	-0.158
CYS	-	-	-	-	5.406	-1.510	3.079	0.322	2.041	-0.628	3.885	2.836	3.098	0.518	1.456	0.965	5.066	0.512	6.124	2.614
GLN	-	-	-	-	-	-0.720	-1.840	0.111	-0.197	3.443	1.313	2.394	6.734	-0.405	-1.431	-0.628	1.014	-0.926	4.234	1.435
GLU	-	-	-	-	-	-	-1.763	0.041	-2.409	2.840	-0.210	-0.956	4.629	-0.544	-0.765	-0.195	1.636	2.347	1.963	-0.278
GLY	-	-	-	-	-	-	-	1.059	-0.009	0.603	0.338	0.549	-2.221	-0.269	-0.128	-0.832	0.853	2.307	-0.653	0.386
HIS	-	-	-	-	-	-	-	-	-0.068	1.993	-0.251	4.281	-0.595	2.827	2.334	0.438	0.855	-0.349	-0.281	2.910
ILE	-	-	-	-	-	-	-	-	-	3.144	5.030	1.627	3.709	2.266	1.182	2.246	3.081	-0.778	6.530	3.399
LEU	-	-	-	-	-	-	-	-	-	-	-0.100	1.173	5.956	5.702	-0.187	-0.172	-0.163	6.058	-0.659	0.649
LYS	-	-	-	-	-	-	-	-	-	-	-	0.037	-0.366	-0.048	-0.401	0.286	-0.841	8.442	-0.037	0.652
MET	-	-	-	-	-	-	-	-	-	-	-	-	-0.305	1.530	-1.966	-0.350	2.956	4.745	-0.017	5.035
PHE	-	-	-	-	-	-	-	-	-	-	-	-	-	2.684	2.241	1.147	3.540	7.383	6.564	2.303
PRO	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-1.197	-0.524	-0.477	0.185	0.859	-1.460
SER	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.337	1.024	4.761	1.575	0.241
THR	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.764	2.486	2.206	-0.134
TRP	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	5.889	7.143	3.585
TYR	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	5.901	0.489
VAL	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2.327

Table B.3: MRG DNA model bond parameters. All the $k_{\text{bond},n}$ are in unit kcal/mol/nm^2, and r_0 unit is nm.

$k_{\text{bond},2}$	$k_{\text{bond},3}$	$k_{\text{bond},4}$	r_0
262.5	-226	149	0.496

Table B.4: MRG DNA model angle parameters. All the $k_{\text{angle},n}$ are in unit kcal/mol/degree^2, and θ_0 unit is degree.

$k_{\text{angle},2}$	$k_{\text{angle},3}$	$k_{\text{angle},4}$	θ_0
9.22	4.16	1.078	156

Table B.5: MRG DNA model fan bond parameters. All the $k_{\text{fan bond},n}$ are in kcal/mol/nm^2, and $r_{\Delta,0}$ unit is nm. Δ means the fan bond between CG nucleotide i and $j + \Delta$, where nucleotide i and j are WC-paired.

Δ	$k_{\text{bond},2}$	$k_{\text{bond},3}$	$k_{\text{bond},4}$	$r_{\Delta,0}$
-5	4.67	2.1	1.46	1.71
-4	0.0001324	-12.2	18.5	1.635
-3	8.5	-44.4	50.0	1.47
-2	12.3	-40.0	37.0	1.345
-1	4.0	-10.0	8.0	1.23
0	292.0	410.0	720.0	1.13
1	11.5	-41.0	58.0	0.99
2	9.55	-45.9	50.2	0.92
3	13.78	-52.7	50.0	1.02
4	13.86	-56.8	50.0	1.25
5	36.26	-77.0	50.0	1.69

Table B.6: The normalized KR scale and Urry hydropathy scale values (i.e. λ_i parameters in equation B.14).

Amino acid	KR scale	Urry scale
ALA	0.730	0.602942
ARG	0.0	0.558824
ASN	0.432	0.588236
ASP	0.378	0.294119
CYS	0.595	0.64706
GLN	0.514	0.558824
GLU	0.459	0.0
GLY	0.649	0.57353
HIS	0.514	0.764707
ILE	0.973	0.705883
LEU	0.973	0.720589
LYS	0.514	0.382354
MET	0.838	0.676471
PHE	1.0	0.82353
PRO	1.0	0.758824
SER	0.595	0.588236
THR	0.676	0.588236
TRP	0.946	1.0
TYR	0.865	0.897059
VAL	0.892	0.664707

Table B.7: Mpipi parameter ϵ values as defined in equation B.25. Due to limited space, the numbers are rounded to 3 decimal places. The values are in unit kJ/mol.

	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL	A	C	G	U
ALA	0.207	0.738	0.509	0.327	0.249	0.523	0.345	0.305	0.872	0.104	0.127	0.158	0.186	0.923	0.268	0.232	0.168	1.255	0.980	0.115	1.175	0.804	1.183	0.742
ARG	-	0.376	1.040	0.006	0.780	1.054	0.007	0.837	0.520	0.636	0.658	0.510	0.718	2.280	0.799	0.764	0.699	2.894	2.543	0.647	2.518	1.778	2.535	1.652
ASN	-	-	0.811	0.629	0.551	0.825	0.647	0.607	1.174	0.406	0.429	0.460	0.488	1.225	0.570	0.534	0.470	1.557	1.282	0.417	1.477	1.106	1.485	1.044
ASP	-	-	-	0.331	0.368	0.643	0.345	0.425	0.007	0.224	0.246	0.002	0.306	1.043	0.388	0.352	0.288	1.375	1.100	0.235	1.237	0.866	1.245	0.804
CYS	-	-	-	-	0.290	0.564	0.387	0.347	0.914	0.146	0.168	0.199	0.228	0.964	0.310	0.274	0.209	1.296	1.022	0.157	1.216	0.846	1.224	0.783
GLN	-	-	-	-	-	0.839	0.661	0.621	1.188	0.420	0.442	0.473	0.502	1.239	0.584	0.548	0.484	1.571	1.296	0.431	1.490	1.120	1.499	1.057
GLU	-	-	-	-	-	-	0.358	0.444	0.007	0.243	0.265	0.002	0.325	1.061	0.406	0.371	0.306	1.393	1.119	0.253	1.250	0.880	1.259	0.817
GLY	-	-	-	-	-	-	-	0.404	0.970	0.203	0.225	0.256	0.285	1.021	0.366	0.331	0.266	1.353	1.079	0.213	1.273	0.903	1.281	0.840
HIS	-	-	-	-	-	-	-	-	0.114	0.769	0.792	0.366	0.851	1.588	0.933	0.898	0.833	1.920	1.646	0.780	1.128	0.758	1.136	0.695
ILE	-	-	-	-	-	-	-	-	-	0.002	0.024	0.055	0.084	0.820	0.165	0.130	0.065	1.152	0.878	0.012	1.072	0.702	1.080	0.639
LEU	-	-	-	-	-	-	-	-	-	-	0.046	0.077	0.106	0.842	0.188	0.152	0.087	1.174	0.900	0.035	1.094	0.724	1.102	0.661
LYS	-	-	-	-	-	-	-	-	-	-	-	0.080	0.137	0.482	0.219	0.183	0.118	0.438	0.430	0.066	0.667	0.444	0.672	0.407
MET	-	-	-	-	-	-	-	-	-	-	-	-	0.166	0.902	0.247	0.212	0.147	1.234	0.960	0.094	1.154	0.784	1.162	0.721
PHE	-	-	-	-	-	-	-	-	-	-	-	-	-	1.639	0.984	0.948	0.884	1.971	1.696	0.831	1.890	1.520	1.899	1.457
PRO	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.329	0.293	0.229	1.316	1.042	0.176	1.236	0.865	1.244	0.803
SER	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.258	0.193	1.280	1.006	0.141	1.200	0.830	1.208	0.767
THR	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.129	1.216	0.941	0.076	1.135	0.765	1.144	0.702
TRP	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2.302	2.028	1.163	2.222	1.852	2.231	1.789
TYR	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1.754	0.889	1.948	1.578	1.956	1.515
VAL	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.023	1.083	0.712	1.091	0.650
A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2.142	1.772	2.151	1.709
C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1.402	1.780	1.339
G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2.159	1.718
U	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1.276

Table B.8: Mpipi parameter σ values as defined in equation B.25. Due to limited space, the numbers are rounded to 3 decimal places. The values are in unit nm.

	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL	A	C	G	U
ALA	0.527	0.605	0.559	0.555	0.550	0.577	0.572	0.498	0.580	0.589	0.588	0.597	0.587	0.594	0.554	0.534	0.558	0.616	0.600	0.571	0.686	0.675	0.689	0.672
ARG	-	0.684	0.638	0.633	0.628	0.656	0.651	0.577	0.659	0.664	0.664	0.674	0.664	0.673	0.632	0.612	0.635	0.695	0.679	0.648	0.764	0.753	0.767	0.750
ASN	-	-	0.592	0.587	0.582	0.610	0.605	0.531	0.613	0.619	0.619	0.628	0.619	0.628	0.586	0.567	0.590	0.649	0.633	0.602	0.718	0.707	0.722	0.705
ASP	-	-	-	0.582	0.577	0.605	0.600	0.526	0.608	0.615	0.615	0.624	0.614	0.622	0.581	0.562	0.585	0.644	0.628	0.598	0.713	0.702	0.717	0.700
CYS	-	-	-	-	0.572	0.600	0.595	0.521	0.603	0.611	0.610	0.619	0.609	0.617	0.577	0.557	0.580	0.639	0.622	0.594	0.708	0.697	0.712	0.695
GLN	-	-	-	-	-	0.628	0.623	0.549	0.631	0.637	0.637	0.646	0.637	0.645	0.604	0.584	0.607	0.667	0.651	0.620	0.736	0.725	0.739	0.722
GLU	-	-	-	-	-	-	0.618	0.544	0.626	0.632	0.632	0.641	0.632	0.640	0.599	0.579	0.603	0.662	0.645	0.616	0.731	0.720	0.734	0.717
GLY	-	-	-	-	-	-	-	0.470	0.552	0.558	0.558	0.567	0.558	0.566	0.525	0.505	0.528	0.588	0.571	0.541	0.657	0.646	0.660	0.643
HIS	-	-	-	-	-	-	-	-	0.634	0.639	0.639	0.649	0.639	0.648	0.607	0.587	0.610	0.670	0.654	0.623	0.739	0.728	0.742	0.725
ILE	-	-	-	-	-	-	-	-	-	0.692	0.661	0.662	0.650	0.654	0.614	0.595	0.621	0.676	0.659	0.672	0.768	0.757	0.772	0.755
LEU	-	-	-	-	-	-	-	-	-	-	0.653	0.659	0.648	0.654	0.614	0.595	0.620	0.676	0.659	0.639	0.749	0.738	0.752	0.735
LYS	-	-	-	-	-	-	-	-	-	-	-	0.667	0.657	0.663	0.623	0.604	0.628	0.685	0.668	0.643	0.756	0.745	0.759	0.742
MET	-	-	-	-	-	-	-	-	-	-	-	-	0.647	0.654	0.613	0.594	0.618	0.676	0.659	0.632	0.745	0.734	0.749	0.732
PHE	-	-	-	-	-	-	-	-	-	-	-	-	-	0.663	0.621	0.602	0.625	0.685	0.668	0.637	0.753	0.742	0.757	0.740
PRO	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.581	0.561	0.584	0.643	0.627	0.597	0.712	0.701	0.716	0.699
SER	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.541	0.565	0.623	0.607	0.578	0.693	0.682	0.696	0.679
THR	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.589	0.646	0.630	0.604	0.716	0.705	0.720	0.703
TRP	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.707	0.690	0.659	0.775	0.764	0.779	0.762
TYR	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.673	0.642	0.759	0.748	0.762	0.745
VAL	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.627	0.735	0.724	0.739	0.722
A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.844	0.833	0.848	0.831
C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.822	0.837	0.820
G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.851	0.834
U	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.817

Table B.9: SMOG MJ potential parameter ϵ as defined in equation B.31. Due to the limited space, the numbers are rounded to 3 decimal places. The values are in unit kJ/mol.

	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA	0.436	0.293	0.295	0.273	0.572	0.303	0.242	0.370	0.386	0.734	0.787	0.210	0.632	0.771	0.326	0.322	0.372	0.613	0.539	0.648
ARG	-	0.249	0.263	0.367	0.412	0.289	0.364	0.276	0.346	0.582	0.646	0.095	0.500	0.638	0.273	0.260	0.305	0.547	0.507	0.492
ASN	-	-	0.269	0.269	0.415	0.274	0.242	0.279	0.334	0.520	0.600	0.194	0.473	0.601	0.245	0.253	0.301	0.492	0.443	0.454
ASP	-	-	-	0.194	0.386	0.234	0.164	0.255	0.372	0.508	0.545	0.269	0.412	0.558	0.213	0.261	0.289	0.455	0.443	0.398
CYS	-	-	-	-	0.872	0.457	0.364	0.507	0.577	0.882	0.935	0.313	0.800	0.930	0.492	0.459	0.499	0.794	0.667	0.795
GLN	-	-	-	-	-	0.247	0.228	0.266	0.318	0.589	0.648	0.207	0.529	0.657	0.277	0.239	0.305	0.499	0.476	0.492
GLU	-	-	-	-	-	-	0.146	0.196	0.345	0.524	0.576	0.289	0.463	0.571	0.202	0.237	0.279	0.479	0.447	0.428
GLY	-	-	-	-	-	-	-	0.359	0.345	0.606	0.667	0.184	0.544	0.662	0.300	0.292	0.334	0.548	0.483	0.542
HIS	-	-	-	-	-	-	-	-	0.489	0.664	0.728	0.216	0.638	0.765	0.361	0.338	0.388	0.638	0.564	0.574
ILE	-	-	-	-	-	-	-	-	-	1.049	1.129	0.483	0.965	1.097	0.603	0.564	0.646	0.927	0.842	0.970
LEU	-	-	-	-	-	-	-	-	-	-	1.182	0.540	1.028	1.167	0.674	0.629	0.696	0.985	0.909	1.039
LYS	-	-	-	-	-	-	-	-	-	-	-	0.019	0.398	0.539	0.156	0.168	0.210	0.431	0.417	0.399
MET	-	-	-	-	-	-	-	-	-	-	-	-	0.876	1.052	0.553	0.486	0.563	0.890	0.787	0.853
PHE	-	-	-	-	-	-	-	-	-	-	-	-	-	1.164	0.682	0.645	0.686	0.988	0.908	1.009
PRO	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.281	0.252	0.305	0.598	0.512	0.532
SER	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.268	0.314	0.479	0.446	0.489
THR	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.340	0.516	0.483	0.555
TRP	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.811	0.747	0.831
TYR	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.669	0.741
VAL	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.885

Table B.10: Comparison of potential energies computed with OpenMM and GROMACS using MOFF for ten configurations of HP1α dimer. The energy unit is kJ/mol. See text *Section:Validating the force field implementation in OpenMM* for simulation details.

Frame ID	Software	Bonds	Angles	Dihedrals	Native pairs	Contacts	Electrostatics
1	OpenMM	4.97	2.55	1.67	-3346.36	-103.74	-53.22
	GROMACS	4.97	2.55	1.67	-3346.36	-103.74	-53.24
2	OpenMM	386.61	375.45	454.22	-2936.22	-87.27	-65.96
	GROMACS	386.61	375.44	454.22	-2936.22	-87.26	-65.95
3	OpenMM	444.53	384.39	400.96	-2925.66	-109.99	-62.71
	GROMACS	444.53	384.39	400.96	-2925.66	-109.98	-62.70
4	OpenMM	387.01	384.20	422.79	-2877.72	-96.31	-53.14
	GROMACS	387.01	384.20	422.79	-2877.72	-96.30	-53.13
5	OpenMM	440.40	383.35	450.34	-2934.92	-88.74	-59.70
	GROMACS	440.40	383.35	450.34	-2934.92	-88.73	-59.70
6	OpenMM	428.49	352.92	445.23	-2913.69	-90.77	-78.06
	GROMACS	428.49	352.92	445.23	-2913.69	-90.77	-78.06
7	OpenMM	473.75	439.53	480.74	-2915.47	-128.57	-54.99
	GROMACS	473.75	439.53	480.74	-2915.47	-128.56	-54.98
8	OpenMM	433.29	375.09	427.08	-2915.08	-127.77	-65.20
	GROMACS	433.29	375.09	427.08	-2915.08	-127.77	-65.19
9	OpenMM	423.03	382.15	455.44	-2982.15	-99.48	-84.06
	GROMACS	423.03	382.15	455.44	-2982.15	-99.48	-84.05
10	OpenMM	452.80	417.20	494.68	-2896.97	-126.01	-85.76
	GROMACS	452.80	417.20	494.68	-2896.97	-126.00	-85.76

Table B.11: Comparison of potential energies computed with OpenMM and GROMACS using MOFF for proteins and MRG for DNA for ten configurations of HP1α dimer bound to a dsDNA. The energy unit is kJ/mol. See text *Section: Validating the force field implementation in OpenMM* for simulation details.

		Protein				DNA			
Frame ID	Software	Bonds	Angles	Dihedrals	Native pairs	Bonds & fan bonds	Angles	Contacts	Electrostatics
1	OpenMM	5.08	2.70	2.33	-3346.29	98.90	3.21	-106.02	759.21
	GROMACS	5.08	2.70	2.33	-3346.29	98.90	3.21	-106.02	759.04
2	OpenMM	420.00	344.70	459.40	-2940.25	1264.66	411.22	-140.32	686.32
	GROMACS	420.00	344.70	459.40	-2940.25	1264.65	411.22	-140.31	686.27
3	OpenMM	404.52	387.80	480.73	-2913.34	1341.51	438.40	-161.86	673.97
	GROMACS	404.52	387.80	480.73	-2913.34	1341.51	438.40	-161.85	673.91
4	OpenMM	418.64	416.13	472.84	-2878.63	1352.10	416.04	-137.46	633.35
	GROMACS	418.64	416.13	472.84	-2878.63	1352.10	416.04	-137.45	633.32
5	OpenMM	437.58	376.80	452.27	-2869.10	1387.26	413.92	-119.47	665.30
	GROMACS	437.58	376.80	452.27	-2869.10	1387.26	413.92	-119.47	665.26
6	OpenMM	425.41	416.51	502.90	-2874.56	1375.09	390.28	-108.54	600.42
	GROMACS	425.41	416.51	502.90	-2874.56	1375.09	390.28	-108.53	600.37
7	OpenMM	415.22	385.06	466.07	-2922.99	1292.21	390.08	-144.63	641.36
	GROMACS	415.22	385.06	466.07	-2922.99	1292.21	390.08	-144.62	641.30
8	OpenMM	427.79	401.11	488.72	-2922.98	1284.19	393.42	-160.31	595.48
	GROMACS	427.79	401.11	488.72	-2922.98	1284.19	393.42	-160.30	595.43
9	OpenMM	423.46	397.75	537.37	-2917.23	1376.37	381.50	-142.74	570.76
	GROMACS	423.46	397.75	537.37	-2917.23	1376.37	381.50	-142.73	570.71
10	OpenMM	391.07	332.69	509.94	-2903.83	1303.68	410.59	-162.62	567.65
	GROMACS	391.07	332.69	509.94	-2903.83	1303.68	410.59	-162.61	567.60

Table B.12: Comparison of potential energies computed with OpenMM and HOOMD-Blue using HPS with Urry or KR scales for ten configurations of protein DDX4. The energy unit is kJ/mol. See text *Section: Validating the force field implementation in OpenMM* for simulation details.

Frame ID	Software	Bonds	Contacts (Urry)	Contacts (KR)	Electrostatics
1	OpenMM	399.36	-62.28	-76.41	12.69
	HOOMD-Blue	399.36	-62.28	-76.41	12.69
2	OpenMM	333.40	-67.89	-83.48	13.89
	HOOMD-Blue	333.40	-67.89	-83.48	13.89
3	OpenMM	314.07	-66.08	-82.68	12.95
	HOOMD-Blue	314.07	-66.08	-82.68	12.95
4	OpenMM	341.34	-68.28	-86.45	13.69
	HOOMD-Blue	341.34	-68.28	-86.45	13.69
5	OpenMM	346.92	-68.09	-85.46	14.10
	HOOMD-Blue	346.92	-68.09	-85.46	14.10
6	OpenMM	342.02	-74.57	-94.67	13.08
	HOOMD-Blue	342.02	-74.56	-94.67	13.08
7	OpenMM	252.22	-73.01	-93.11	15.66
	HOOMD-Blue	252.22	-73.01	-93.11	15.66
8	OpenMM	296.98	-63.08	-84.78	11.16
	HOOMD-Blue	296.98	-63.08	-84.78	11.16
9	OpenMM	338.76	-68.30	-88.34	12.56
	HOOMD-Blue	338.76	-68.30	-88.34	12.56
10	OpenMM	386.13	-72.54	-94.43	13.99
	HOOMD-Blue	386.13	-72.54	-94.43	13.99

Table B.13: Comparison of potential energies computed with OpenMM and LAMMPS using Mpipi force field for a polyR+polyK+polyU system. The system consists of a chain of 10 arginines, a chain of 10 lysines, and 2 individual chains of 10 uracils. The energy unit is kJ/mol. See text *Section: Validating the force field implementation in OpenMM* for simulation details.

Frame ID	Software	Bonds	Contacts	Electrostatics
1	OpenMM	0.07	-12.49	19.22
	LAMMPS	0.07	-12.49	19.22
2	OpenMM	35.10	-22.14	12.26
	LAMMPS	35.10	-22.14	12.26
3	OpenMM	32.01	-37.80	12.77
	LAMMPS	32.01	-37.80	12.77
4	OpenMM	24.04	-62.50	6.19
	LAMMPS	24.04	-62.49	6.19
5	OpenMM	51.21	-29.63	14.43
	LAMMPS	51.20	-29.63	14.43
6	OpenMM	52.29	-33.85	11.85
	LAMMPS	52.29	-33.85	11.85
7	OpenMM	56.81	-37.85	2.63
	LAMMPS	56.81	-37.85	2.63
8	OpenMM	50.78	-28.42	11.88
	LAMMPS	50.78	-28.42	11.88
9	OpenMM	69.27	-72.28	2.20
	LAMMPS	69.27	-72.28	2.20
10	OpenMM	55.98	-36.91	9.33
	LAMMPS	55.98	-36.91	9.33

Table B.14: The coexistence concentrations of HP1α and HP1β dimers measured by slab simulations with MOFF at different temperatures below the critical temperature. The cutoff distance for searching the largest cluster is 5 nm.

Protein	T (K)	ρ_{L} (mg/mL)	ρ_{H} (mg/mL)
HP1α dimer	260	0.02	387.37
	270	0.16	327.11
	280	0.51	305.77
	290	2.51	242.44
	300	13.45	207.17
HP1β dimer	210	0.00	347.36
	220	0.82	310.57
	230	2.31	252.44
	240	7.64	199.42

Table B.15: The coexistence concentrations of HP1α and HP1β dimers measured by slab simulation with MOFF. The concentrations were similarly determined as those shown in Table B.14 but the cutoff distance for searching the largest cluster set as 8 instead of 5 nm. The results are almost identical to the ones shown in Table B.14, supporting the robustness of phase diagrams with respect to the cutoff distance used for protein clustering.

Protein	T (K)	ρ_{L} (mg/mL)	ρ_{H} (mg/mL)
HP1α dimer	260	0.02	387.37
	270	0.16	327.11
	280	0.51	305.77
	290	2.51	242.44
	300	13.45	207.17
HP1β dimer	210	0.00	347.36
	220	0.82	310.58
	230	2.31	252.54
	240	7.63	199.00

Table B.16: The coexistence concentrations of FUS LC and DDX4 proteins measured by performing slab simulations with HPS model Urry scale and the optimal parameter set ($\mu = \mu_{\text{Urry}}^{\text{opt}} = 1$ and $\Delta = \Delta_{\text{Urry}}^{\text{opt}} = 0.08$) at different temperatures below the critical temperature. The cutoff distance for searching the largest cluster is 5 nm.

Protein	T (K)	ρ_{L} (mg/mL)	ρ_{H} (mg/mL)
FUS LC	260	0.00	695.41
	270	0.00	653.23
	280	0.01	610.82
	290	0.28	569.50
	300	0.28	525.02
	310	1.25	480.18
	320	4.96	425.01
	330	14.57	361.59
DDX4	260	0.00	560.53
	270	0.00	514.19
	280	0.10	474.87
	290	0.57	431.43
	300	1.89	379.44
	310	7.69	326.18
	320	47.67	271.05

Appendix C

Supplementary information for chapter 4

C.1 Details of Coarse-grained Simulations

We carried out all molecular dynamics simulations with the software LAMMPS [208]. Umbrella sampling was performed using collective variables implemented by the Plumed package [310]. We applied the weighted histogram method [133], [311] and fastmbar [76] to process the data and computed the unbiased extension length at a given force.

C.1.1 System Setup

We built a structural model for the chromatin with 12 nucleosomes and 20-bp linker DNA following two steps. We first connected 12 individual nucleosomes into a continuous segment without much regard to the overall chromatin topology. We then aligned the DNA model to a template that closely resembles the cryo-EM structure with a two-start fibril organization [9].

We connected individual nucleosomes to build a 12mer chromatin as follows. The nucleosome unit with 167-bp of DNA was extracted from the tetranucleosome X-ray structure (PDB ID: 1ZBB) [8]. The DNA was taken as residues 158-324 of chain I and the corresponding complementary segment from chain J. There are no extra base pairs at the entry side of the nucleosome in this setup, but 20-bp linker DNA exists at the exiting end. The resulting sequence with the core DNA (147bp) in the underline is

ACAGGATGTAACCTGCAGATACTACCAAAAGTGTATTTGGAAACTGCTCCATCAAAAGGCATGTTCAGCTGGATTCCAGCTGAACATGCCTTTTGATGGAGCAGTTTCCAAATACACTTTTGGTAGTATCTGCAGGTGATTCTCCAGGGCGGCCAGTACTTACATGC

We further replaced the coordinates for histone proteins with that from PDB ID: 1KX5 [4], which resolved the coordinates for histone tails.

We added one additional DNA base pair at the end of the linker DNA as one sticky end using the software 3DNA [312] for alignment between neighboring nucleosomes. This 168-bp segment is the building block for constructing the dodecamer. For example, to extend chromatin with n nucleosomes, we align the 168-th bp of the n-th nucleosome with the first bp of the $(n + 1)$-th nucleosome. The alignment determines the orientation of the $(n + 1)$-th nucleosome, and the fiber is extended by one nucleosome after removing the overlapping nucleotides. For the last (12-th) nucleosome, we deleted the linker DNA to build the dodecamer with 1984 bp of DNA. The resulting all-atom model was converted into the coarse-grained model with in-house scripts.

While the above procedure succeeds at building an all-atom model for the 12mer, the precise topology of the resulting structure cannot be controlled easily. To construct a two-start fibril configuration resembling the compact and twisted Cryo-EM structure [9], we aligned the model to a two-start fiber structure built by the software FiberModel, as detailed below. The structural alignment was performed using MDAnalysis [306], [307] with RMSD coordinate fitting [313], [314].

The template was generated by the software fiberModel as the lowest energy configuration [201]. FiberModel optimized fiber configurations by utilizing a series of geometric parameters, including the height per nucleosome along the fiber axis (h), the rotation angle per nucleosome around the fiber axis (θ), the radius of the fiber (R), and three Euler angles that determine the direction of each nucleosome (α, β, γ). To build a fibril chromatin structure, we set initial values for the parameters $(h, \theta, R, \alpha, \beta, \gamma)$ as (2.34 nm, 2.88, 7.3 nm, -3.14, 0.622, 0) as estimated from the cryo-EM structure [9]. Also, each nucleosome was treated as a cylinder with a radius and height of 5.2 and 4.5 nm. We then used FiberModel to optimize the chromatin structure based on parameters α and γ, keeping the other parameters fixed. The optimization utilized the basin hopping global search technique [315]. The final structure aligned with the FiberModel template is shown in Figure 2.1A.

C.1.2 Force Field Setup

We used the same force fields as in the tetra-nucleosome study [46] to simulate the 12mer. The 3SPN.2C DNA model [206] was adopted to model each nucleotide with three coarse-grained beads for phosphate, sugar, and base, respectively. The Cα structure-based model [204] was adopted to simulate the conformational dynamics of individual histone proteins. Both bonded and nonbonded interactions were generated based on the nucleosome crystal structure (PDB ID: 1KX5). For nonbonded contact potentials, two residues were considered in contact when their minimum distance is smaller than 6Å, implemented using the Shadow algorithm [299]. We further scaled the default interaction strength [133] by 2.5 to prevent proteins from unfolding at 300K. To model the disordered portions of the histones, we removed the dihedral and contact potentials for disordered residues not included in the core histones (residue ID: 44-135, 160-237, 258-352, 401-487, 531-622, 647-724, 745-839, 888-974). Including the secondary structure motifs in the disordered regions of histone proteins does not quantitatively change nucleosome stability and protein-DNA interactions (Figure C.1). The IDs continuously index residues from chain A to chain H of the crystal structure with PDB ID: 1KX5.

In addition, residue-specific protein-protein interactions were introduced with the Miyazwa-Jernigan (MJ) potential [205] and scaled by a factor of 0.4. In a previous study, we showed that the scaled MJ potential provides a balanced modeling of the radius of gyration for both folded and disordered proteins [46].

Protein-DNA interactions include the electrostatic potential modeled at the Debye-Hückel level with a monovalent salt concentration of 150 mM. In addition, a weak, non-specific Lennard-Jones potential was applied between all protein-DNA beads. Detailed expression for these potentials can be found in Ref. [211].

Our group and the de Pablo group have shown that the force field can reproduce the energetic cost of nucleosomal DNA unwinding [148], [150], [211], the dependence of the unwinding barrier on applied tension [148], and the sequence-specific DNA binding strength to the histone octamer [316]. The de Pablo group further showed that the model could reproduce the binding strength between a pair of nucleosomes measured in DNA origami-based force spectrometer experiments [207], [317].

The quantitative accuracy of the coarse-grained model in reproducing single nucleosome stability and inter-nucleosome interactions strongly supports its application to longer chromatin segments. Our prior study of tetranucleosomes further supports the model's accuracy in studying connected nucleosomes. For example, we simulated two di-nucleosomes with different link lengths and succeeded in resolving the structural difference, quantitatively reproducing FRET measurements from the van Noort group [318]. More details about this comparison can be found in Figure S10 of Ref. [46]. For the tetra-nucleosome, we predicted that stacked conformations with the two columns of nucleosomes more aligned have lower free energy than the PDB structure. This prediction was validated by independent simulations performed with an explicit solvent force field SIRAH [319] and by all-atom simulations from the Wereszczynski group [320]. More details about this comparison can be found in Figure S5 of Ref. [46].

C.1.3 Free Energy Profiles for Chromatin Under Tension

We defined two collective variables to explore chromatin configurations and compute free energy profiles. The unwrapping variable, q_{wrap}, quantifies DNA unwrapping using the distance between neighboring nucleosome $d_{i,i+1}$. It is defined as

$$q_{\text{wrap}} = \frac{1}{11}\sum_{i=1}^{11} \exp\left[-\frac{(\max(d_{i,i+1}, d_o) - d_o)^2}{2\sigma_w^2}\right]. \tag{C.1}$$

$d_o = 15$ nm is close to the distance between two neighboring nucleosomes in the PDB structure (PDB ID: 1KX5) [4], and we used $\sigma_w = 4$ nm. The function max selects the larger value of the two distances. The above definition makes use of the geometric constraint that increase in the distances between neighboring nucleosomes $(d_{i,i+1})$ can only arise from nucleosome unwrapping. The unstacking variable, d_{stack}, measures the mean distance between i-th and $(i+2)$-th nucleosomes as

$$d_{\text{stack}} = \frac{1}{10}\sum_{i=1}^{10} d_{i,i+2}. \tag{C.2}$$

Umbrella simulations with the two collective variables at forces 0-3 pN were carried out to compute the free energy profiles. To compare our simulations with force-extension experiments, we applied force f_{ext} along the z-axis projection of the DNA end-to-end distance (L_z). The two DNA ends were defined as the geometric centers of all the coarse-grained beads for the first and last five base pairs. The potential energy function of these simulations at center (q_o, d_o) and force f_{ext} is defined as

$$U_{\text{biased}} = U(\mathbf{r}) + \frac{\kappa_q}{2}(q_{\text{wrap}} - q_o)^2 + \frac{\kappa_d}{2}(d_{\text{stack}} - d_o)^2 - f_{\text{ext}} L_z, \tag{C.3}$$

where $U(\mathbf{r})$ corresponds to the interaction energy defined by the force field. The umbrella centers (q_o, d_o) were initially placed on a uniform grid [0.45:0.90:0.15]×[10:30:5] nm. We introduced additional centers to improve the overlap between umbrella simulations. A complete list of the umbrella centers and the restraining constants is provided in Table C.1.

At the extension force larger than 3 pN, the 12mer can adopt configurations that cover a wide range of q_{wrap} and d_{stack}. Uniform sampling of the entire accessible phase space becomes too costly computationally. Therefore, we only carried out one-dimensional umbrella simulations with d_{stack} as the collective variable. The corresponding potential energy function was defined as

$$U_{\text{biased}} = U(\mathbf{r}) + \frac{\kappa_d}{2}(d_{\text{stack}} - d_o)^2 - f_{\text{ext}} L_z. \tag{C.4}$$

We used $\kappa_d = 0.05$ kcal/(mol $\cdot$ nm^2), and d_o spans from 10 to 50 nm with a step size of 2.5 nm.

Initial configurations from the neural network model

Conformational sampling of the coarse-grained model is challenging due to strong but nonspecific electrostatic interactions. We initialized the umbrella sampling simulations with the most probable configurations predicted by a neural network under a similar setup to alleviate the sampling challenge. As detailed in the *Section: Neural Network Model for the 12mer Chromatin*, the neural network model quantifies the stability of chromatin configurations using inter-nucleosome distances. It is computationally efficient and allows equilibrium sampling of chromatin configurations.

For a coarse-grained umbrella simulation centered at (q_o^N, d_o^N) with extension force $f_{\text{ext}} \leq$ 3 pN, we carried out replica exchange Monte Carlo sampling of the following biased free energy

$$F_{\text{biased}} = F_{12}(\mathbf{d}) + \frac{\kappa_q}{2}(q_{\text{wrap}} - q_o^N)^2 + \frac{\kappa_d}{2}(d_{\text{stack}} - d_o^N)^2 - f_{\text{ext}} L, \tag{C.5}$$

with $\kappa_q = 47.8$ kcal/mol and $\kappa_d = 4.78 \times 10^{-2}$ kcal/(mol $\cdot$ nm^2). $F_{12}(\mathbf{d})$ quantifies the free energy of the 12-mer as function of inter-nucleosome distances, $\mathbf{d}$, with a neural network model. More details about the free energy function can be found in the *Section: Neural Network Model for the 12mer Chromatin.* L is the distance between the first and the last nucleosomes. See *Section: Numerical Simulations of the Neural Network Model* for sampling details. We used the samples collected in the final 300000 steps of the 300 K replica for a

K-means clustering analysis with 10 centers. The configuration closest to the center of the largest cluster was selected as the most probable configuration.

For simulations with 3.5 pN force, we only performed umbrella sampling of the neural network model at a limited set of values for $d_o^N = 10.0$ nm, 15.0 nm, 20.0 nm, 25.0 nm, and 30.0 nm. q_o^N was set as 0.45. A total of five neural network configurations were constructed. We assigned these configurations to initialize coarse-grained simulations by minimizing the difference between d_o^N and the corresponding umbrella center of coarse-grained simulations.

For simulations with 4 pN force, we performed two sets of coarse-grained simulations. These simulations were initialized with neural network configurations obtained from umbrella sampling with centers located at [q_o^N=0.45, 0.6] × [d_o^N=10:30:5] nm.

The neural network model represents chromatin structures with inter-nucleosome distances. We performed short targeted molecular dynamics simulations starting from the two-helix fiber to build coarse-grained model structures consistent with the most probable configurations from the neural network sampling. These simulations bias on all the inter-nucleosome distances with a restraining constant of 23.9 kcal/(mol $\cdot$ nm^2) for approximately 300000 steps. The end configurations of these simulations were used to initialize the coarse-grained umbrella simulations.

One-dimensional free energy calculations at 4.5 pN force

For simulations with 4.5 pN force, the neural network model is no longer sufficient for producing equilibrated, most probable starting configurations. It was trained using tetra-nucleosome configurations with a maximum extension of 50 nm, so the model can at most predict an end-to-end distance of 183 nm. This value is smaller than that anticipated in the linear regime ($\sim$ 270 nm). We performed two independent sets of umbrella simulations using different initial configurations as detailed below.

In the first set of simulations, we initialized the trajectories using the end configurations from the second set of 4 pN simulations presented in the main text (Table C.1). A total of

17 simulations were performed.

From the first set of umbrella simulations, we observed that chromatin at large end-to-end distances tends to fall into configurations with clusters formed by neighboring nucleosomes. To sample more extended configurations, we introduced another collective variable, $d_{i,i+1}^{\min}$, defined as $\min(d_{i,i+1})$, to initialize the second set of umbrella simulations from the most probable configuration predicted from the neuronal network model at 4 pN. The new variable quantifies the minimal distance between nearest-neighbor nucleosomes. Explicit biases on $d_{i,i+1}^{\min}$ help overcome the energetic barrier associated with breaking these clusters. Specifically, we ran nine umbrella-sampling simulations using harmonic biases on $d_{i,i+1}^{\min}$, with umbrella centers placed on a uniform grid of [5.0:25:2.5] nm and an umbrella bias of 0.0005 kcal/(mol $\cdot$ nm^2). Each simulation lasted for 12.75 million steps and was performed with the presence of a 4.5 pN force on the DNA end-to-end distance. The end structures of these simulations were used to initialize production simulations at the 4.5 pN force with harmonic biases on d_{stack} as all simulations presented in the main text. The production runs lasted 25 million steps.

We combined the two data sets to estimate the chromatin extension at 4.5 pN.

Estimating the extension per nucleosome from experimental data

We processed the force-extension curve from single-molecule force spectroscopy experiments to compute the extension per nucleosome. The extension length from experiments includes contributions from the DNA handle and the chromatin. Following previous study [190], we estimate the DNA handle extension as

$$L_{z,\text{handle}} = L_{\text{c,handle}} \times \left(1 - \frac{1}{2}\sqrt{\frac{k_B T}{f_{\text{ext}} A}} + \frac{f_{\text{ext}}}{S}\right) \tag{C.6}$$

where k_B is Boltzmann constant, T is temperature, A is the persistence length of DNA, f_{ext} is the extension force along the z-axis, and S is the stretching modulus. We used $A = 50$ nm,

$S = 900$ pN, and $T = 300$ K. The contour length of the DNA handle, $L_{c,\text{handle}}$, is estimated as

$$L_{c,\text{handle}} = [n_{\text{bp}} - \text{NRL} \times (n_{\text{nucl}} - 1) - 147]b \tag{C.7}$$

where n_{bp} is the total number of base pairs in DNA, NRL is nucleosomal repeat length, n_{nucl} is the total number of nucleosomes, and b is the length of each base pair. We used NRL $=$ 167 bp, $n_{\text{nucl}} = 25$, $n_{\text{bp}} = 7045$ bp, and $b = 0.34$ nm.

Subtracting the extension of the DNA handle from the total extension length L_z, the extension per nucleosome can be estimated as

$$L_{z,\text{nucl}} = \frac{L_z - L_{z,\text{handle}}}{n_{\text{nucl}} - 1}. \tag{C.8}$$

Theoretical predictions of chromatin extension along the *z*-axis

To better understand the linear extension of chromatin at small forces, we introduced an analytical model based on simulation results without extension force.

We approximate the unbiased free energy profile for chromatin extension at zero force with a harmonic function, $F(L) = a(L{-}L_0)^2{+}b$, where L is the extension length, i.e., the end-to-end distance. The parameters were obtained by a least-squares fitting to the simulation data presented in Figure 2.1C, resulting in $a = 1.200 \times 10^{-2}\ k_BT/\text{nm}^2$, $L_0 = 26.83$ nm, and $b = 0.3820\ k_BT$. The corresponding free energy profile with an extension force f along the z-axis can be defined as $F_f(L) = F(L) - fL\cos\theta$, where θ is the azimuthal angle (i.e. the angle between the fiber end-to-end distance direction and the z-axis). From this expression, the average extension along the z-axis can be computed as

$$\langle L_z \rangle_f = \frac{\int_0^\infty \int_0^\pi L\cos\theta e^{-\beta F_f(L)} L^2 dL \sin\theta d\theta}{\int_0^\infty \int_0^\pi e^{-\beta F_f(L)} L^2 dL \sin\theta d\theta} \tag{C.9}$$

Numerical integration of the above equation led to $\langle L_z \rangle_f = 0$, 33.87, 45.76, and 56.36 nm for extension force of 0, 1, 2, and 3 pN, respectively. The extension per nucleosome along the z-axis (Z_{ext} per nucleosome) is defined as $\langle L_z \rangle_f / 11$ and shown in Figure C.7.

C.1.4 Decomposing Inter-nucleosome Distances into Shear and Normal Motions

As discussed in the main text, two distinct motions can increase the distance between i-th and $(i+2)$-th nucleosomes and the collective variable d_{stack}. To characterize these two motions quantitatively, we introduced a coordinate system for each nucleosome. Following de Pablo and coworkers [207], we defined the origin of the coordinate system using the geometric center of residues 63-120, 165-217, 263-324, 398-462, 550-607, 652-704, 750-811, and 885-949. The IDs continuously index residues from chain A to chain H of PDB 1KX5. Two additional points were introduced to define the nucleosomal plane using the geometric center of the dyad that includes CG atoms 81-131, 568-618, and the geometric center of CG atoms 63-120, 165-217, 750-811, and 885-949. Two unit vectors, $\mathbf{u}$ and $\mathbf{v}$, can then be defined using the vectors pointing from the origin to the dyad and the third point. Atoms in the third point were chosen such that $\mathbf{u}$ and $\mathbf{v}$ are approximately orthogonal to each other. The unit normal vector $\mathbf{w}$ for nucleosome plane can then be defined as parallel to the cross product, $\mathbf{u} \times \mathbf{v}$. An illustration of the various axes is provided in Figure C.11.

With the nucleosomal axes defined above, the distances between two nucleosomes can be decomposed to the distances within the nucleosomal plane, i.e., shearing, and the distance perpendicular to the plane, i.e., unstacking. Denoting the vector from nucleosome i to nucleosome $i+2$ as $\mathbf{d}_{i,i+2}$ (here we use the distance between the coordinate origins for the two nucleosomes), the corresponding normal and shear distances are $d^n_{i,i+2} = |\mathbf{d}_{i,i+2} \cdot \mathbf{w}_i|$ and $d^s_{i,i+2} = \sqrt{|\mathbf{d}_{i,i+2}|^2 - {d^n_{i,i+2}}^2}$. The normal and shear distances for the 12mer chromatin, d_n and d_s, are defined using the mean values of all nucleosome i and $i+2$ pairs as $d_n = \frac{1}{10}\sum_{i=1}^{10} d^n_{i,i+2}$ and $d_s = \frac{1}{10}\sum_{i=1}^{10} d^s_{i,i+2}$.

C.1.5 Free Energy Calculations for Two Interacting 12mers

To quantify the impact of chromatin-chromatin interactions and crowding on the stability of fibril configurations, we carried out simulations with two chromatin segments. Umbrella sampling was performed using two collective variables. The first variable quantifies the average extension of the two 12mers with $\bar{d}_{\text{stack}}$ defined as

$$\bar{d}_{\text{stack}} = \frac{1}{2}(d^1_{\text{stack}} + d^2_{\text{stack}}). \tag{C.10}$$

d_{stack} is defined in Eq. C.2 and $1, 2$ index the two 12mers. The second variable measures the number of contacts between the two chromatins. Contacts were defined at the nucleosome level, and a pair of nucleosomes is denoted as in-contact if the distance between their geometric centers $(d_{i,j})$ is less than 15 nm. Mathematically, the interchain contacts, C, is defined as

$$C = \sum_{i=1}^{12} \sum_{j=13}^{24} \frac{1-(d_{i,j}/d_o)^6}{1-(d_{i,j}/d_o)^{12}}, \tag{C.11}$$

where i and j indices over nucleosomes from the two 12mers and $d_o = 15$ nm.

We biased the simulations towards various collective variable values for a comprehensive exploration of the phase space. Details of the umbrella centers and force restraints used in our simulations are provided in Table C.2. Simulations with umbrella centers $\bar{d}_{\text{stack}}$ biased to values $\leq$ 10nm were initialized using the two-helix fibril configuration for each chromatin placed at $\geq$ 20 nm apart. The rest of the simulations were initialized with extended chromatin configurations extracted from the neural network simulations. Chromatin configurations in simulations with 4 pN extension force at umbrella centers d_o were adopted here for simulations that biased $\bar{d}_{\text{stack}}$ to the same values. Only chromatin configurations for the first set of simulations with 4 pN extension force were used here (see *Section: One-dimensional free energy calculations at 4 pN*). The initial five million steps of each trajectory were discarded as equilibration in our free energy calculations.

C.2 Neural Network Model for the 12mer Chromatin

To facilitate conformational sampling of the 12mer, we introduced a neural network model to quantify the free energy of chromatin configurations as a function of inter-nucleosomal distances. The neural network model is a generalization of the free energy surface for a tetra-nucleosome determined in a previous study [46].

C.2.1 Parameterizing the Tetra-nucleosome Free Energy Landscape with Neural Networks

In Ref. [46], we parameterized a neural network model to compute the free energy of a tetra-nucleosome (A) from the six internuclesome distances ($\mathbf{d}$).

To make the neural network's output invariant with respect to nucleosome indexing order, i.e., $A(\mathbf{d} = (d_{12}, d_{13}, d_{14}, d_{23}, d_{24}, d_{34})) = A(\tilde{\mathbf{d}} = (d_{34}, d_{24}, d_{14}, d_{23}, d_{13}, d_{12}))$, we further converted the inter-nucleosome distances into symmetrical features $\mathbf{s}(\mathbf{d}) = (s_1(\mathbf{d}),\ s_2(\mathbf{d}),\ s_3(\mathbf{d}),\ s_4(\mathbf{d}),\ s_5(\mathbf{d}),\ s_6(\mathbf{d}))$ as follows:

$$
\begin{aligned}
s_1 &= d_{12} + d_{34} \\
s_2 &= d_{13} + d_{24} \\
s_3 &= d_{14} \\
s_4 &= d_{23} \\
s_5 &= d_{12} \cdot d_{13} + d_{24} \cdot d_{34} \\
s_6 &= d_{12} \cdot d_{13}^2 + d_{24}^2 \cdot d_{34}.
\end{aligned}
\tag{C.12}
$$

From the above definition, it is straightforward to verify that $\mathbf{s}(\mathbf{d}) = \mathbf{s}(\tilde{\mathbf{d}})$. In addition, given any $\mathbf{s}$ in the range of $\mathbf{s}(\mathbf{d})$, two solutions of $\mathbf{d}$ exist for Eq. (C.12) and these two solutions corresponds to the two different ways of indexing nucleosome. Specifically, if one of the solution is $\mathbf{d} = (d_{12}, d_{13}, d_{14}, d_{23}, d_{24}, d_{34})$, the other solution will be $\tilde{\mathbf{d}} = (d_{34}, d_{24}, d_{14}, d_{23}, d_{13}, d_{12})$. Therefore, the features $\mathbf{s}(\mathbf{d})$ are symmetric and only symmetric to the two ways of indexing nucleosomes.

Using the symmetric features as input, i.e., $A(\mathbf{d}) = A(\mathbf{s}(\mathbf{d}))$, a neural network with two fully connected hidden layers, each of which has 200 nodes, was used to parameterize the free energy. The neural network was trained by minimizing the loss function

$$||(-\nabla A(\mathbf{d})) - \mathbf{F}(\mathbf{d})||^2 + \lambda||\mathbf{w}||^2, \tag{C.13}$$

where $\mathbf{w}$ are weight parameters of the neural network. $\mathbf{F}(\mathbf{d})$ are mean forces at $\mathbf{d}$ estimated using restrained molecular dynamics simulations (see below). $\lambda = 6 \times 10^{-4}$ is the weight decay factor and acts as a regularizer of optimization. Overall, the neural network has 41801 ($7 \times 200 + 201 \times 200 + 201 \times 1 = 41801$) parameters, which is smaller than the total number of constraints $10000 \times 6 = 60000$. The Adam optimizer [321] was used to train the neural network for 100000 steps with a learning rate of 0.001. To prevent over fitting and improve the robustness of neural networks, we trained 30 models independently and used the average results to estimate the final free energy.

To estimate mean forces at different chromatin configurations with inter-nucleosome distances $\mathbf{d_o}$, we carried out restrained molecular dynamics simulations with the harmonic biasing potential

$$V_b = \frac{1}{2}\sum_{i=1}^{3}\sum_{j=i+1}^{4} k(d_{ij}(\mathbf{r}) - d_{ij}^o)^2, \tag{C.14}$$

where i and j are indexes of the four nucleosomes, $k = 1000$ kJ/(mol·nm^2), and d_{ij}^o is the inter-nucleosome distances between nucleosome i and j for the selected center $\mathbf{d_o}$. The mean forces were estimated as

$$\mathbf{F}_{ij}^o = \frac{1}{T}\sum_{t=1}^{T} k(d_{ij}^t - d_{ij}^o). \tag{C.15}$$

Here $T = 50,000$ represents the number of configurations collected from a 500,000 step-long trajectory.

Ten thousand tetra-nucleosome configurations were selected to compute mean forces and parameterize the neural network. To ensure that these configurations cover relevant structures for chromatin folding, we selected them from simulation trajectories that repeatedly probe chromatin folding and unfolding. These trajectories were performed by combining metadynamics with temperature accelerated molecular dynamics (TAMD) to bias the simulations along two collective variables R_g and Q. The radius of gyration, R_g, is defined as

$$R_g = \sqrt{\frac{1}{4}\sum_{i=1}^{4}(\mathbf{r}_i - \mathbf{r}_{\text{com}})^2}, \tag{C.16}$$

where $\mathbf{r}_i$ is the geometric center of the i-th nucleosome using the coordinates of nucleosome core histone residues. $\mathbf{r}_{\text{com}}$ is the center of mass coordinate for all nucleosomes. Q measures the similarity of a given tetra-nuclesome configuration to the crystal structure (PDB ID: 1ZBB) and is defined as

$$Q = \frac{1}{6}\sum_{i=1}^{3}\sum_{j=i+1}^{4}\exp\left[-\frac{(r_{ij} - r_{ij}^{o})^2}{2\sigma^2}\right], \tag{C.17}$$

where r_{ij} measures the distance between the center of the two nucleosomes. More simulation details can be found in Ref. [46].

C.2.2 Generalizing Tetra-nucleosome Results to 12mer Chromatin

We generalized the tetra-nucleosome neural network model to estimate the free energy of 12mer. We defined $F_n(1\ldots,n)$ as the free energy of an oligomer including n nucleosomes of indices $1,\ldots,n$. We assumed that each nucleosome with index i could only interact with nucleosomes $i \pm 1/2/3$, ignoring nucleosome pair interactions beyond tetramers. As shown in Figure C.4, mean distances for these nucleosome pairs are much larger and no interactions are expected among them. Under this assumption, the free energy of $n+1$ nucleosomes (F_{n+1}) can be determined from the following recursive relationship as

$$F_{n+1}(1,\ldots,n+1) = F_n(1,\ldots,n) + F_4(n-2,n-1,n,n+1) - F_3(n-2,n-1,n). \quad \text{(C.18)}$$

Subtracting the free energy (F_3) avoids the double-counting from adding the tetrameric contribution (F_4).

The trimer free energy is estimated as follows. Assuming that the fourth nucleosome is far away from the rest of the three and its interaction with them can be ignored, the free energy difference between the two should be a constant. Therefore, we have

$$F_3(d_{1,2}, d_{1,3}, d_{2,3}) = F_4(d_{1,2}, d_{1,3}, d_{2,3}, d_{1,4} = d_{2,4} = d_{3,4} = 15\text{nm}) + \text{const.} \quad \text{(C.19)}$$

Here $d_{i,j}$ refers the distance between nucleosome i and j. The distances from the fourth nucleosome to the other three (i.e. $d_{1,4}, d_{2,4}, d_{3,4}$) were set as 15 nm. For $d_{3,4}$, this value is comparable to the distance between neighboring nucleosomes in the PDB structure for a tetra-nucleosome to avoid significant DNA unwrapping or DNA overstretching. It is also large enough to unstack i and $i \pm 2$ nucleosomes and to dissociate i and $i \pm 3$ nucleosome contacts, based on previous computational results [207]. The effectiveness of this generalized neural network model is verified based on the fact that it can accurately predict the extension at different forces (Figure C.5).

Given all the $d_{i,i\pm1/2/3}$ and assuming a left-handed helix, the relative position of each nucleosome can be uniquely determined geometrically, as long as the distances satisfy some geometric requirements such as triangle inequality. After determining the relative position of each nucleosome, the full set of distances (i.e. distances between any two different nucleosomes) was used to bias coarse-grained simulations towards the most probable configurations predicted by the neural network sampling.

C.2.3 Numerical Simulations of the Neural Network Model

We used the replica-exchange Monte Carlo algorithm to explore the free energy surface defined by the neural network. 20 Replicas with temperatures as the geometric sequence from 300 K to 2000 K were used. 500000 steps of simulations were performed for each replica. The initial 20000 steps were used to optimize the MC simulation step size so that the mean acceptance rate of MC movement is $\sim$ 0.20-0.25. The exchange between two neighboring replicas was attempted every 50 steps. We used the samples collected in the final 300000 steps of the replica at 300 K for analysis.

C.3 Details of Validation Simulations

C.3.1 Simulations Starting from Uniformly Extended Chromatin Configurations

As detailed in the *Section: Free Energy Profiles for Chromatin Under Tension*, we used chromatin configurations predicted by the neural network model to initialize the umbrella simulations. From the initial configurations, we performed extensive, long molecular dynamics simulations to alleviate any biases that the neural network model might have introduced.

We acknowledge that, despite our best effort, it remains possible that the simulations are not sufficient to remove biases that the neural network might introduce. As an additional test, we carried out a new set of umbrella simulations starting from uniformly extended chromatin structures. By design, the initial configurations are free of clutches. These simulations were again performed with the presence of a 4 pN extension force for direct comparison with results presented in the main text.

To facilitate the conformational sampling of clutched versus uniform chromatin conformations, we performed two dimensional umbrella simulations using both d_{stack} and α. As mentioned in the main text, d_{stack} uses the average distance between 1-3 nucleosomes to measure the average chromatin extension. α is defined as the ratio of the maximum and minimum distance between 1-3 nucleosomes, i.e., $\alpha = d^{\max}_{i,i+2}/d^{\min}_{i,i+2}$. For clutched configurations, the distance at the interface between two nucleosome clusters is expected to be much larger than the distance between nucleosomes within the same cluster, and α will be much larger than one. On the other hand, for more uniformly extended configurations, α will approach one. The maximal and minimal values of $d_{i,i+2}$ were computed with the following expressions with analytical derivatives

$$\begin{aligned} d^{\max}_{i,i+2} =& \beta_1 \ln \left(\sum_{i=1}^{10} e^{d_{i,i+2}/\beta_1} \right) \\ d^{\min}_{i,i+2} =& \beta_2 / \ln \left(\sum_{i=1}^{10} e^{\beta_2/d_{i,i+2}} \right), \end{aligned} \tag{C.20}$$

where $\beta_1 = 0.1$ nm and $\beta_2 = 1000$ nm. The list of umbrella centers is provided in Table C.3.

We initialized these umbrella simulations with two uniform chromatin configurations that lack nucleosome clutches (Figure C.13A). Both configurations were obtained from biased simulations initialized with the fibril structures. The less extended uniform structure with an end-to-end distance per nucleosome of 7.74 nm was produced by restricting pair-wise nucleosome distances $d_{i,i+1}, d_{i,i+2}, d_{i,i+3}$ to 15, 20, 25 nm, respectively. The more extended uniform chromatin structure with an end-to-end distance per nucleosome of 13.64 nm was prepared with a constant-velocity pulling simulation that stretches the end-to-end distance to 150 nm. The entire 147 bp nucleosomal DNA and the histone core proteins were rigidified during the biasing simulations to prevent DNA unwrapping and clutch formation.

From the two initial configurations, we first carried out 0.5 million steps equilibration simulations to relax them towards individual umbrella centers. Simulations with $d_{\text{stack}} \leq 20$ nm started from the structure with a end-to-end distance per nucleosome of 7.74 nm (Figure C.13A, top), and the rest of the simulations started from the second structure. The relaxation was achieved with a moving harmonic restraint $U_{\mathrm{relax}}(t)$ defined as

$$U_{\mathrm{relax}}(t) = \frac{1}{2}\left[\kappa_{d_{\mathrm{stack}}}(t)(d_{\mathrm{stack}} - d_{\mathrm{stack},0}(t))^2 + \kappa_\alpha(t)(\alpha - \alpha_0(t))^2\right], \tag{C.21}$$

where $\kappa_{d_{\mathrm{stack}}}(t)$ and $\kappa_\alpha(t)$ are time-dependent harmonic restraint constants. $d_{\mathrm{stack},0}(t)$ and $\alpha_0(t)$ are time-dependent moving restraint centers. Values for these time-dependent quantities are provided in Table C.4.

During the relaxation period, for umbrella centers with $d_{\mathrm{stack}} \leq 25$ nm and $\alpha \leq 4$, we kept the entire 147 bp nucleosomal DNA and the histone core rigidified to avoid DNA unwrapping and clutch formation. For umbrella simulations at larger d_{stack} values, no such restrictions were applied since doing so may prevent chromatin extension.

After equilibration, we launched production simulations that lasted 10 million steps. The production simulations used the same force field setup as those presented in the main text. The first three million steps were discarded and the rest of the data were used for free energy calculations.

C.3.2 Simulations with Fully Rigidified Nucleosomes

To explore the role of DNA unwrapping on nucleosome clutch formation and inter-chain contacts, we performed additional simulations with fully rigidified nucleosomes. Unlike simulations presented in the main text, the entire 147 bp nucleosomal DNA and the histone core were constrained together as rigid bodies in the native configurations using the same algorithms. Only linker DNA and histone tails remain flexible.

Chromatin extension under 4 pN force

To more directly evaluate the impact of DNA unwrapping on clutch formation, we carried out a new set of simulations with fully rigidified nucleosomes to study chromatin extension under 4 pN force. A total of 10 umbrella simulations were performed to bias d_{stack} to values between

10 nm to 32.5 nm, with an increment of 2.5 nm. We set the umbrella restraining constant $\kappa = 100.0$ kcal/(mol·nm^2) in the first 400,000 steps to drive chromatin configurations towards the restraining centers. After that, the umbrella bias was relaxed to $\kappa = 0.05$ kcal/(mol·nm^2) and the simulations continued for another 15 millions steps. The initial 1 million steps were excluded when calculating the free energy profile. Umbrella centers and force restraints used in these simulations are provided in Table C.5.

Two sets of configurations were used to initialize the above simulations. They were produced by constant-velocity pulling simulations over 5 million steps initialized from a fibril structure. The pulling bias was applied to the z-axis projection of the end-to-end distance. Fiver pulling simulations were performed using independent random seeds with a target bias of 75 nm, and another five with a target bias of 150 nm. In total, these pulling simulations produced ten configurations. The first five 75 nm configurations were used to initialize umbrella simulations centered between 10 nm to 20 nm. The second five 150 nm configurations were used to initialize umbrella simulations centered between 22.5 nm to 32.5 nm.

Inter-chain contacts with two 12mer simulations

To explore the contribution of nucleosomal DNA unwrapping to inter-chain contacts, we carried out additional simulations following the same protocol as that described in *Section: Free Energy Calculations for Two Interacting 12mers*, but with fully rigidified nucleosomes. Initial configurations of these simulations were obtained from a constant-velocity pulling simulation that drives the chromatin z-axis extension towards 75 nm over 20 million steps. The two 12mers adopt identical configurations at the beginning of the simulations and were separated 20 nm part as measured by the center-of-mass distance. Umbrella centers and harmonic restraining constants used in these simulations are provided in Table C.5. The initial 5 million steps of the umbrella sampling were discarded as equilibration.

Table C.1: Summary of umbrella simulation details for free energy calculations at various extension forces. The format for umbrella centers, “start:end:step”, indicates the a series of values from “start” to “end” with a spacing of “step”. The two restraining constants are shown in the format “($\kappa_{q_{\text{wrap}}}$ (kcal/mol), $\kappa_{d_{\text{stack}}}$ (kcal/(mol $\cdot$ nm^2)))”.

Extension force (pN)	Umbrella center: q_{wrap}	Umbrella center: d_{stack} (nm)	Restraining constants	Simulation length (million steps)
0	0.45:0.90:0.15	10.0:30.0:5.0	(50, 0.05)	10.5
0	1.00	6.0:10.0:0.5	(47.8, 1.20)	10
0	1.00	10.0:15.0:2.5	(47.8, 0.120)	10
0	1.00	12.5:15.0:2.5	(47.8, 0.478)	10
0	0.75:0.95:0.05	6.0:10.0:0.5	(120, 1.20)	10
0	0.90:0.95:0.05	10.0:15.0:2.5	(120, 0.120)	10
0	0.90:0.95:0.05	12.5:15.0:2.5	(120, 0.478)	10
0	0.80:0.85:0.05	10.0:20.0:2.5	(120, 0.0120)	10
0	0.75:0.85:0.05	12.5:20.0:2.5	(120, 0.478)	10
1	0.45:0.90:0.15	10.0:30.0:5.0	(50, 0.05)	10
2	0.45:0.90:0.15	10.0:30.0:5.0	(50, 0.05)	10
3	0.45:0.90:0.15	10.0:30.0:5.0	(50, 0.05)	15
3	0.45	10.0:20.0:5.0	(50, 0.2)	15
3	0.60	10.0:20.0:5.0	(50, 0.2)	15
3	0.75	10.0:30.0:5.0	(50, 0.2)	15
3	0.90	10.0:30.0:5.0	(50, 0.2)	15
3.5 (1st set)	n.a.	10.0:50.0:2.5	(0, 0.05)	25
4 (1st set)	n.a.	10.0:50.0:2.5	(0, 0.05)	24.5
4 (2nd set)	n.a.	10.0:50.0:2.5	(0, 0.05)	25
4.5 (1st set)	n.a.	10.0:50.0:2.5	(0, 0.05)	25
4.5 (2nd set)	n.a.	10.0:50.0:2.5	(0, 0.05)	25
4.5 (3rd set)	n.a.	47.5:50.0:2.5	(0, 0.5)	25

Table C.2: Summary of umbrella simulation details for free energy calculations with two 12-mers. The same format as in Table A.1 is adopted here. The two restraining constants are shown in the format “(κ_C (kcal/mol), $\kappa_{\bar{d}}$ (kcal/(mol $\cdot$ nm^2)))"

Umbrella center: C	Umbrella center: $\bar{d}$ (nm)	Restraining constants	Simulation length (million steps)
30.0:45.0:5.0	10.0:25.0:2.5	(0.1, 0.05)	20
10.0:20.0:5.0	6.0:10.0:0.5	(0.120, 1.20)	10
10.0:20.0:5.0	10.0:25.0:2.5	(0.478, 0.239)	10
10.0:20.0:5.0	10.0:25.0:2.5	(0.120, 0.0478)	10
10.0:20.0:5.0	9.0:9.5:0.5	(0.120, 4.78)	10
25.0:45.0:5.0	6.0:10.0:0.5	(0.120, 1.20)	20
25.0	10.0:25.0:2.5	(0.120, 0.0478)	20
25.0:45.0:5.0	9.0:9.5:0.5	(0.120, 4.78)	20
25.0	10.0	(0.120, 4.78)	20
25.0:45.0:5.0	10.0	(0.478, 0.239)	20
30.0:45.0:5.0	10.0	(0.120, 0.0478)	20
30.0:40.0:5.0	12.5:15.0:2.5	(0.478, 0.239)	20

Table C.3: Summary of umbrella simulation details for free energy calculations using d_{stack} and α as collective variables. The same format as in Table A.1 is adopted here. The two restraining constants are shown in the format "($\kappa_{d_{\text{stack}}}$ (kcal/(mol $\cdot$ nm^2)), κ_α (kcal/mol))"

Umbrella center: d_{stack} (nm)	Umbrella center: α	Restraining constants	Simulation length (million steps)
10.0:35.0:2.5	2.0:8.0:2.0	(0.05, 0.2)	10
22.5:27.5:2.5	2.0:8.0:2.0	(0.2, 0.2)	10
10.0:25.0:2.5	2.0:4.0:2.0	(0.05, 0.5)	10

Table C.4: Summary of simulations with moving restraints to target chromatin configurations towards specific umbrella centers using d_{stack} and α as collective variables. The format for "restraining constants and centers" is $(\kappa_{d_{\text{stack}}}$ (kcal/(mol $\cdot$ nm^2)), κ_α (kcal/mol)), ($d_{\text{stack},0}$ (nm), α_0 (1)). We only listed the restraining constants and centers at simulation time of zero, 4×10^5 and 5×10^5 steps, and values in between these time points were updated via linear interpolation during the simulation.

Umbrella center: d_{stack} (nm)	Umbrella center: α	Restraining constants and centers at $t = 0$	Restraining constants and centers at $t = 4 \times 10^5$ steps	Restraining constants and centers at $t = 4 \times 10^5$ steps
10.0:25.0:2.5	2.0:4.0:2.0	(50, 200), (30, 1)	(50, 200), (d_{stack}, α)	(0.05, 0.2), (d_{stack}, α)
10.0:20.0:2.5	6.0:8.0:2.0	(50, 200), (20, 1)	(50, 200), (d_{stack}, α)	(0.05, 0.2), (d_{stack}, α)
22.5:25.0:2.5	6.0:8.0:2.0	(50, 200), (30, 1)	(50, 200), (d_{stack}, α)	(0.05, 0.2), (d_{stack}, α)
27.5:35.0:2.5	2.0:8.0:2.0	(50, 200), (30, 1)	(50, 200), (d_{stack}, α)	(0.05, 0.2), (d_{stack}, α)

Table C.5: Summary of umbrella simulation details for free energy calculations with two 12-mers with fully rigidified nucleosomes. The same format as in Table A.1 is adopted here, and the units for the two restraining constants are κ_C (kcal/mol), $\kappa_{\bar{d}_{\text{stack}}}$ (kcal/(mol $\cdot$ nm^2)).

Umbrella center: C	Umbrella center: $\bar{d}_{\text{stack}}$ (nm)	Restraining constants	Simulation length (million steps)
30.0:55.0:5.0	6.0:10.0:0.5	(0.0478, 0.478)	20
30.0:55.0:5.0	10.0:17.5:2.5	(0.0478, 0.239)	20

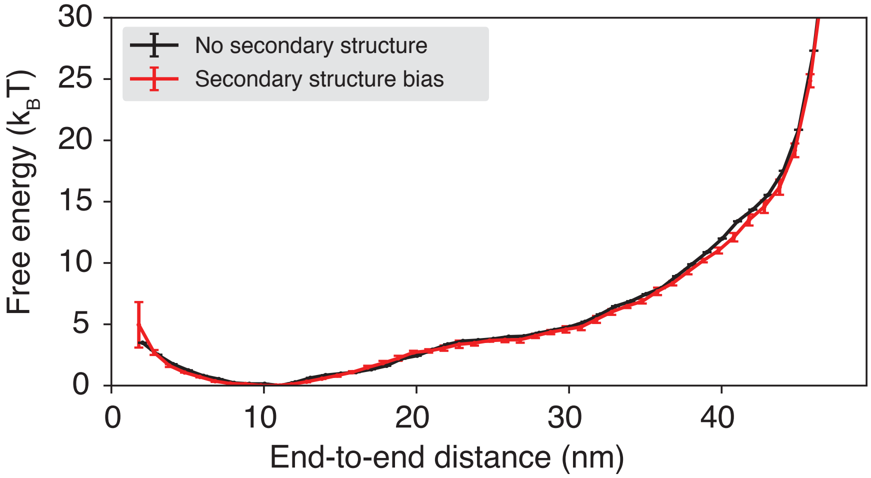


Figure C.1: Secondary structure motifs for disordered histone tails negligibly impact nucleosome stability and protein-DNA interactions. The two curves correspond to the free energy profiles of the outer layer nucleosomal DNA unwrapping as a function of the DNA end-to-end distance. These profiles were determined from replica-exchange umbrella simulations with biases on the end-to-end distance of the nucleosomal DNA. The two sets of simulations only differ in the treatment of histone tails but otherwise share identical settings. The black curve was computed using simulations performed with the same model as that presented in the main text. On the other hand, the red curve was determined using simulations that explicitly accounted for secondary structure biases in the disordered histone tails. In particular, we used AlphaFold2 [322] to predict the structure of all the histone tails. We built new structure-based models for histone tails that account for the bonds, angles, and dihedrals from these initial structures. Therefore, the new models should reproduce the residue folding of histone tails and their tendency to form any secondary/tertiary structures. The umbrella centers were placed on a uniform grid [5.0:70.0:5.0] nm. The temperature replica exchange was applied between temperatures from 300 K to 410 K with a spacing of 10 K. Each simulation replica lasted for 5.5 million steps with a time step of 10.0 fs, and the first 250k steps were excluded for equilibration. We used the WHAM algorithm [311] to process the simulation data from all temperatures and compute the free energy profiles. Error bars correspond to the standard deviation of the means estimated from three independent data blocks.

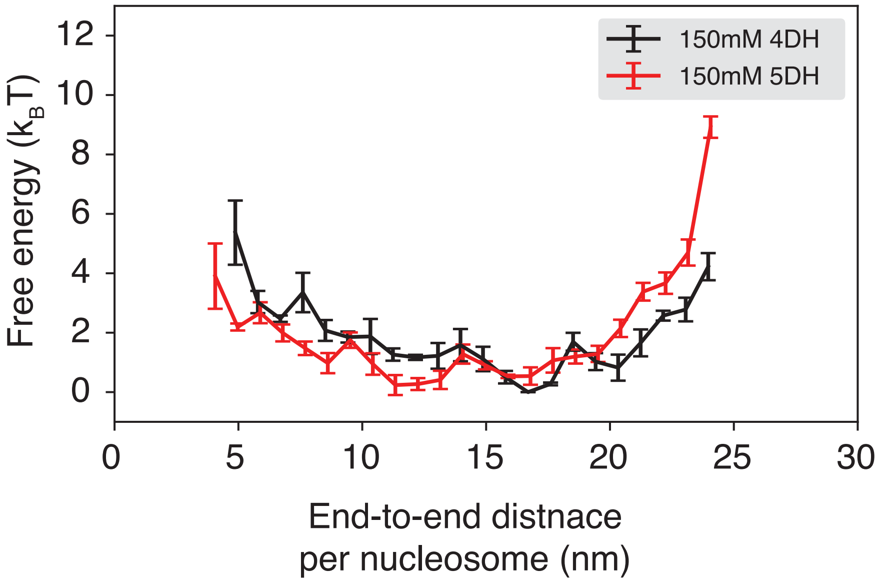


Figure C.2: The cutoff distance used for the Debye Hückel potential has negligible impact on the computed free energy profile. The black line is identical to the one presented in Figure 2.1. The red curve was computed with a new set of simulations that adopted a cutoff distance of five times Debye screening length. The new simulations were carried out following the same simulation protocol as those presented in the main text with the presence of 4 pN force.

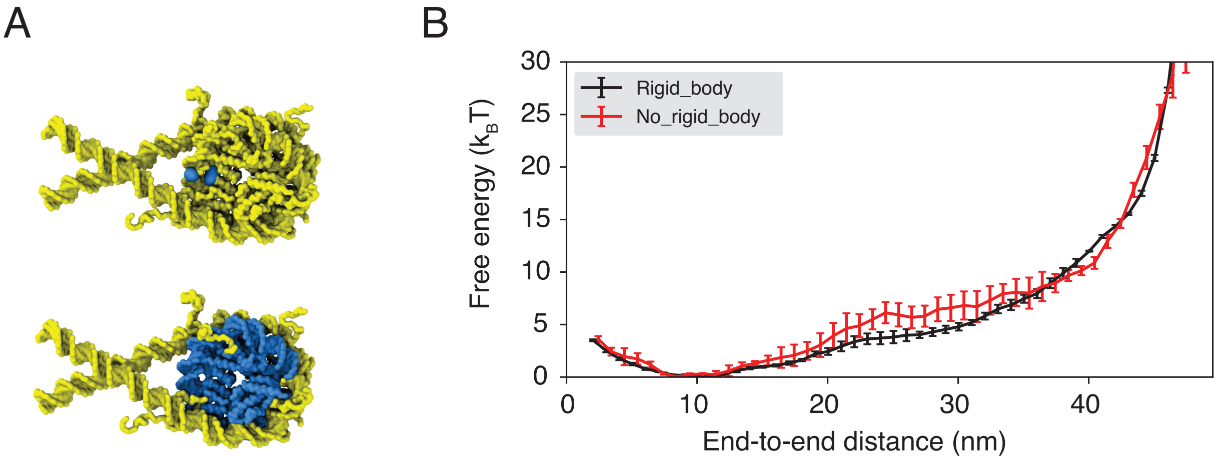


Figure C.3: Rigidifying the inner layer nucleosomal DNA does not impact the energetics of outer layer DNA unwrapping. (A) Illustration of the groups of atoms rigidified in simulations. For simulations presented in the main text (bottom), both the histone core and inner layer (73 bp) of nucleosome DNA (shown in blue) are treated together as one rigid body. As an alternative treatment (top), we only rigidified the four residues and two nucleotides (shown in blue) located on the dyad axis to avoid nucleosomal DNA sliding. (B) Free energy profiles of outer layer nucleosomal DNA unwrapping as a function of the DNA end-to-end distance. These profiles were determined from replica-exchange umbrella simulations with biases on the end-to-end distance of the nucleosomal DNA. The two sets of simulations only differ in the treatment of rigid groups, as illustrated in part A, but otherwise share identical settings. The umbrella centers were placed on a uniform grid [5.0:70.0:5.0] nm. The temperature replica exchange was applied between temperatures from 300 K to 410 K with a spacing of 10 K. Exchanges among the replicas were attempted every 100 steps. Each simulation replica lasted for at least 5.5 million steps. The simulations that rigidified both the histone core and inner layer of nucleosomal DNA used a time step of 10.0 fs. The simulations that only rigidified the four residues and two nucleotides on the dyad axis require a smaller time step of 1.0 fs to ensure energy conservation. In both cases, the first 250k steps were excluded for equilibration. We used the WHAM algorithm [311] to process the simulation data from all temperatures and compute the free energy profiles. Error bars correspond to the standard deviation of the means estimated from three independent data blocks.

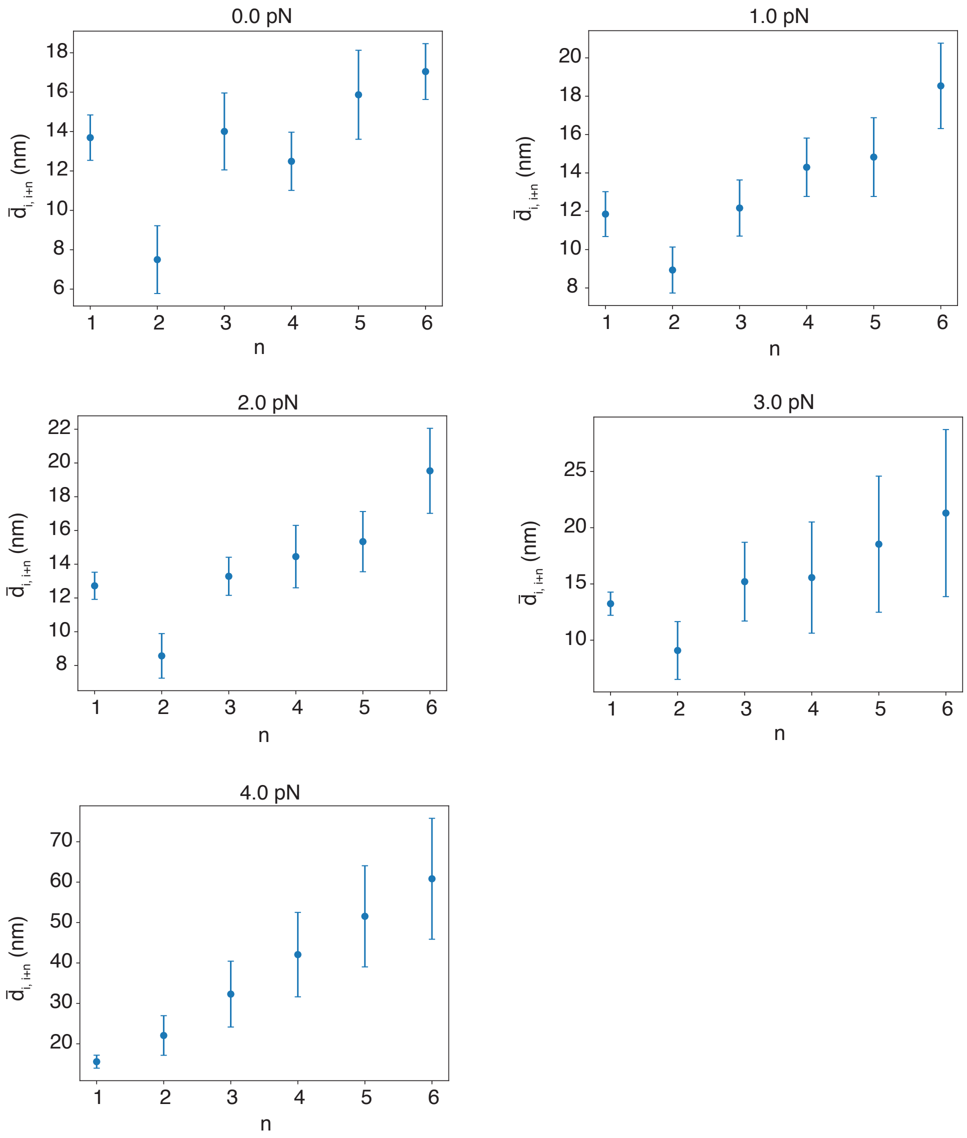


Figure C.4: Mean distances between pairs of nucleosomes at various values of nucleosome separation n. Error bars correspond to the standard deviation of the mean estimated from three independent data blocks. These data suggest that the average distance between nucleosome pairs separated by four or more nucleosomes is larger than 13 nm. Therefore, nonbonded interactions between these nucleosomes contribute negligibly to the overall potential energy and stability of the chromatin structure. Therefore, neglecting their contribution to the chromatin conformational free energy in the neural network model is a reasonable approximation. See *Section: Neural Network Model for the 12mer Chromatin* for more details on the neural network model.

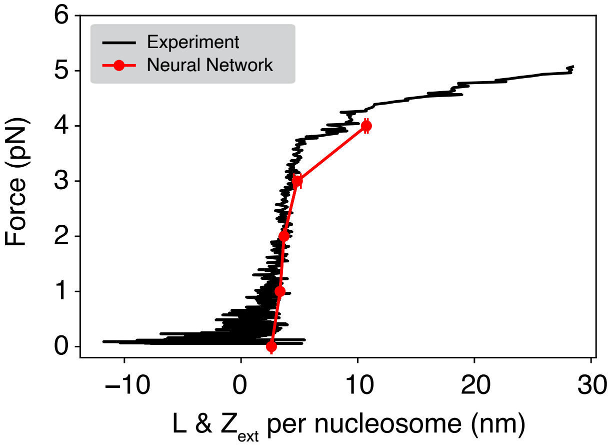


Figure C.5: Comparison between experimental [202] force-extension curve (black) and the one predicted by the neural network model. The neural network model quantifies chromatin stability as a function of inter-nucleosome distances. Based on the derivation shown in Eq. C.9, when the extension force is larger than 1 pN, the extension along z-axis (L_z) is very close to the end-to-end distance (L), so that we approximated the z-axis extension per nucleosome using the distance between first and last nucleosome (L) divided by 11. L at different extension forces was calculated using umbrella simulations of the neural network model. See text *Section: Initial configurations from the neural network model* and *Section: Neural Network Model for the 12mer Chromatin* for simulation details.

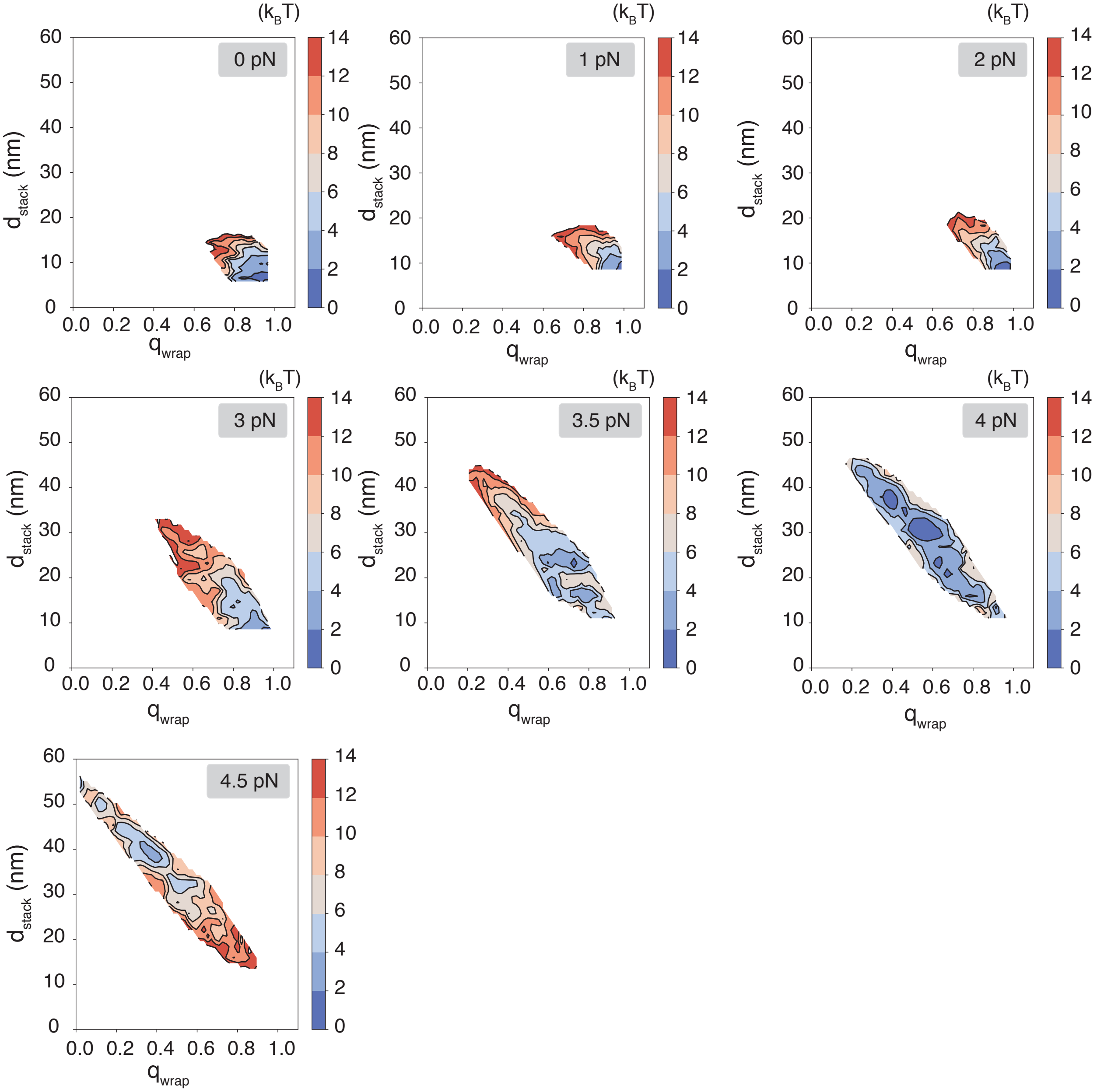


Figure C.6: Two dimensional free energy profiles as a function of nucleosome unwrapping (q_{wrap}) and unstacking (d_{stack}) at various extension forces determined from umbrella simulations. See text *Section: Free Energy Profiles for Chromatin Under Tension* for simulation details.

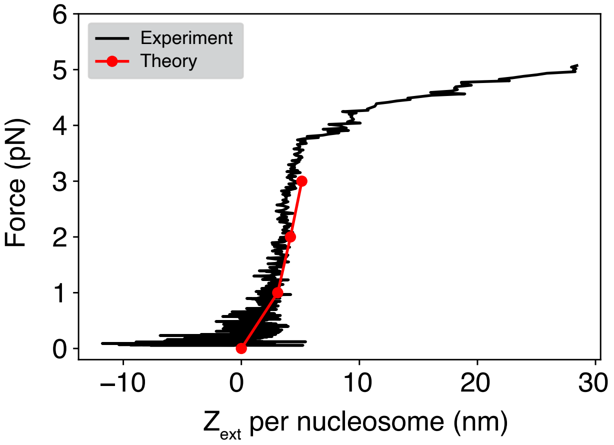


Figure C.7: Theoretical predictions of chromatin extension along the z-axis, Z_{ext}. We assumed a harmonic potential for the end-to-end distance of the unbiased chromatin. Parameters in the potential were obtained from a least-square fitting to the simulation results shown in Figure 2.1C at 0 pN. From the harmonic potential, Z_{ext} can be computed with the analytical expression provided in Eq. C.9. See *Section: Theoretical predictions of chromatin extension along the z-axis* for a detailed discussion.

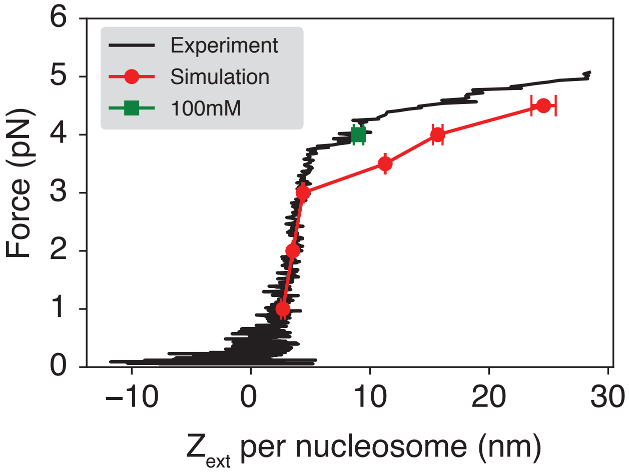


Figure C.8: Comparison between the simulated (red) and experimental [202] (black) force-extension curves. The results for simulations performed with 150 mM monovalent ions are reproduced from Fig. 1B. The green dot corresponds to chromatin extension at 4pN force obtained from simulations with 100 mM monovalent ions. We note that while previous experimental studies [285] have shown that lower salt concentrations lead to chromatin decompaction, our results do not contradict them. A critical difference between the results presented here and previous experimental studies is the presence of force. In previous studies, chromatin was probed without any tension and should, in general, adopt compact conformations. For compact chromatin, linker DNAs come in close contact and contribute significantly to chromatin stability. Therefore, factors that affect their repulsion, such as increasing salt concentration, will dramatically impact chromatin extension. However, with 4 pN force, chromatin adopts much more extended configurations with very few contacts between linker DNA (Figure 2.2). Histone-DNA interactions become more important for chromatin stability and extension in these configurations as many nucleosomes have unwrapped. Therefore, lowering the salt concentration would enhance attraction between histone proteins and DNA to stabilize individual nucleosomes and reduce chromatin extension. Consistent with this interpretation, many experimental studies have shown that nucleosome unwrapping becomes more prevalent at higher salt concentrations [323]–[326].

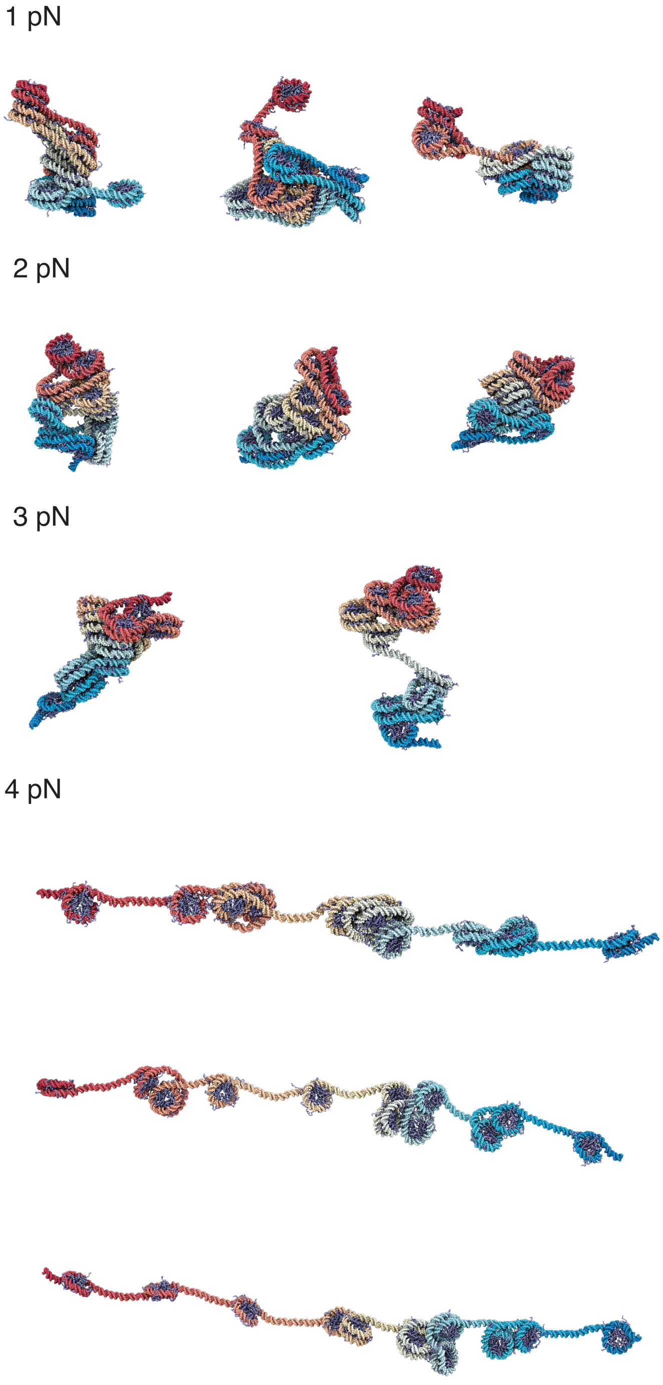


Figure C.9: Additional representative chromatin structures from simulations performed under various extension forces. The values for the extension force are provided next to the structures. Similar to the ones shown in Figure 2.2, these structures correspond to the central configurations of the clusters identified by the single-linkage algorithm using root mean squared distance (RMSD) as the distance between structures.

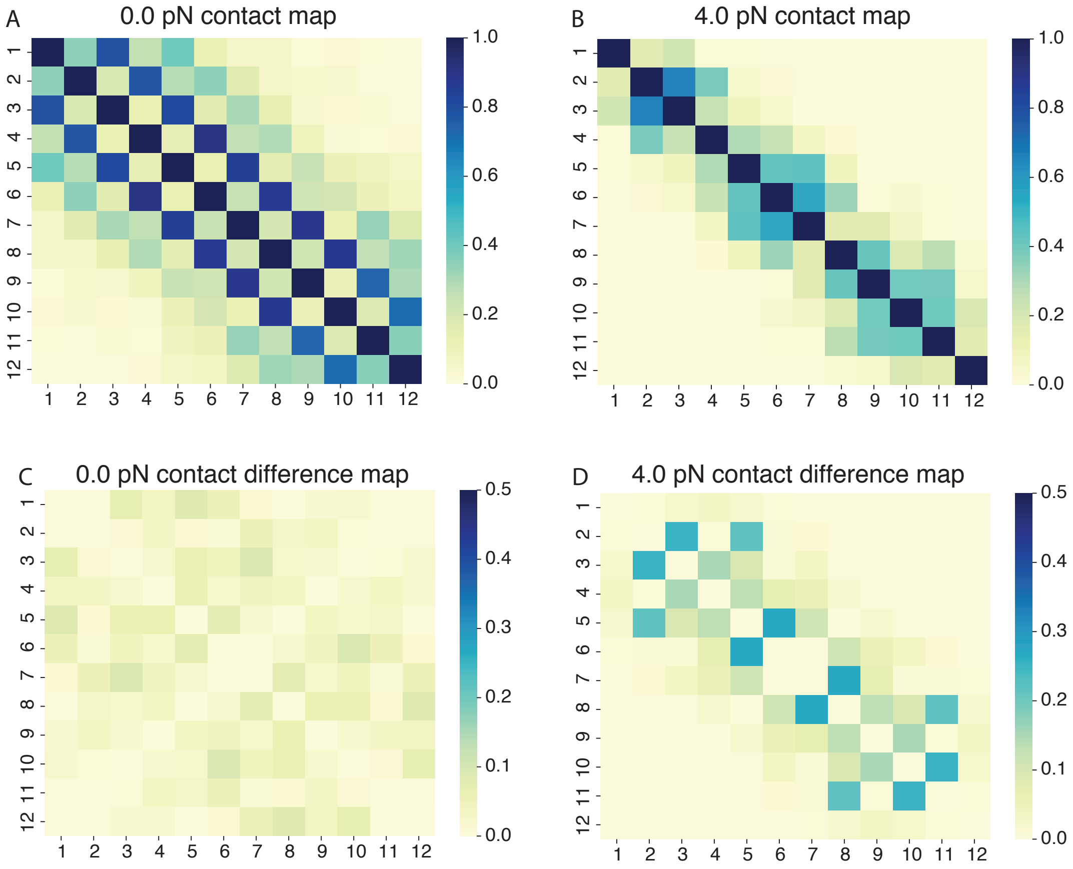


Figure C.10: The ensemble of simulated chromatin configurations at different forces satisfy the C_2 symmetry. (A, B) Average nucleosome pair-wise contact maps computed using chromatin structures simulated with the presence of 0 and 4 pN force. The contact between nucleosome pairs (i, j) is defined as $c_{ij} = \left\langle \frac{1-\left(\frac{r_{ij}-d_0}{r_0}\right)^n}{1-\left(\frac{r_{ij}-d_0}{r_0}\right)^m} \right\rangle$ with $d_0 = 3$ nm, $r_0 = 8$ nm, $n = 6$, and $m = 12$. The angular brackets $\langle\cdot\rangle$ represent ensemble averaging. (C, D) Difference in contacts between pairs of nucleosomes defined as $\Delta c_{ij} = |c_{ij} - c_{13-i,13-j}|$. The difference in contacts was designed to examine the C_2 symmetry of the system. For example, we anticipate that for the 12mer, 1-2 nucleosomes should have comparable contacts as 11-12, 1-3 nucleosomes should have similar contacts as 10-12, etc. We note that the 12mer does not have translational symmetry, since n and $n + m$ nucleosomes are not identical due to the boundary effects and the finite length of chromatin.

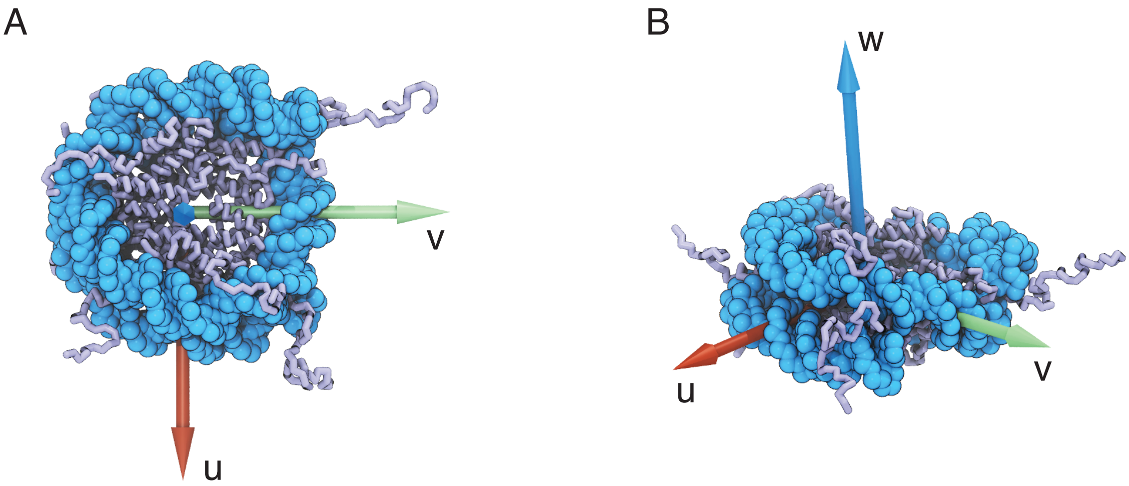


Figure C.11: Illustration of the nucleosome coordinate system used to distinguish shearing and normal motions. The nucleosome is shown in the coarse-grained representation derived from the crystal structure (PDB ID: 1KX5) [4]. The origin of the coordinate system is defined as the center of residues 63-120, 165-217, 263-324, 398-462, 550-607, 652-704, 750-811, and 885-949. The red arrow points from the origin to the center of residues 63-120, 165-217, 750-811, and 885-949. The green arrow points towards the nucleosome dyad defined as the center of residues 81-131 and 568-618. The blue arrow is defined as the cross product of vectors along the red and the green arrows. See text *Section: Decomposing Inter-nucleosome Distances into Shear and Normal Motions* for further discussions.

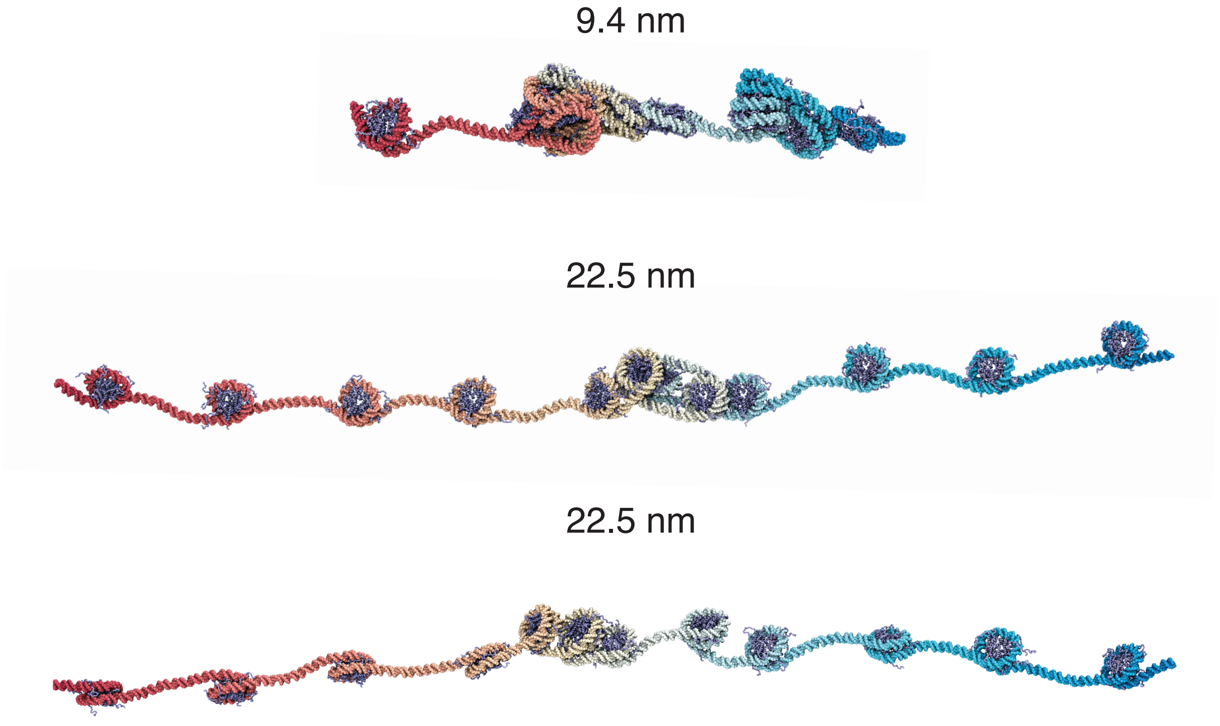


Figure C.12: Additional representative chromatin structures at smaller and larger distances than the average extension at 4 pN force. The end-to-end distances are provided above the structures.

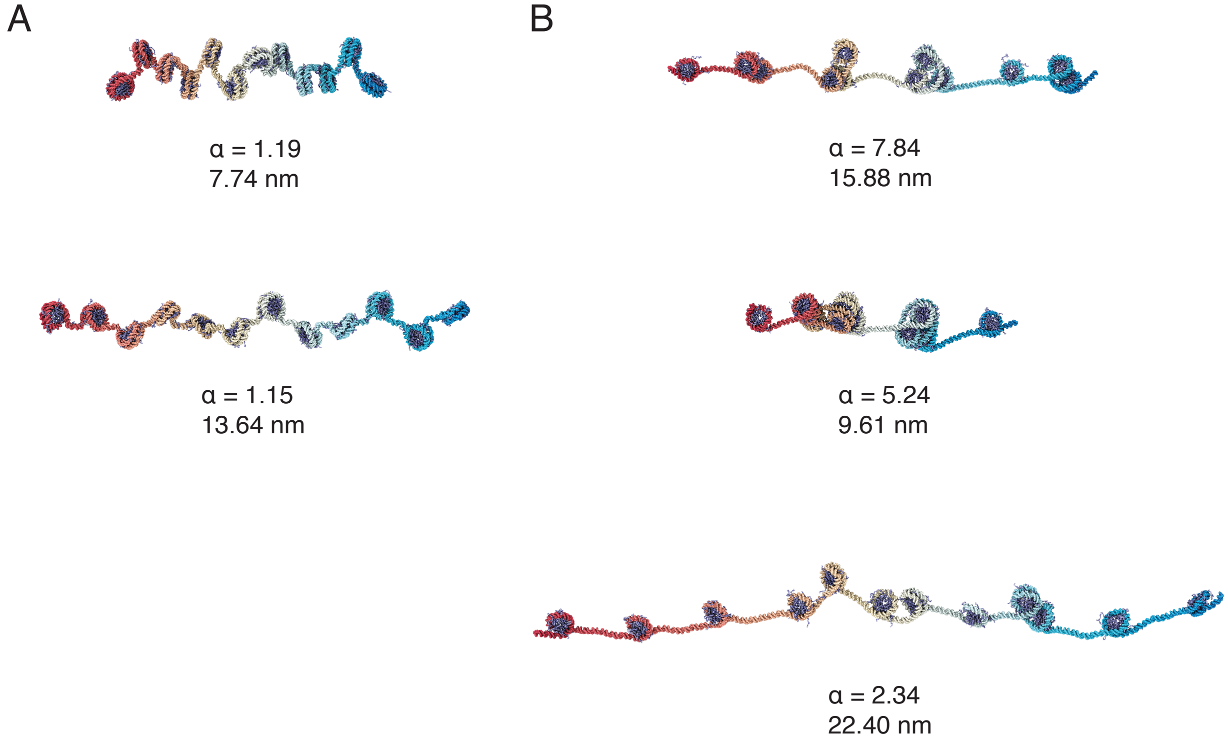


Figure C.13: Simulations initialized from uniform chromatin configurations produce clutched structures. See text *Section: Simulations starting from uniformly extended chromatin configurations* for additional simulation details. (A) Illustration of the two uniformly extended configurations used to initialize the umbrella simulations. (B) Representative chromatin structures with different end-to-end distances per nucleosome produced by umbrella simulations. We selected configurations with the most likely α values. Numbers below the structures correspond to values for α and the end-to-end distance per nucleosome.

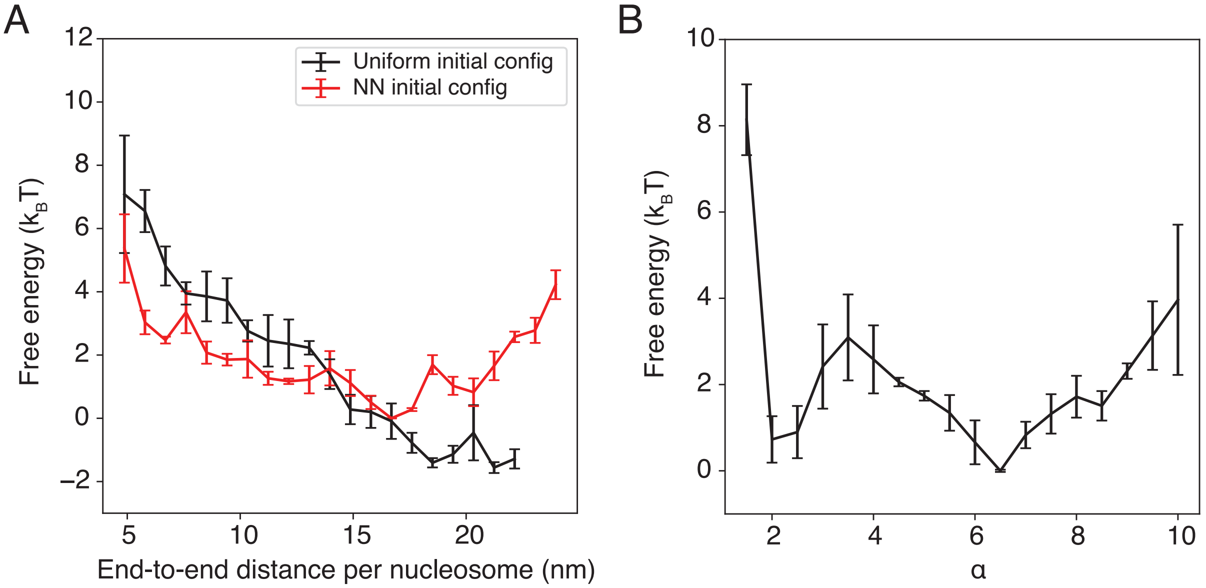


Figure C.14: Simulations with uniform chromatin configurations reproduce findings presented in the main text. See text *Section: Simulations starting from uniformly extended chromatin configurations* for additional simulation details. (A) Comparison of the two free energy profiles as a function of end-to-end distance per nucleosome obtained from simulations with uniform chromatin configurations (black) and with configurations predicted by the neural network model (red). The red curve is identical to that presented in Figure 2.1C of the main text. The statistical equivalence of two independent sets of simulations initialized with different configurations within error bars supports the convergence of our results. We note that the residual differences between the two free energy profiles highlight the challenges of sampling chromatin configurations, which motivated our use of initial configurations predicted by the neural network model for simulations presented in the main text. (B) Free energy profile as a function of α. The global minimum at large α value supports the formation of clutched chromatin configurations.

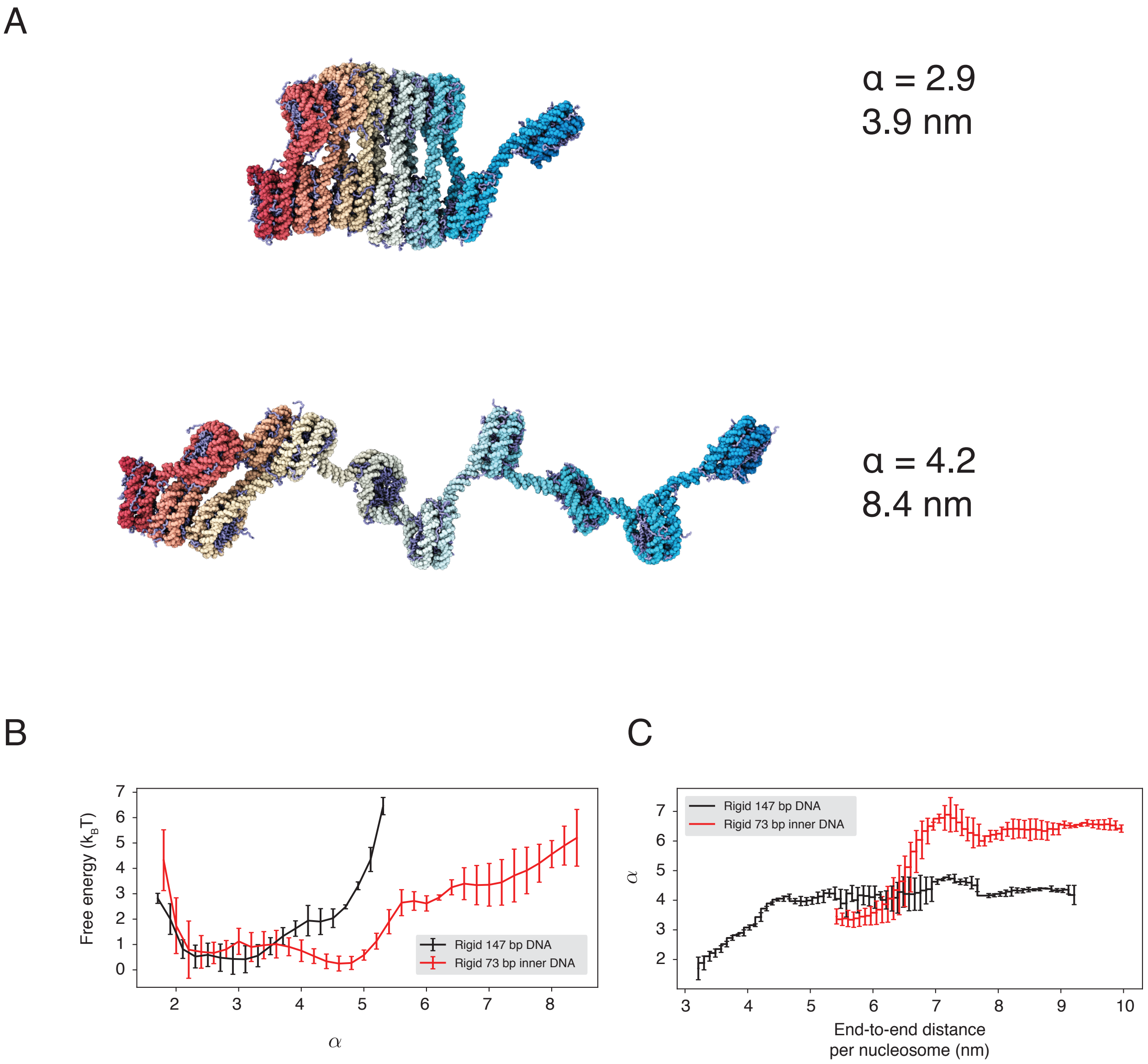


Figure C.15: Restricting nucleosomal DNA unwrapping reduces clutch formation. See text *Section: Simulations with Fully Rigidified Nucleosomes* for additional simulation details. (A) Representative chromatin structures with different end-to-end distances produced by umbrella simulations. We selected configurations with the most likely α values. Numbers next to the structures correspond to values for α and the end-to-end distance per nucleosome. (B) Free energy profiles as a function of $\alpha = d^{\max}_{i,i+2}/d^{\min}_{i,i+2}$ calculated from simulations under 4 pN tension with the entire 147 bp nucleosomal DNA rigidified (black) and with only the inner 73 bp nucleosomal DNA rigidified (red). (C) The average value of α calculated as a function of the per-nucleosome DNA end-to-end distance from simulations under 4 pN tension with the entire 147 bp nucleosomal DNA rigidified (black) and with only the inner 73 bp nucleosomal DNA rigidified (red). Error bars are calculated from the standard deviation estimated via block averaging. For a better comparison between these two sets of simulations, we only show data with per-nucleosome end-to-end distance below 10 nm.

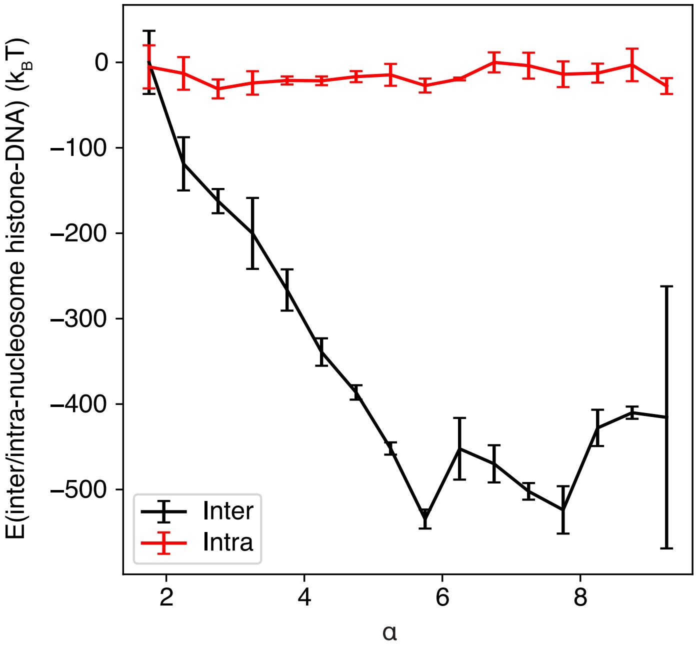


Figure C.16: Correlation between $\alpha = d_{i,i+2}^{\max}/d_{i,i+2}^{\min}$, the ratio between maximum and minimum values of the 1-3 nucleosome stacking distance, and the inter or intra-nucleosome histone-DNA interaction energies. α was introduced to quantify the degree of irregularity in chromatin structure. As the name suggests, The intra-nucleosome energy (red) only accounts for the interactions between histone proteins and DNA segments from the same nucleosome, while the inter-nucleosome energy (black) quantifies interactions from different nucleosomes. The two curves were computed using data from simulations with the 4 pN force presented in the main text. They were shifted to set the maximum values as zero. The errorbars correspond to the standard deviation of the mean computed via block averaging.

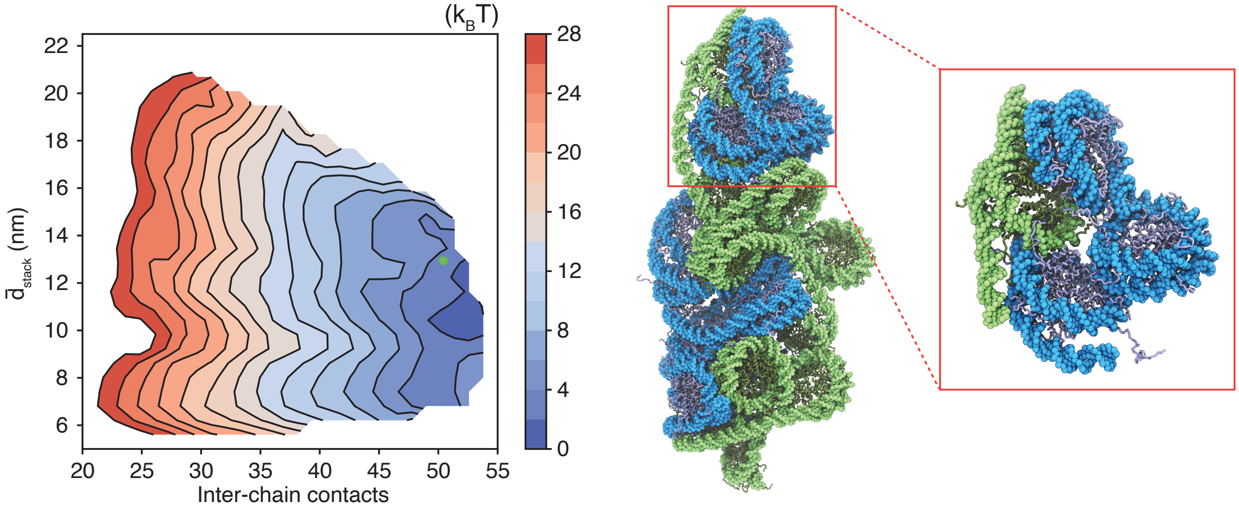


Figure C.17: Representative structure of two contacting chromatin segments that adopt more extended configurations. Extension leads to more interdigitation between the two chains. The inset highlights the interactions between inter-chain nucleosomes. The free energy and collective variable value are indicated as the green dot in the free energy profile.

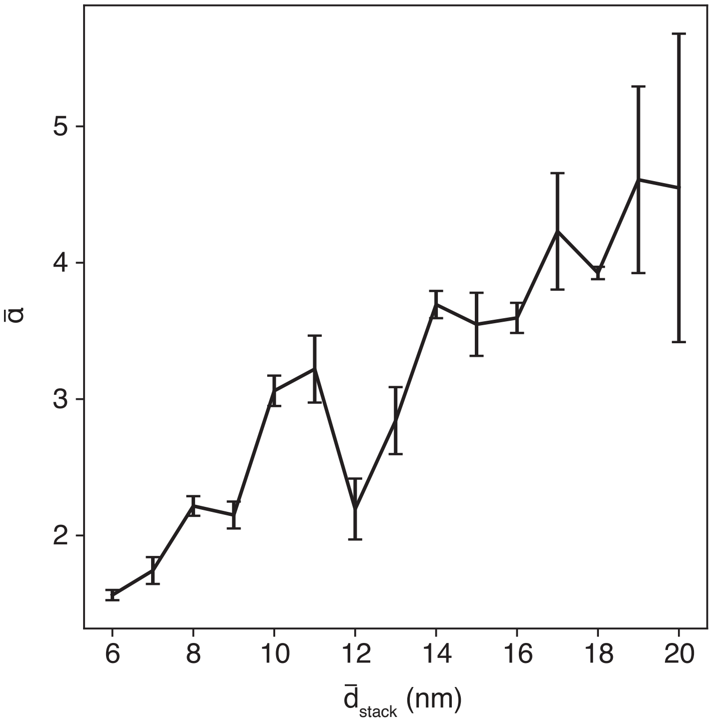


Figure C.18: The average value of $\bar{\alpha}$ as function of $\bar{d}_{\text{stack}}$ determined using the same simulations presented in Figure 2.5. $\alpha = d^{\max}_{i,i+2}/d^{\min}_{i,i+2}$ was introduced to quantify the degree of irregularity in chromatin structure. We averaged over two chromatin segments to define the mean value as $\bar{\alpha} = (\alpha_{\text{fiber 1}} + \alpha_{\text{fiber 2}})/2$. The errorbars measure the standard deviation of the mean and were estimated from three independent data blocks. This plot supports the formation of irregular chromatin configurations with nucleosome clutches (larger $\bar{\alpha}$ values) as chromatin extends to break stacking interactions (higher $\bar{d}_{\text{stack}}$ values).

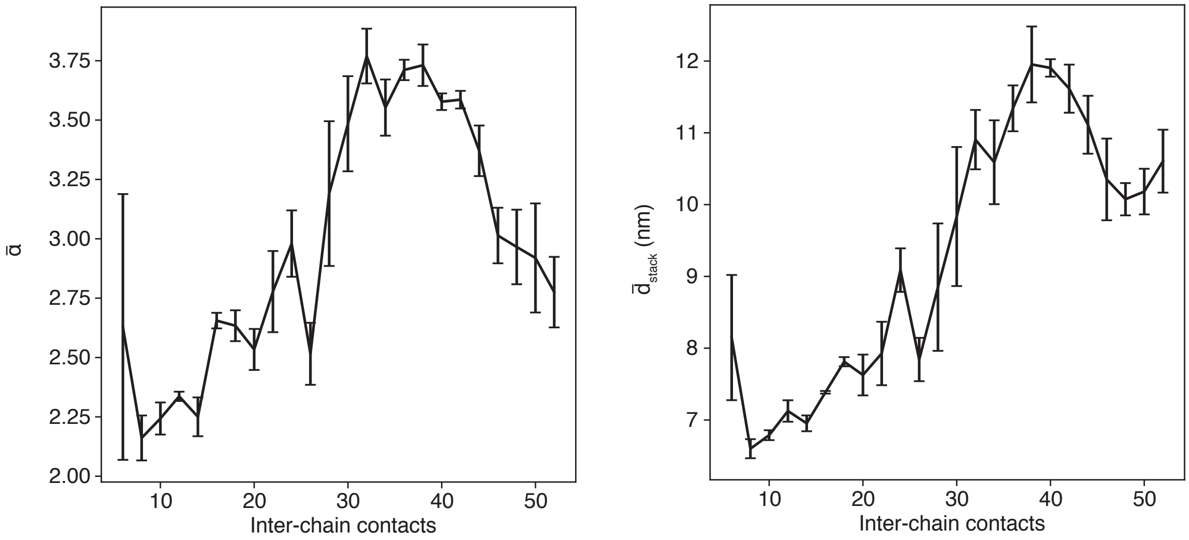


Figure C.19: Average $\bar{\alpha}$ (Left) and $\bar{d}_{\text{stack}}$ (Right) as a function of inter-chain contact numbers determined using simulations presented in Figure 2.5. The error bars measure the standard deviation of the mean and were estimated from three independent data blocks. The two plots support that chromatin become more irregular (larger $\bar{\alpha}$ values) and extended (larger $\bar{d}_{\text{stack}}$ values) as contacts form. The slight decrease in $\bar{\alpha}$ for very large contacts arises from chromatin compaction as seen in the drop for $\bar{d}_{\text{stack}}$. More contacts necessitate more compact chromatin configurations.

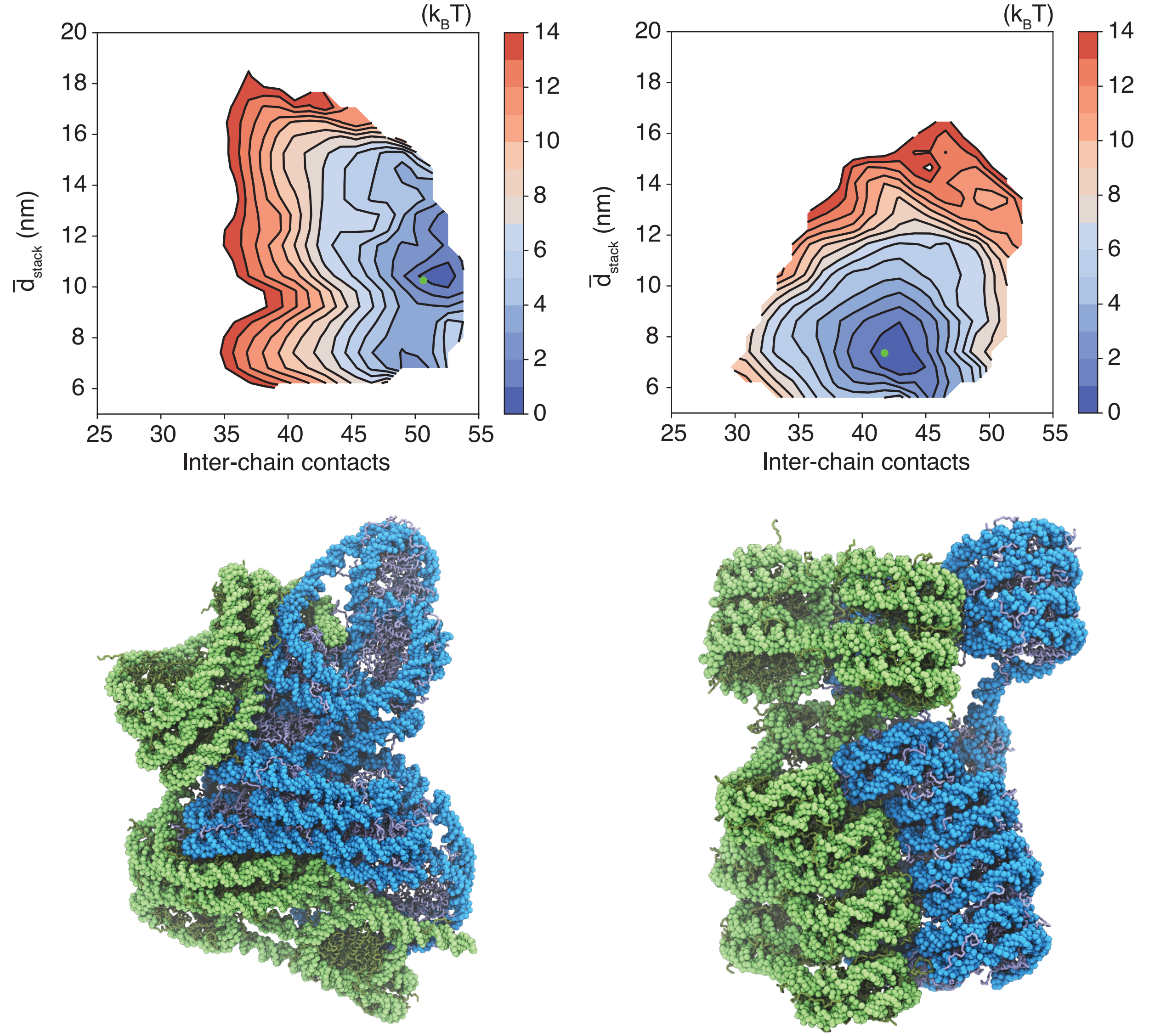


Figure C.20: Free energy surface as a function of the inter-chain contacts and the average extension of the two 12mers determined with simulations that permit (left) or prohibit (right) outer nucleosomal DNA unwrapping. The left plot is identical to Figure 2.5A but with a different color scale. The right plot was computed with a new set of umbrella simulations in which the entire 147 bp nucleosomal DNA was rigified together with the histone core. Representative structures near the free energy minimum are shown below, with the collective variable values indicated as green dots in free energy surfaces. See *Section: Simulations with Fully Rigidified Nucleosomes* for simulation details.

Appendix D

Supplementary information for chapter 5

D.1 Simulation Details

D.1.1 System Setup

We built three tetra-nucleosome models with DNA linker lengths of 20, 25, and 30 bp, respectively. No additional DNA was attached to either end of the tetra-nucleosome. This workflow was introduced in a previous publication [49], and we briefly summarize it here.

We built tetra-nucleosome models by connecting nucleosomes with given DNA linker lengths. Single nucleosome structure is extracted from the tetra-nucleosome x-ray structure (PDB: 1ZBB) [8] with 147 bp wrapped around the histone and an additional 20 bp at the exiting end. The histones were then replaced with the structure from PDB 1KX5 [4], which includes histone tails. Such a 167-bp nucleosome is the building block for all tetra-nucleosome models.

To concatenate the 167 bp blocks, we used 3dna [312] to build small segments of DNA and align them with the 167 bp blocks. We also removed additional DNA linkers at either end. By adjusting the lengths of the short DNA segments, we constructed tetra-nucleosomes with different linker lengths.

D.1.2 Force Field setup

We combined 3SPN.2C for DNA [146], [206] and SMOG for protein [133]. 3SPN.2C models each nucleotide as 3 CG beads, and SMOG models each amino acid as 1 CG bead. Relative to the default SMOG parameters, we scaled all the bonded interaction strength by 2.5 to prevent unfolding at 300 K. The nucleosome core is defined as the histone core (residue ID: 44-135, 160-237, 258-352, 401-487, 531-622, 647-724, 745-839, 888-974) and the middle 73 bp wrapped around histone (i.e. inner layer DNA that wraps around the histone). Each nucleosome core is set as a rigid body to prevent the inner layer DNA from unwrapping and sliding. This also stabilizes the histone core. All the protein dihedral potentials involving histone tail atoms (i.e. histone atoms not within histone core belong to histone tails) are removed, as tails are intrinsically disordered. Specific nonbonded interactions between amino acids are captured with Miyazwa-Jernigan (MJ) potential [205] scaled by 0.4. Protein-DNA nonbonded interactions include Lennard-Jones and Debye-Hückel potentials. Debye length is computed under 150 mM monovalent salt and 300 K. We used the same force field to successfully capture the folding dynamics of tetra-nucleosome and chromatin 12mer [46], [49]. The force field with the parameters introduced before has been implemented in a CG simulation force field package named OpenABC and can be readily used [61].

D.1.3 Select Initial Structures from Enhanced Sampling Trajectories

We intend to start our simulations from diverse configurations explored with enhanced sampling techniques. The sampling method and configurations are reported in the previous work [46]. Here we briefly summarize the protocol applied in the previous study.

In the previous study [46], the tetra-nucleosome structure was initialized as the x-ray structure (PDB: 1ZBB) [8]. This structure has linkers of lengths 20 bp. The same CG force field was applied. Two collective variables (CVs), the radius of gyration (R_g, equation D.1) and the fraction of native contacts (Q, equation D.2) were defined to help enhance sampling.

$$R_g = \sqrt{\frac{1}{4}\sum_{i=1}^{4}(\mathbf{r}_i - \mathbf{r}_{\text{COM}})^2} \tag{D.1}$$

$$Q = \frac{1}{6}\sum_{i=1}^{3}\sum_{j=i+1}^{4}\exp\left(-\frac{(r_{ij}-r_{ij}^{o})^2}{2\sigma^2}\right) \tag{D.2}$$

$\mathbf{r}_i$ is the position of the i-th nucleosome, and $\mathbf{r}_{\text{COM}}$ is the center of mass (COM) of all the nucleosomes. r_{ij} is the distance between the i-th and j-th nucleosomes, and r_{ij}^{o} is the distance within the native structure (i.e. x-ray structure). R_g measures whether the structure expands, and Q measures how much the structure deviates from the native one.

The enhanced sampling method called unified free energy dynamics (UFED) [281] was applied to explore diverse configurations. This method combines temperature accelerated molecular dynamics (TAMD) [327] and metadynamics [328]. The two target CVs, R_g and Q, were coupled to fictitious variables, r and q, respectively, with harmonic bias (equation D.3). Here force constants $k_r = 200$ kJ/mol/nm^2 and $k_q = 10000$ kJ/mol.

$$V_{\text{harmonic}} = \frac{k_r}{2}(R_g(\mathbf{r}) - r(t))^2 + \frac{k_q}{2}(Q(\mathbf{r}) - q(t))^2 \tag{D.3}$$

The two fictitious variables, r and q, were coupled to a thermostat of $T_c = 1000$ K, while the real coordinate variables $\mathbf{r}$ were coupled to a thermostat of $T = 300$ K. The higher temperature for fictitious variables together with the harmonic potential accelerated dynamics along R_g and Q and helped cross free energy barriers. Additionally, similar to metadynamics [328], a history-dependent bias is applied to r and q (equation D.4), thus avoiding revisiting explored states. Potential height $h = 0.1$ kJ/mol, potential deposit interval Δt is 500 steps, width parameters $\sigma_r = 0.5$ nm and $\sigma_q = 0.02$.

$$V_{\mathrm{metad}} = h \sum_{i, i\Delta t < t} \left[\exp\left(-\frac{(r(t) - r(i\Delta t))^2}{2\sigma_r^2} \right) + \exp\left(-\frac{(q(t) - q(i\Delta t))^2}{2\sigma_q^2} \right) \right] \tag{D.4}$$

20 independent UFED simulations were performed, and more than 50 million steps were run for each simulation. The snapshots were saved every 5000 steps, and 293,291 configurations were collected in all. 10,000 representative configurations were selected from the saved configurations with K-means algorithm.

In the current study, considering the symmetry that nucleosome indices (1, 2, 3, 4) is equivalent to (4, 3, 2, 1), we only picked the configurations with $d_{13} \geq d_{24}$ (d_{ij} means the distance between the i and j-th nucleosomes) from the 10,000 configurations, leading to 4643 configurations. We computed d_{ij} for all 4643 configurations. Due to the symmetry that $(d_{12}, d_{13}, d_{14}, d_{23}, d_{24}, d_{34})$ is equivalent to $(d_{34}, d_{24}, d_{14}, d_{23}, d_{13}, d_{12})$, samples collected by MD simulation can be essentially doubled by this mapping, and we do this mapping to double our samples before doing analyses (see *Section: Trajectory Preprocessing* for details). Such symmetry was also enforced in the architecture of the neural network free energy estimator reported before to prevent overfitting [46].

D.1.4 Single Tetra-nulcoeosome Simulation Protocols

Given the tetranucleosome linker lengths as 20 or 25 or 30 bp, we first ran restrained MD to drag the structures towards the target d_{ij} values given by the 4643 selected configurations. Then we ran an unbiased NVT simulation. Here we provide detailed protocols for running single tetra-nucleosome restrained MD and unbiased NVT simulations. All the single tetra-nucleosome simulations were performed with LAMMPS [127] in a cubic box of length 200 nm with Nosé-Hoover integrator [209], [329], the temperature was 300 K, the damping parameter was 1 ps, and the timestep was 10 fs. Each nucleosome core was always fixed as a rigid body.

Restrained MD Simulation

The position of each nucleosome is defined as the geometric center of the following Cα atoms: 44-135:3, 160-237:3, 258-352:3, 401-487:3, 531-622:3, 647-724:3, 745-839:3, 888-974:3. Here n_1-n_2:n_3 means atoms start from index n_1 and end at index n_2 (inclusive) with step size n_3. These Cα atoms are all within the ordered domain of histone. We applied strong umbrella bias on d_{ij} with a strong force constant $\kappa = 1000$ kJ/mol/nm^2 to shift the distances to the target values (equation D.5). Here $d_{ij}^{(k)}$ means the distance between nucleosome i and j from the k-th selected configuration. All the restrained MD simulations lasted 0.2 million steps.

$$V_{\text{restrain}}^{(k)} = \sum_{i<j} \frac{\kappa}{2}(d_{ij} - d_{ij}^{(k)}) \tag{D.5}$$

For the NRL = 172 tetra-nucleosome system, since our initial structure built by the protocol introduced before has slight overlap between CG atoms, we first dragged the 3dna built structure to $(d_{12}, d_{13}, d_{14}, d_{23}, d_{24}, d_{34})$ = (15, 15, 25.98, 15, 15, 15) nm with the same bias shown in equation D.5, and force constant $\kappa = 100$ kJ/mol/nm^2. This preliminary step ensures the structure can extend to a configuration without nonphysical overlap. Then all the restrained simulations started from this extended configuration. For NRL = 167 and 177 systems, the restrained MD directly started from the structure built by 3dna.

Unbiased MD simulation

For a single tetra-nucleosome, we continued the simulation from the final snapshot of the restrained MD by releasing the restraint (i.e. removing the restraint bias shown in equation D.5). Unbiased NVT simulations were performed under 300 K. For NRL = 167 system, at least 0.8 million steps were performed for each simulation. For NRL = 172, 177 systems, 1 million steps were executed for each simulation. The simulations converged after about 20,000 steps. The ensembles of unbiased and parallel trajectories were further analyzed with Markov State Models (MSMs) introduced below.

D.1.5 Tetra-nucleosome in Sea of Nucleosomes Simulation Protocols

We also performed the simulation for a single tetra-nucleosome with 20 bp linkers within a sea of nucleosomes. A tetra-nucleosome and 26 single nucleosomes were placed in a cubic box of length 55 nm, resulting in an overall nucleosome concentration of 0.3 mM. This concentration is close to the one in chromatin condensate [17] and sufficient inter-nucleosomal contacts are observed. Such a large system is slow to run with CPU parallelization. It has been manifested that OpenMM [62] GPU acceleration can execute large systems efficiently [61], [147], and we run all the dense phase simulations with OpenMM and GPUs. We combined SMOG [330] and 3SPN2 [147], which have been implemented into OpenMM separately, into one force field following the OpenABC framework [61]. All the OpenMM NVT simulations were performed in a cubic box of length 55 nm with Langevin middle integrator [130], The friction coefficient was 0.01 ps^{-1}, the temperature was 300 K, and the timestep was 10 fs. Note we used the Langevin middle integrator instead of the Nosé-Hoover integrator as the OpenMM rigid body algorithm is more robust with the Langevin middle integrator (https://github.com/openmm/openmm/issues/3993).

Since the condensate simulation is more expensive and requires longer time to observe equilibration and condensation, we ran simulations with fewer individual trajectories but extended simulation time. We selected 530 representative initial structures from each MSM microstate constructed from unbiased trajectories of the single NRL = 167 tetra-nucleosome (See section *Construction and Validation of Microstate-MSM* for details). The selection guaranteed that all of the chosen structures fell within the 10% standard deviation from the mean value of the Root Mean Squared Distances (RMSD) of each cluster in the MSM. We placed the tetra-nucleosome with the selected configuration at the center of the box, then randomly inserted single nucleosomes into the box with the tools provided by OpenABC [61]. The first stage is to relax the single nucleosomes while fixing the whole tetra-nucleosome. We fixed the whole tetra-nucleosome at given configurations as a whole rigid body, and the nucleosome core of every single nucleosome is a rigid body. We ran NVT simulations for 0.2 million steps to relax the positions of single nucleosomes relative to the tetra-nucleosome. Figure D.2A demonstrates that the relaxation reached equilibrium after about 50,000 steps, thus the relaxation stage is long enough. Next, we released the constraints on tetra-nucleosomes. Each nucleosome core of the tetra-nucleosome is an individual rigid body, while other parts are flexible. Nucleosome cores of single nucleosomes are still kept as rigid bodies. We did energy minimization, then ran NVT simulations for at least 7 million steps (i.e. at least 70 ns) as the production simulation, and analyzed these production trajectories with MSMs. Figure D.2B shows the production run reached equilibrium at about 40,000 steps. The distances between two nucleosomes within tetra-nucleosome, as well as all the nucleosome positions, were recorded every 500 steps with PLUMED [309]. Since the simulation box is not too large, we measured the size distributions of tetra-nucleosome in the production run, and validated that the tetra-nucleosome did not touch its periodic images (Figure D.2C). To ensure the distances between two nucleosomes were computed within the same intact tetra-nucleosome, we first used the PLUMED “WHOLEMOLECULES" command to rebuild the tetra-nucleosome so the whole tetra-nucleosome was built as a complete molecule, then the distances between two nucleosomes within the tetra-nucleosome were computed with this rebuilt configuration by adding “NOPBC” flag.

D.2 Markov State Model and Non-Markovian Dynamics Model Construction

We built three independent MSMs based on the unbiased simulations of single tetra-nucleosomes with different DNA linker lengths (NRL = 167, 172, 177), respectively. For the simulation of the tetra-nucleosome in the sea of single nucleosomes, we constructed the MSM by only considering the features of the tetra-nucleosome. All MSM constructions followed similar protocols:

(a) Select the converged segments of the trajectories and duplicate the converged trajectories according to the reflection symmetry of nucleosome indices. All the following analyses were applied to the converged and duplicated trajectories.

(b) With six inter-nucleosomal distances d_{ij} as input features, apply the time-lagged independent component analysis (tICA) method with kinetic mapping algorithm to identify the collective variables [275]–[278].

(c) Group MD conformations into microstates by the K-Means algorithm according to their kinetic similarities based on the geometric distances in the tICA collective variable space. The hyperparameters of the tICA and clustering (tICA relaxation time, number of tICs, and number of microstates) were optimized using the cross-validation tool: the generalized matrix Rayleigh quotient (GMRQ) [279].

(d) Construct and validate the microstate-MSM by the Chapman-Kolmogorov (CK) test and implied-timescale analysis [265], [266] to ensure the Markovian properties.

(e) Employ the Transition Path Theory (TPT) [260], [261], [274] to elucidate the folding kinetic pathways and the corresponding fluxes.

(f) Lump multiple parallel kinetic pathways into a small set of metastable and representative path channels using latent-space path clustering (LPC) algorithm to facilitate the understanding of folding mechanisms [267], [272], [282].

(g) Group microstates into a few macrostates with the Robust Perron Cluster Analysis (PCCA+) algorithm [331], [332] and implement the Integrative Generalized Master Equation (IGME) theorem [63] to calculate thermodynamical (such as stationary populations of macrostates) and kinetic properties (such as mean first passage time).

All analyses were conducted using in-house Python codes and codes based on MSMBuilder version 3.8.1. [333]. We explain more details in the following sections.

D.2.1 Trajectory Preprocessing

Since the unbiased simulations of three individual tetra-nucleosome systems began with conformations generated by restrained MD simulations, it was essential to exclude the initial portion of each trajectory to ensure that only equilibrium trajectories were considered for subsequent analysis. Consequently, we removed the initial 30,000 steps from each trajectory of all three single tetra-nucleosome systems based on the convergence of their energies (Figure D.1A-C). For the simulation of the tetra-nucleosome in the sea of nucleosome, we discarded the first 50,000 steps of each production run according its energy convergence (Figure D.2B).

Because we focused on the tetra-nucleosome folding process, we neglected the detailed local conformational changes of the histone proteins or DNA segments. Previous studies have shown that six distances between the geometric centers of each pair of nucleosomes can sufficiently describe the large-scale folding of the tetra-nucleosome [46]. Thus, we adopted the strategy of incorporating these six distances as embedded features for each conformation. Since the initial and terminal points of the tetra-nucleosome are indistinguishable, to ensure consistency with the symmetry of the tetra-nucleosome in our analysis, we duplicated the six-distance trajectories by mapping $(d_{12}, d_{13}, d_{14}, d_{23}, d_{24}, d_{34})$ to $(d_{34}, d_{24}, d_{14}, d_{23}, d_{13}, d_{12})$, leading to 4643×2 trajectories. The duplicated ensembles of converged six-distance trajectories were used for further analysis.

D.2.2 Identification of Collective Variables by tICA

To reduce the dimension of features for further analyses, we further employed tICA coupled with kinetic mapping to identify the low-dimensional collective variables (CVs) by recombining the six pairwise distances. The Euclidean distances in the tICA-reconstructed space can be viewed as approximations of the kinetic distances, which facilitate the clustering of kinetically close conformations. To determine the optimal hyperparameters including the number of CVs, the tICA relaxation time, and the number of microstates, we employed the

GMRQ cross-validation method to select the hyperparameters, thus balancing systematic and statistical error. During the cross-validation process, for each system, all trajectories were randomly divided into four subsets. Among these subsets, three were designated as the training set, while the remaining one was used as the validation set. This procedure was repeated 10 times for each system. We selected the optimal hyperparameters to maximize the validation score and minimize the difference between the training and validation scores. We also verified the number of CVs by checking the reconstruction error of the correlation matrix [277], [334]. GMRQ score supports using the top 3 tICA eigenvectors as the CVs (Figure D.3A, D.4A, D.5A, D.6A). For all the four systems, the top three tICA eigenvectors can reconstruct more than 80% of the eigenvalues of the correlation matrix, so we continued following analyses with the top 3 tICs as CVs. For the NRL = 167, 172, 177 and condenstate systems, tICA lag time of 25,000, 12,500, 25,000, 50,000 steps were adopted based on GMRQ scores (Figure D.3B, D.4B, D.5B, D.6B).

D.2.3 Construction and Validation of Microstate-MSM

We further performed K-means algorithm to cluster samples in the CV space. The optimal number of clusters was determined through GMRQ cross-validation. The dataset was randomly divided into a 3:1 ratio for training and testing purposes, and this cross-validation process was repeated 10 times. All MSMs used for the GMRQ test were constructed using the optimal number of CVs and tICA relaxation time identified before. The optimal cluster numbers for the NRL = 167, 172, and 167 condensate systems were determined to be 530, 500, and 400, respectively (Figure D.3C, D.4C, D.6C). For the NRL = 177 system, the training and testing GMRQ scores exhibited minimal changes over a wide range of K-Means clusters (600-1500). Therefore, we selected 1000 as the midpoint for the number of clusters (Figure D.5C).

We constructed the microstate MSMs for the NRL = 167, 172, and 177 systems using a lag time of 0.25 million steps, while a longer lag time of 1.5 million steps for the NRL =

167 condensate system due to the more dominant memory effect induced by the condensed environment of the nucleosomes. The transition count matrices (TCMs) indicate all the systems have reached detailed balance (Figure D.7A, D.8A, D.9A, D.10A). We verified the Markovian properties with the implied time scale (ITS) and the Chapman-Kolmogorov (CK) test. The ITS plots reach plateaus at the corresponding lag times (Figure D.7B, D.8B, D.9B, D.10B), indicating the corresponding lag times lead to Markovian properties. We further validated microstate MSMs with the CK test. This involves comparing the residence probabilities of the eight most populated microstates, as predicted by the microstate MSMs, with the residence probabilities directly obtained from all-atom MD simulations. The CK test confirmed a consistent agreement between the predicted residence probabilities and those derived from the MD simulations (Figure D.7C, D.8C, D.9C, D.10C).

D.2.4 Characterization of Folding Pathways

To comprehend the folding mechanisms of each system, we discerned the kinetic folding pathways for all four systems using Transition Path Theory (TPT) based on microstate MSMs. To define the source of kinetic pathways, we assigned the microstates with center coordinates ranging from 25.0 nm to 32.0 nm along both d_{13} and d_{24}, which represent extended and unfolded configurations. On the contrary, microstates with center coordinates between 5.0 nm and 6.0 nm along both d_{13} and d_{24} were designated as the sink. Given source and sink, we computed the committor probabilities for each microstate in all four systems at their respective Markovian lag times. These probabilities were then used to construct net flux matrices, which characterize the transition flow through the bottleneck of the pathway. We further applied Dijkstra algorithm to identify kinetic pathways [335].

In all four systems, the intrinsic properties of tetra-nucleosome folding resulted in the elucidation of numerous kinetic pathways with comparable fluxes. The highest flux pathways of NRL = 172 and 177 systems are shown in Figure D.11 D.12, and the similar figures for the NRL = 167 systems are shown in the main text. To enhance our comprehension of the folding mechanisms, we employed the LPC algorithm to condense thousands of pathways into a smaller number of metastable path channels. The LPC algorithm utilizes the Variational AutoEncoder (VAE) neural network to learn and embed the path distributions in the CVs space into a low-dimensional latent space, where each pathway is represented as a single point. Subsequently, the K-Means clustering algorithm was applied to group these points into distinct path channels [267].

In all four systems, the kinetic pathways were initially identified and depicted in the 3-dimensional CV space by linking the microstates traversed by the pathways. By projecting the MD conformations belonging to each microstate onto the CV space, the spatial configuration and distribution of each pathway were captured. To simplify the data structure and facilitate training, each pair of two-dimensional CVs (three CVs were combined to form three two-dimensional subspaces) space was divided into 30 × 30 bins to visualize and embed the distribution of pathways. As a result, the distribution for each pathway was described by a $3 \times 30 \times 30$ one-hot vector [267]. These distributions of pathways would subsequently serve as inputs for training the VAE models. We included and embedded 8000, 8000, 11000, and 3500 kinetic pathways for NRL = 167, 172, 177, and 167 condensate systems, respectively. The number of pathways was chosen to encompass approximately 95% or more of the total flux for each individual system (Figures D.13-D.16A).

To ascertain the dimension of the latent space in VAE, we utilized the fraction of variance explained (FVE) by VAE as a criterion [336], which is defined as

$$\mathrm{FVE} = 1 - \frac{\sum_{i=1}^{N} ||\mathbf{D}(i) - \hat{\mathbf{D}}(i)||^2}{\sum_{i=1}^{N} ||\mathbf{D}(i) - \bar{\mathbf{D}}||^2} \tag{D.6}$$

where $\mathbf{D}(i)$ represents the distribution vector for path i in the CVs space, $\bar{\mathbf{D}}$ is the mean of input path distribution vector, and the $\hat{\mathbf{D}}(i)$ is the reconstructed path distribution vector of path i. For all four systems, we trained multiple VAEs using the same input path distribution but with different dimensions of latent spaces to calculate the FVE values. The training of each VAE was conducted ten times, employing different random seeds to estimate the error bars. We selected the dimensionality when the FVE reached convergence, meaning it did not increase by more than 5% when including an additional dimension. As shown in Figures D.13-D.16 B, the dimensionality for the NRL = 167, 172, 177 and 167 in the nucleosome sea system were selected as 5, 4, 5 and 5, respectively.

To determine the number of path channels, we utilized the average of squared errors (ASE) [337] and the Silhouette score [338] as criteria. The ASE represents the average of the distances between the data points and their corresponding cluster centers, which is defined as follows:

$$\text{ASE} = \frac{1}{N}\sum_{i=1}^{K}\sum_{c\in C_i}^{n_i} ||\mathbf{X}_c - \mathbf{M}(C_i)||^2 \tag{D.7}$$

where K is the number of clusters, n_i is the number of data points belonging to cluster C_i, $\mathbf{X}_c$ is the data point c, and $\mathbf{M}(C_i)$ is the center of cluster C_i. We utilized the Elbow method to identify the optimal number of clusters, where the ASE exhibits the most significant change in slope (Figures D.13-D.16 C). This indicates that increasing the number of clusters beyond the optimal number would not improve the clustering performance further.

The Silhouette score quantifies the distinctiveness and separation of clusters, and it is defined as

$$\text{Sihouette Score} = \frac{1}{K\cdot N}\sum_{i=1}^{K}\sum_{c\in C_i}^{n_i}\frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \tag{D.8}$$

where $b(i) = \min\limits_{k\neq i}\frac{1}{n_k}\sum_{j\in C_k} d(i,j)$ denotes the inter-cluster distance, calculated as the average distance to the nearest cluster other than the one which data point i belongs to, while $a(i) = \frac{1}{n_i-1}\sum_{j\in C_i, i\neq j} d(i,j)$ represents the intra-cluster distance, computed as the average distance to all other points within the same cluster, excluding point i. We chose the number of clusters that maximizes the Silhouette score, indicating the optimal separation of data points. As illustrated in Figure D.13-D.16 D, for all four systems, the optimal number of clusters remains consistent for both of these two scores, and the optimal number is 3. The examples of loss decay for training are shown in Figure D.13-D.16 E. The distribution of path channels as heatmaps are shown in Figure D.17.

D.2.5 Construction and Validation of Macrostates-MSM

To generate explainable and representative models, we lumped the microstate models to macrostate models for all four systems using PCCA+ algorithm [331], [332]. The PCCA+ algorithm categorizes microstates into macrostates by considering the sign of the top eigenvectors of the transition probability matrix, which leads to the recombination of microstates based on their kinetic similarities. Across all four systems, varying number of microstates were grouped into six macrostates. To enhance our comprehension, we visualized the macrostate distribution along d_{13} and d_{24} (Figures D.18-D.21 A).

Larger macrostates require more time for the dynamics between states to reach equilibrium, consequently leading to an extended Markovian lag time for the macrostate model [63], [265]. Since we conducted multiple parallel simulations of limited length and the lag time is bounded by the trajectory length, achieving Markovian for the macrostate model becomes a challenge. To address this issue and consider non-Markovian dynamics, we utilized the Integrative Generalized Master Equation (IGME) method to propagate dynamics and predict thermodynamic and kinetic properties [63]. Unlike MSMs that utilize a first-order master equation to evolve dynamics, IGME employs a generalized master equation which considers memory effects, making it applicable to non-Markovian dynamics.

IGME is derived based on the Nakajima-Zwanzig equation [339] and the Hummer-Szabo projection operator [340]. As demonstrated in our recent study [63], [265], [341], the dynamics of the state model can be characterized by the Generalized Master Equation, which relies on transition probability matrices (TPMs):

$$\dot{T(t)} = T(t)\dot{T}(0) - \int_{0}^{\min[t,\tau_k]} T(t-s)K(s)ds \tag{D.9}$$

Where $T(t)$ is the TPM at lag time t, $K(t)$ is the memory kernel at time t and τ_k is the memory kernel decay time ($K(t \geq \tau_k) = 0$). By applying the Taylor expansion and solving the GME equation in a self-consistent manner, we can derive the solution for long-time dynamics:

$$T(t \geq \tau_k) = A\hat{T}^t \tag{D.10}$$

Where A and $\hat{T}$ are two constant matrices, and $\hat{T}$ satisfies equation:

$$\ln \hat{T} = \dot{T}(0) - M_0 - \sum_{n=1}^{\infty} \frac{(-1)^n}{n!} (\ln \hat{T})^n M_n \tag{D.11}$$

Where $M_n(t)$ are the time integrals of memory kernels:

$$M_n(t) = \int_0^t K(s) s^n ds \tag{D.12}$$

When $t \to \infty$, $\lim_{t\to\infty} T(t) = t \ln \hat{T}$, so that $\hat{T}$ matrix describes the dynamics at the infinite long lag time, while A matrix represents the contribution from fast dynamics. Additionally, the zero-order term of the Equation D.11 can be employed to assess the integral of the memory kernels: $M_0 \approx \dot{T}(0) - \ln \hat{T}$, which also facilitates the calculation of the mean integral of memory kernels (MIKs): $\text{MIK}(\tau_k) = ||M_0||_F/N$, where $||M_0||_F$ is the Frobenius norm of M_0, and N is the number of macrostates.

To acquire long-time transition probabilities according to Equation D.10, it is necessary to estimate the matrices A and $\hat{T}$ using the simulation trajectories. In detail, we employed the logarithm of multiple TPMs at various lag times ($T(\tau_k), T(\tau_k+\Delta t), T(\tau_k+2\Delta t), ..., T(\tau_k+n\Delta t)$) to do least squared fitting to generate each element of A and $\hat{T}$ matrices. For the single tetra-nucleosome systems with NRL = 167, 172, and 177, we selected $\tau_k = 2.5 \times 10^5$ steps and $n\Delta t = 3.5 \times 10^5$ steps as the fitting range, guided by the convergence of the MIKs. For the NRL = 167 tetra-nucleosome system within the sea-of-nucleosomes environment, owing to its condensed nature, we utilized $\tau_k = 1 \times 10^6$ steps and $n\Delta t = 1.5 \times 10^6$ steps for fitting the IGME model. As depicted in Figure D.18-D.21, when using the transition probabilities directly derived from the raw MD data as reference and comparing with the MSM constructed at the lag time $\tau_k + n\Delta t$, it becomes evident that the IGME model exhibits significantly improved accuracy in predicting long-time dynamics. Since the $\hat{T}$ matrix represents the system's dynamical behavior at an infinitely long lag time, we subsequently used the $\hat{T}$ matrix to compute the stationary populations of macrostates and the MFPTs between every pair of macrostates for all four systems, as shown in Figure D.18-D.21.

D.3 Sequences

The DNA sequences used for all the systems are provided. Here only the nucleosomal repeat length (NRL) sequence of the first ssDNA is provided. If the NRL sequence is s and there are n_{nucl} nucleosomes, then the full sequence of the first ssDNA is $s \times (n_{\text{nucl}} - 1) + s[: 147]$ (here $s[: 147]$ means the slice of the first 147 letters in s, and this segment belongs to the nucleosome), and there is no additional DNA on either end of the tetra-nucleosome or single nucleosome. The second ssDNA sequence is Watson-Crick paired.

D.3.1 NRL 167 DNA Sequence

ACAGGATGTAACCTGCAGATACTACCAAAAGTGTATTTGGAAACTGCTCCATCAAAAGGCATGTTCAGCTGGATTCCAGCTGAACATGCCTTTTGATGGAGCAGTTTCCAAATACACTTTTGGTAGTATCTGCAGGTGATTCTCCAGGGCGGCCAGTACTTACATGC

D.3.2 NRL 172 DNA Sequence

ACAGGATGTAACCTGCAGATACTACCAAAAGTGTATTTGGAAACTGCTCCATCAAAAGGCATGTTCAGCTGGATTCCAGCTGAACATGCCTTTTGATGGAGCAGTTTCCAAATACACTTTTGGTAGTATCTGCAGGTGATTCTCCAGGGCGGCCAGTACTTACATGCGGCGG

D.3.3 NRL 177 DNA Sequence

ACAGGATGTAACCTGCAGATACTACCAAAAGTGTATTTGGAAACTGCTCCATCAA
AAGGCATGTTCAGCTGGATTCCAGCTGAACATGCCTTTTGATGGAGCAGTTTCCA
AATACACTTTTGGTAGTATCTGCAGGTGATTCTCCAGGGCGGCCAGTACTTACAT
GCGGCGGCCAGT

D.3.4 Single nucleosome DNA sequence

ACAGGATGTAACCTGCAGATACTACCAAAAGTGTATTTGGAAACTGCTCCATCAA
AAGGCATGTTCAGCTGGATTCCAGCTGAACATGCCTTTTGATGGAGCAGTTTCCA
AATACACTTTTGGTAGTATCTGCAGGTGATTCTCCAG

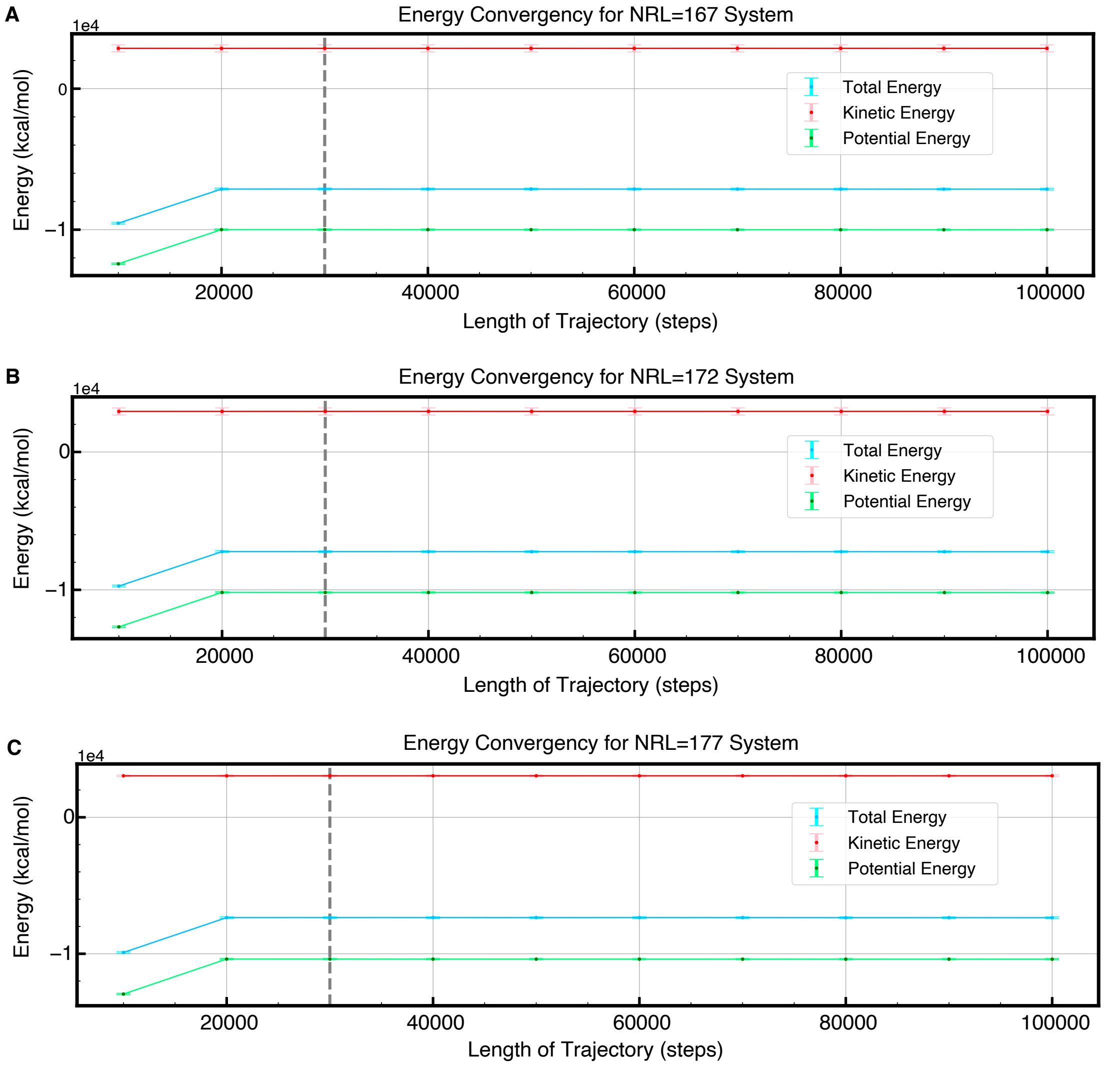


Figure D.1: Validate equilibrium of single tetra-nucleosome simulations. (A)-(C), The energy profiles for NRL = 167, 172, and 177 single tetra-nucleosome unbiased simulations, respectively. Starting from restrained MD final snapshots, the potential energy reaches equilibrium in about 20,000 steps. The energies are averaged over all the trajectories, and error bars are manifested. The first 30,000 steps (left side of the gray dashed line) were removed from every trajectory to ensure that the analyzed data were at equilibrium. The samples collected after the initial 30,000 steps were well-equilibrated and used for analysis.

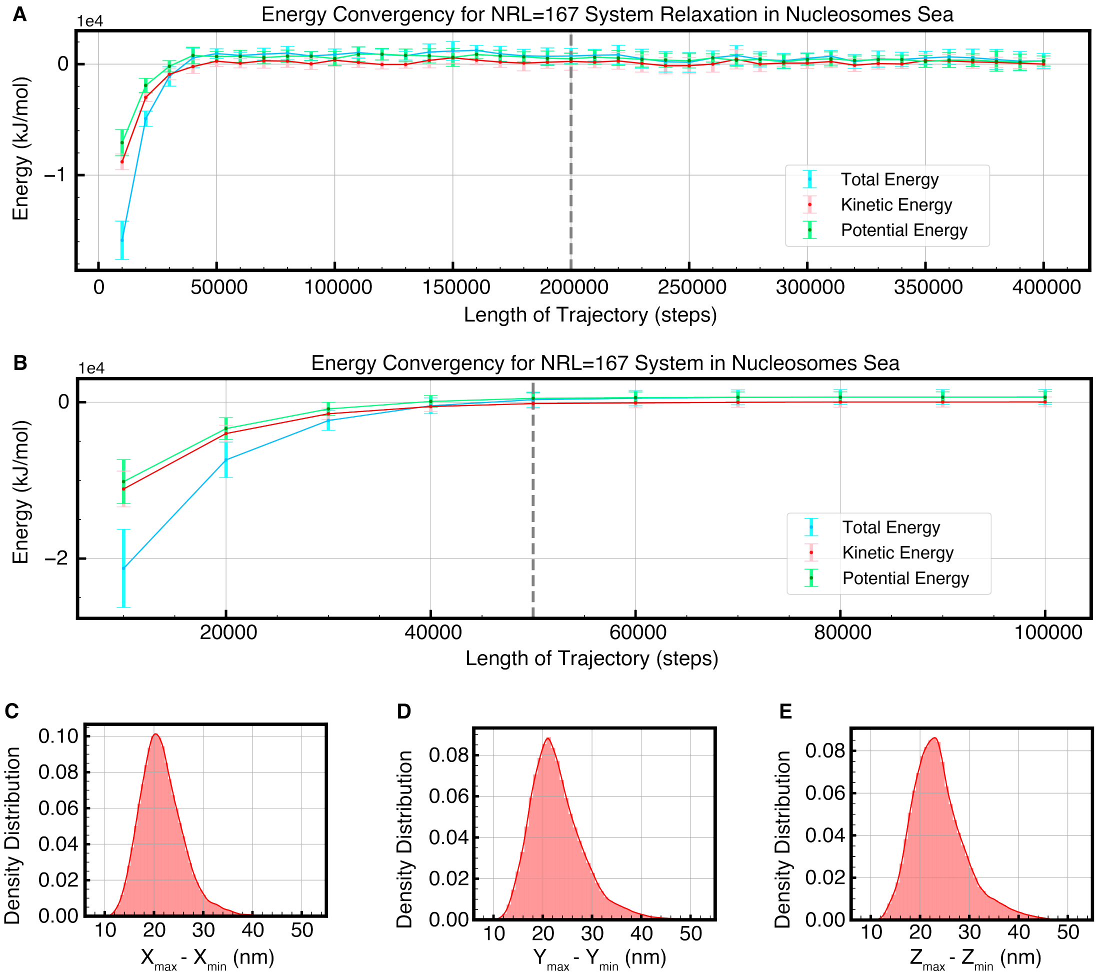


Figure D.2: Validate the energy convergence and the simulation box is large enough for the sea-of-nucleosome simulation. (A) The mean energy profile for relaxing the sea-of-nucleosome simulations. Due to the large constant energy contributed by the rigid bodies, the energy of each trajectory is shifted by removing the value at the final snapshot. Only the first 400,000 steps are shown. (B) The mean energy profile of the production run. Again, the energy of each trajectory is shifted by the value at the final snapshot, and only the first 100,000 steps are shown. The dashed line indicates 50,000 steps, and trajectories before 50,000 steps do not undergo analyses. (C)-(E) The size distributions of the tetra-nucleosome along the x, y, and z directions, respectively. This indicates that the cubic box of length 55 nm is large enough for the tetra-nucleosome to avoid touching its own periodic image.

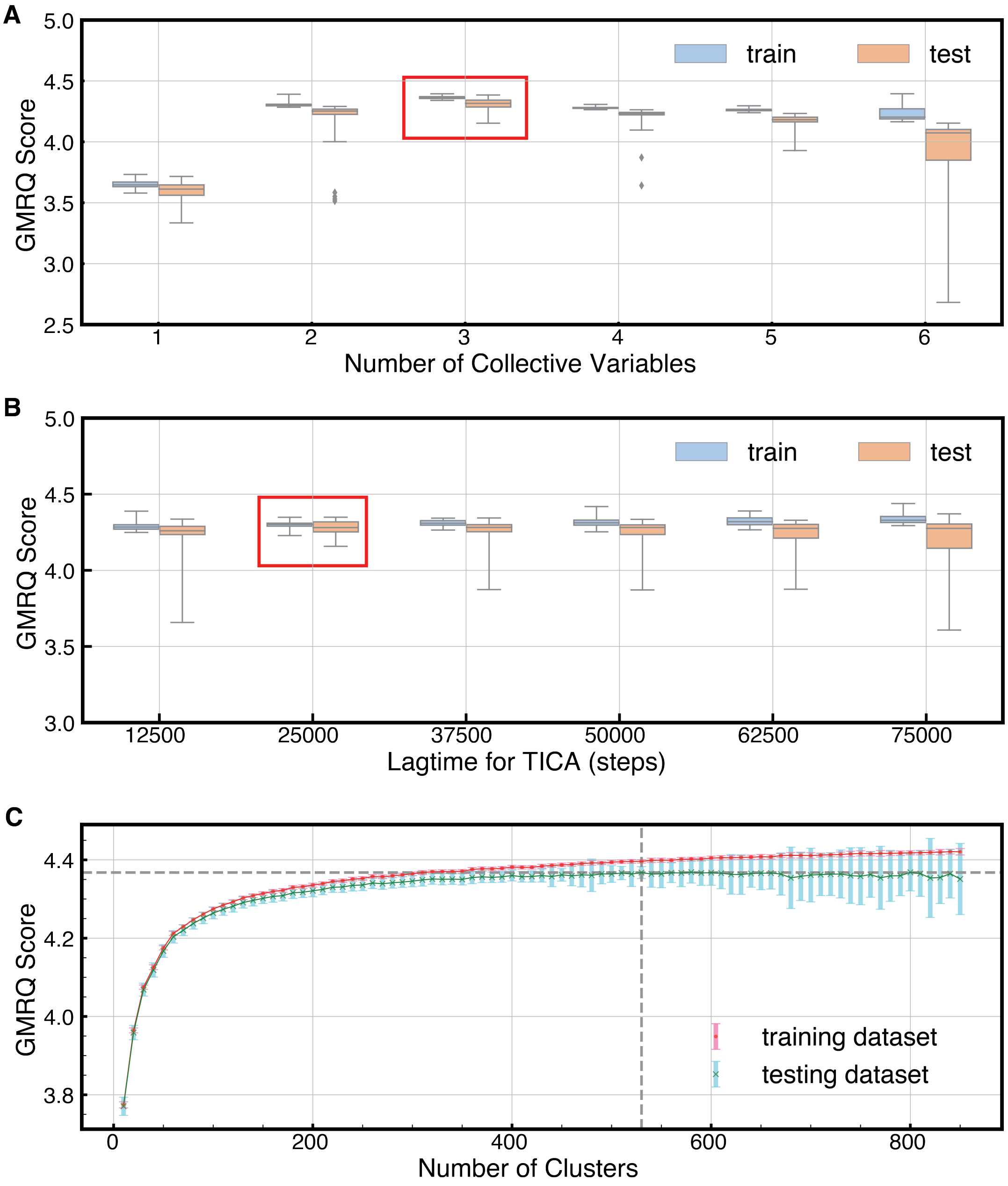


Figure D.3: To construct the MSM for the NRL = 167 system, optimize the hyperparameters for (A) the number of collective variables, (B) tICA lag time, and (C) the number of microstate clusters through cross-validation using GMRQ scores. During all GMRQ cross-validations, the dataset is divided into four subsets. Among these, three subsets are allocated for training, while one subset is reserved for validation. The MSM lag time is set to 0.4 million steps. For each hyperparameter, the cross-validation process is repeated ten times, employing different random seeds.

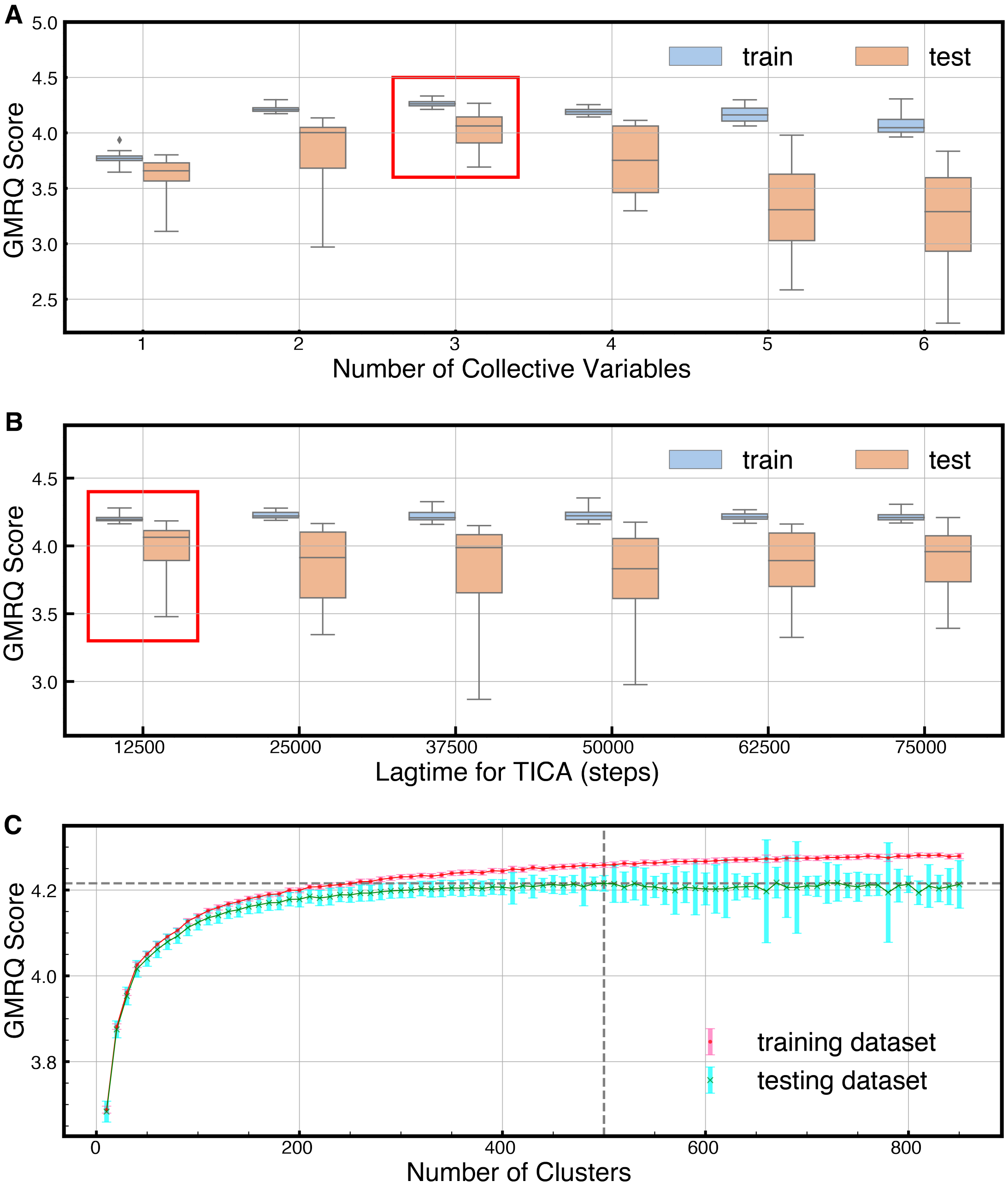


Figure D.4: To construct the MSM for the NRL=172 system, optimize the hyperparameters for (A) the number of collective variables, (B) tICA lag time, and (C) the number of microstate clusters through cross-validation using GMRQ scores. During all GMRQ cross-validations, the dataset is divided into four subsets. Among these, three subsets are allocated for training, while one subset is reserved for validation. The MSM lag time is fixed at 0.4 million steps. For each hyperparameter, the cross-validation process is repeated ten times, employing different random seeds.

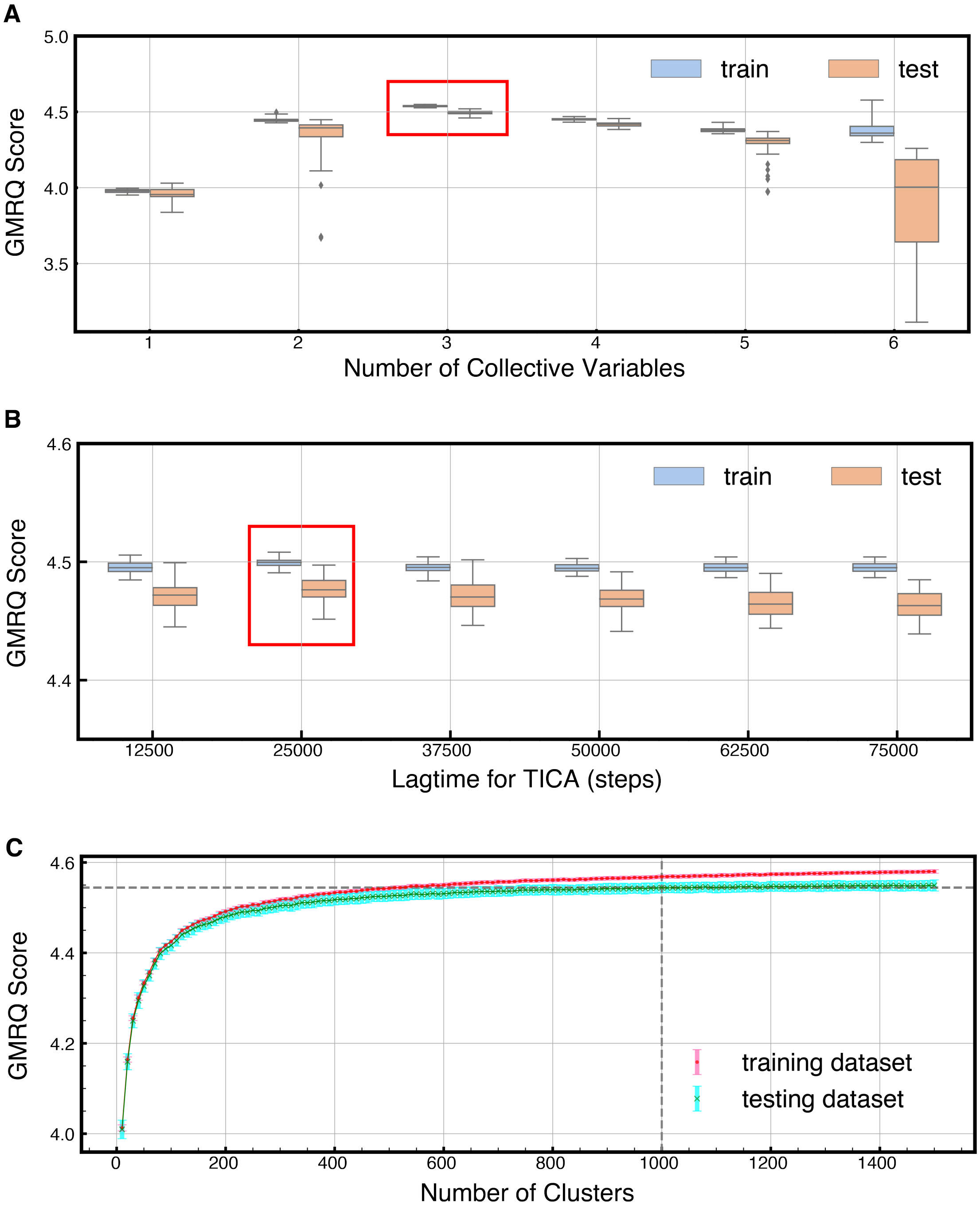


Figure D.5: To construct the MSM for the NRL=177 system, optimize the hyperparameters for (A) the number of collective variables, (B) tICA lag time, and (C) the number of microstate clusters through cross-validation using GMRQ scores. During all GMRQ cross-validations, the dataset is divided into four subsets. Among these, three subsets are allocated for training, while one subset is reserved for validation. The MSM lag time is fixed at 0.4 million steps. For each hyperparameter, the cross-validation process is repeated ten times, employing different random seeds.

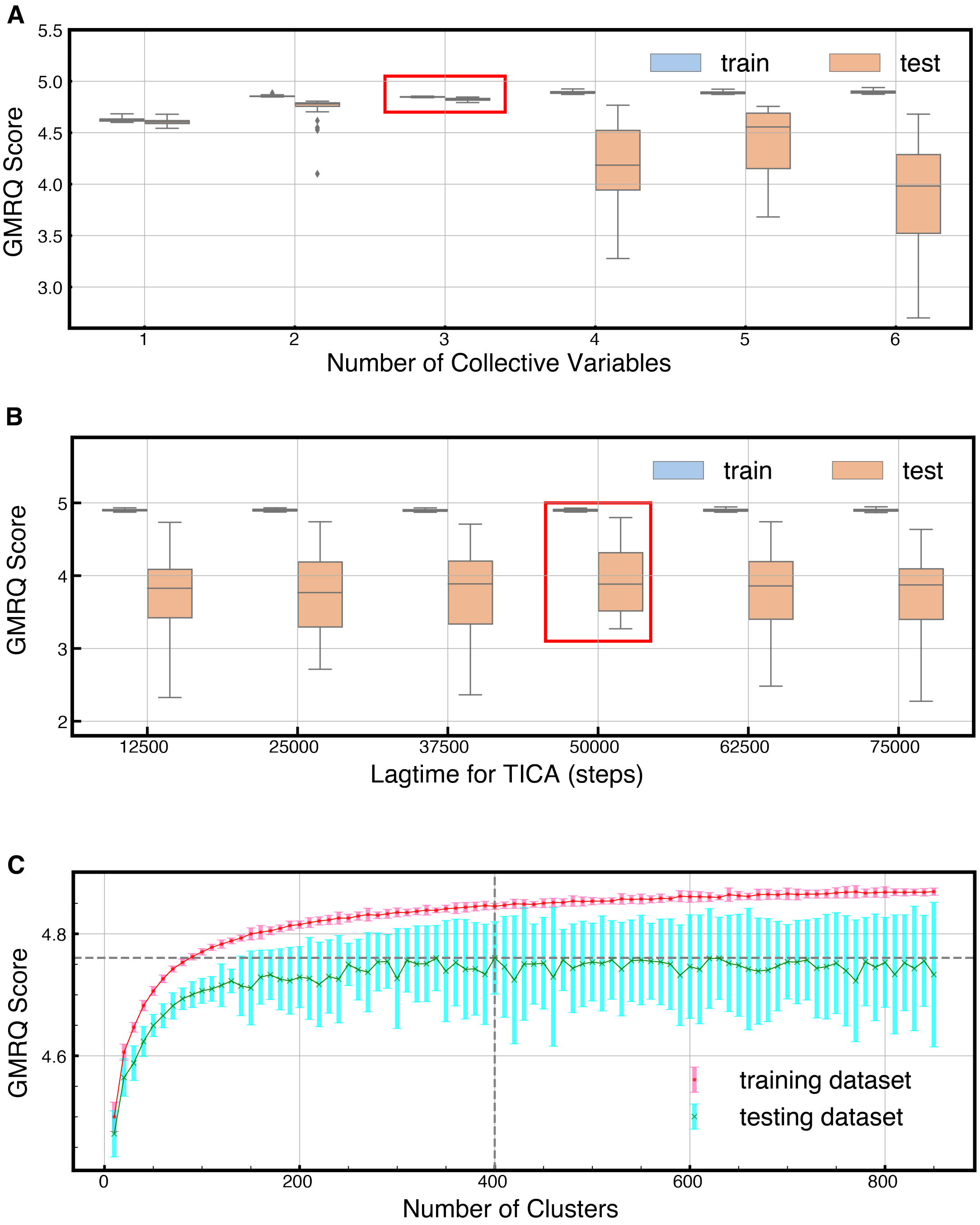


Figure D.6: To construct the MSM for the NRL=167 tetra-nucleosome in the sea-of-nucleosome system, optimize the hyperparameters for (A) the number of collective variables, (B) tICA lag time, and (C) the number of microstate clusters through cross-validation using GMRQ scores. During all GMRQ cross-validations, the dataset is divided into four subsets. Among these, three subsets are allocated for training, while one subset is reserved for validation. The MSM lag time is fixed at 1.5 million steps. For each hyperparameter, the cross-validation process is repeated ten times, employing different random seeds.

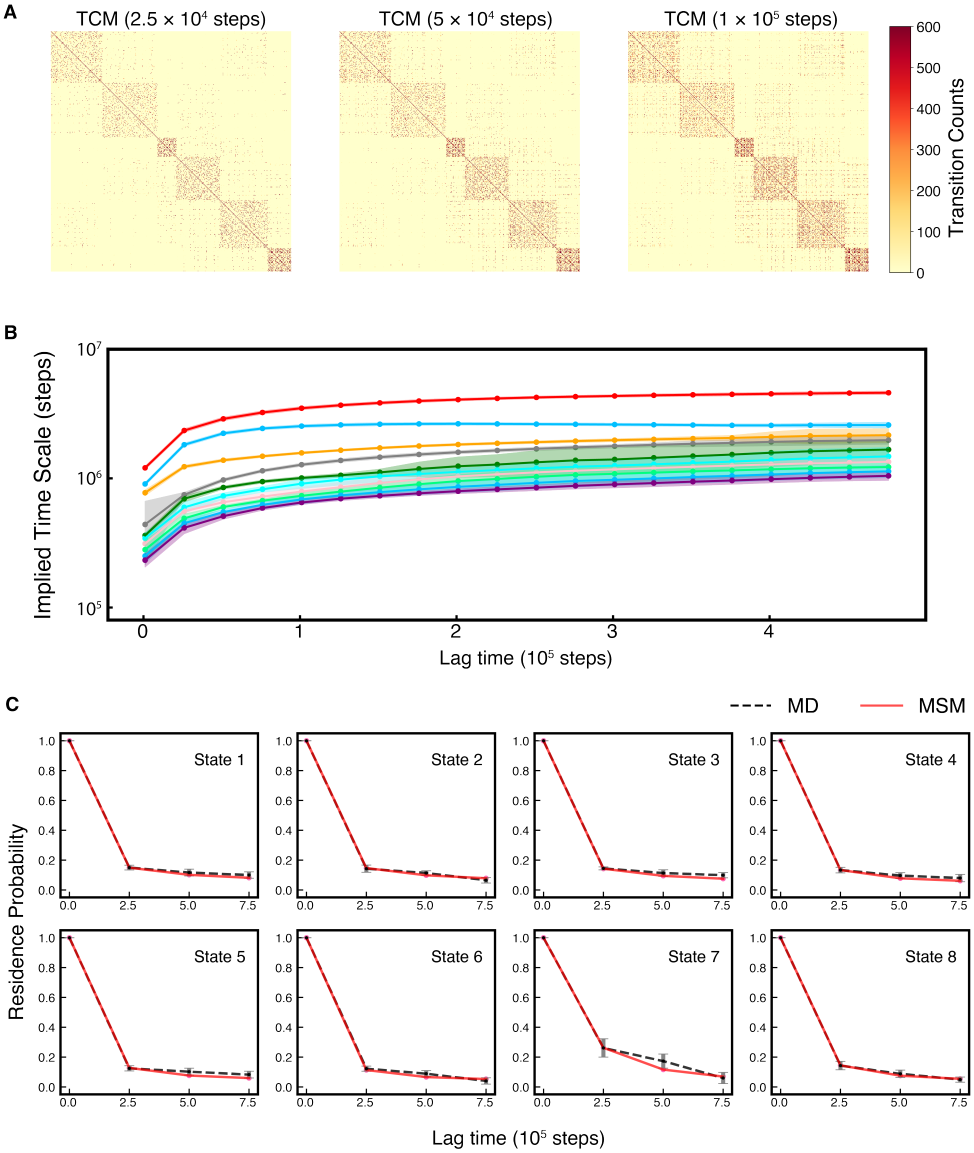


Figure D.7: Validate the microstate MSM for the NRL = 167 system. (A) Transition Count Matrices (TCMs) calculated at various lag times for the 530-microstate MSM. The symmetrical nature of the TCMs substantiates the notion that the folding dynamics satisfies detailed balance. The TCMs have been reorganized in accordance with the PCCA+ lumping results. (B) Implied Time Scales (ITS) plot for the 530-microstate MSM. (C) Chapman-Kolmogorov (CK) test for the 8 most populated microstates utilizing a Markovian lag time of 0.25 million steps. The agreement between the residence probabilities predicted by the MSM and those directly obtained from the MD simulation serves as the validation for the MSM. The error bars in the ITS plot and the CK test plots are calculated by bootstrapping the MD trajectories 20 times with replacements.

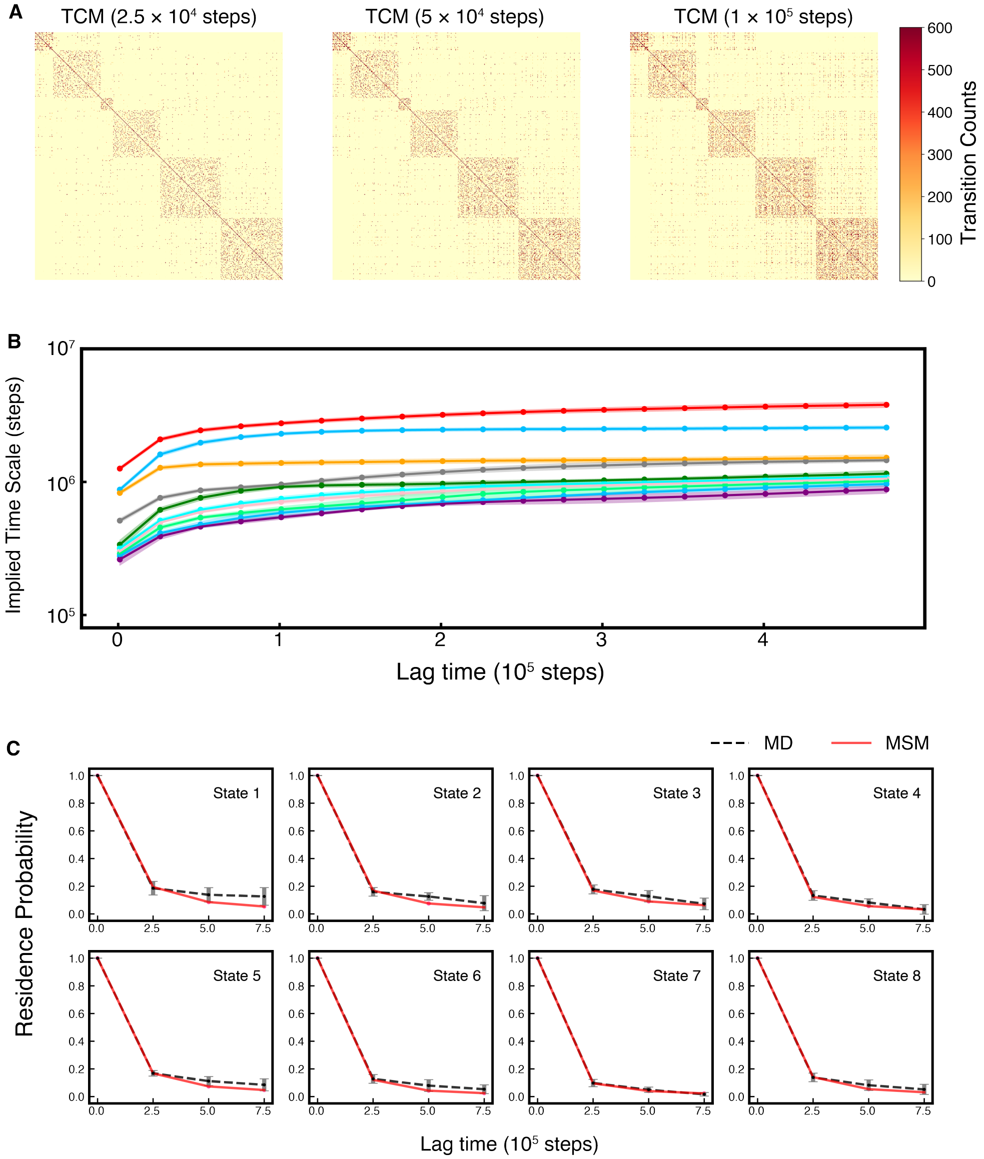


Figure D.8: Validate the microstate MSM for the NRL = 172 system. (A) Transition Count Matrices (TCMs) calculated at various lag times for 500 microstates MSM. The symmetrical nature of the TCMs substantiates the notion that the folding dynamics satisfy detailed balance. The TCMs have been reorganized in accordance with the PCCA+ lumping results. (B) Implied Time Scales (ITS) plot for 500 microstates MSM. (C) Chapman-Kolmogorov (CK) test for the 8 most populated microstates utilizing a Markovian lag time of 0.25 million steps. The agreement between residence probabilites predicted by the MSM and those directly obtained from MD simulation serves as the validation for the MSM. The error bars in the ITS plot and the CK test plots are calculated by bootstrapping the MD trajectories 20 times with replacements.

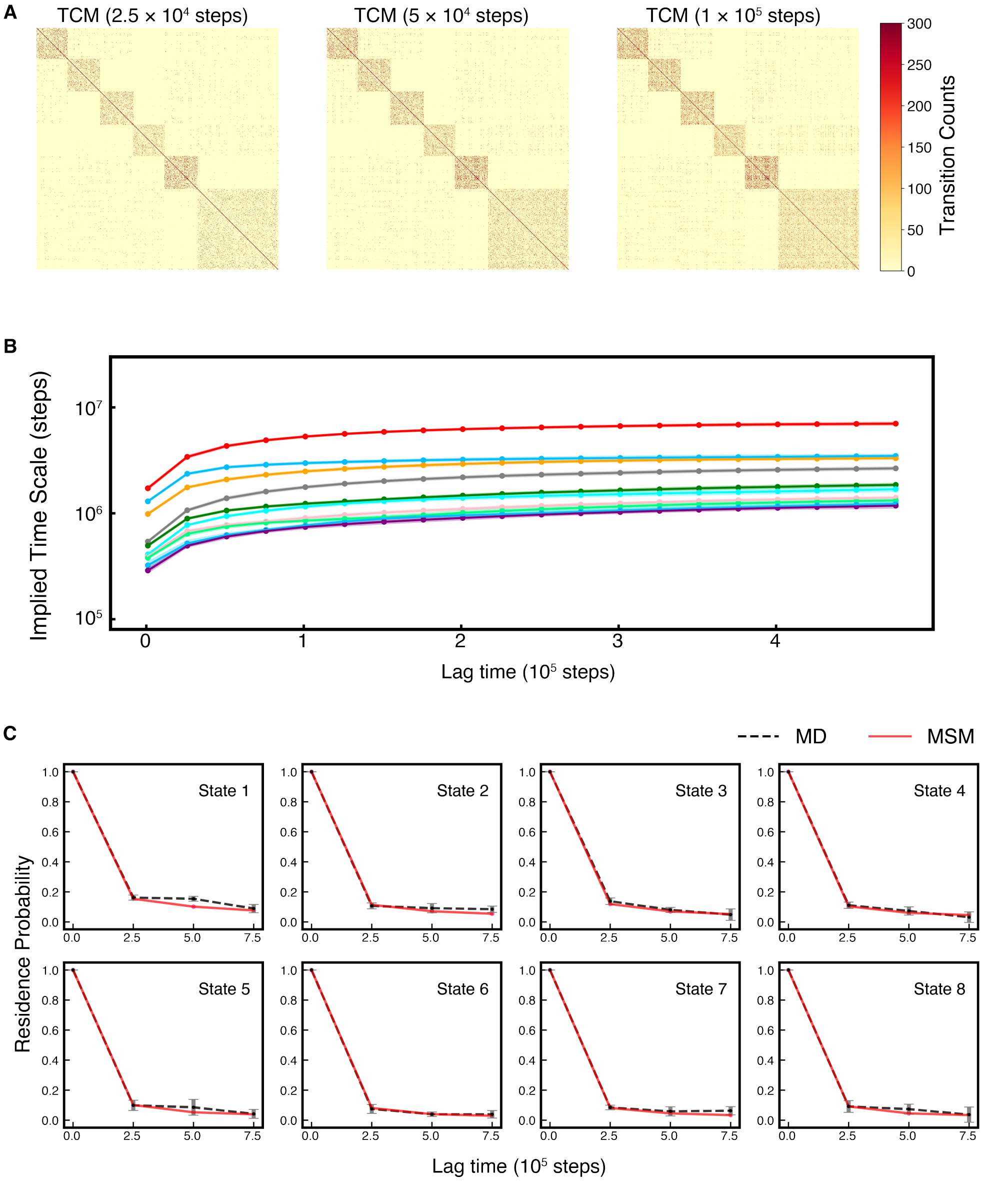


Figure D.9: Validate the microstate MSM for the NRL = 177 system. (A) Transition Count Matrices (TCMs) calculated at various lag times for 1000 microstates MSM. The symmetrical nature of the TCMs substantiates the notion that the folding dynamics satisfy detailed balance. The TCMs have been reorganized in accordance with the PCCA+ lumping results. (B) Implied Time Scales (ITS) plot for 1000 microstates MSM. (C) Chapman-Kolmogorov (CK) test for the 8 most populated microstates utilizing a Markovian lag time of 0.25 million steps. The agreement between residence probabilites predicted by the MSM and those directly obtained from MD simulation serves as the validation for the MSM. The error bars in the ITS plot and the CK test plots are calculated by bootstrapping the MD trajectories 20 times with replacements.

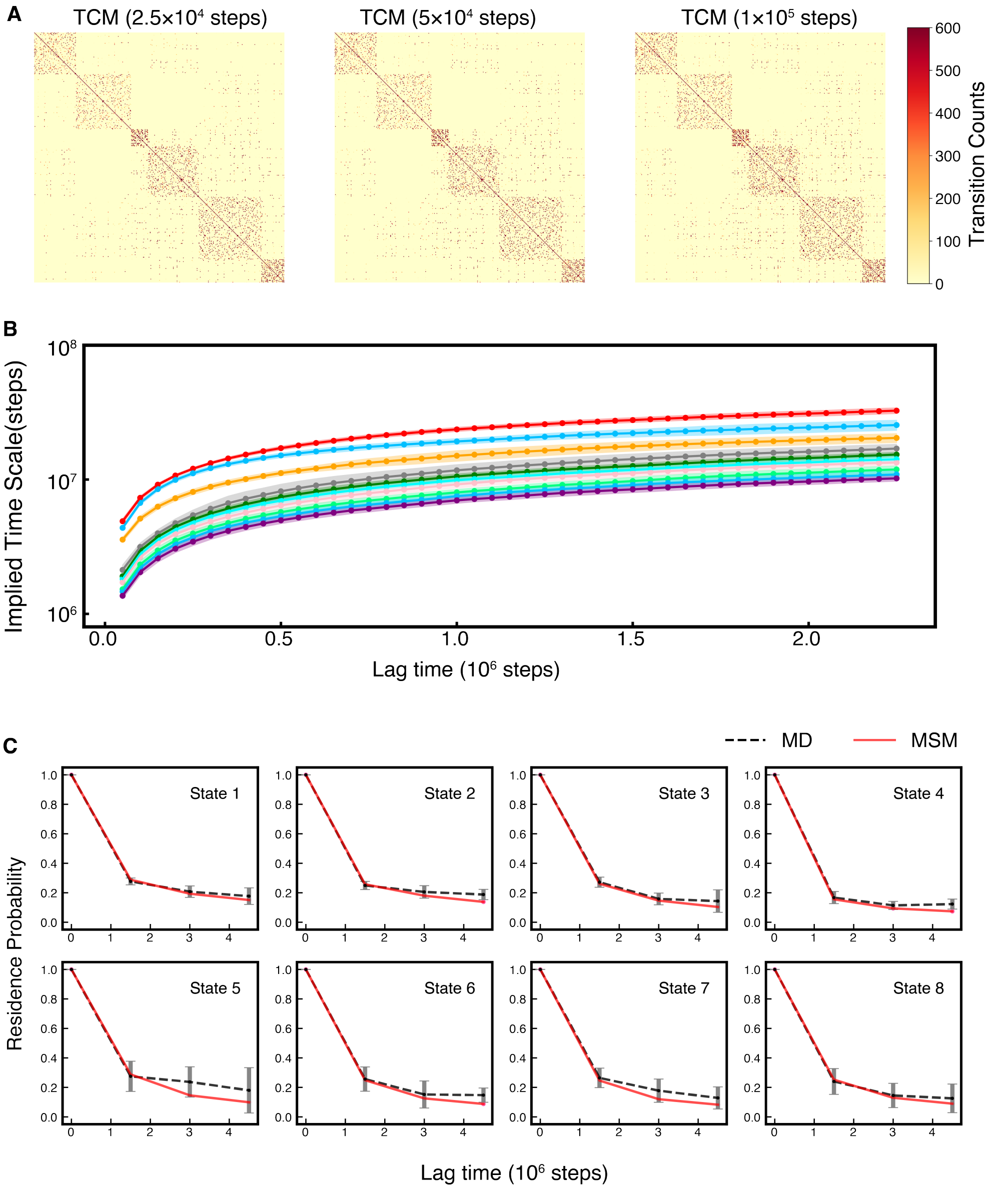


Figure D.10: Validate the microstate MSM for the NRL=167 tetra-nucleosome in the sea-of-nucleosome system. (A) Transition Count Matrices (TCMs) calculated at various lag times for 400 microstates MSM. The symmetrical nature of the TCMs substantiates the notion that the folding dynamics satisfy detailed balance. The TCMs have been reorganized in accordance with the PCCA+ lumping results. (B) Implied Time Scales (ITS) plot for 400 microstates MSM. (C) Chapman-Kolmogorov (CK) test for the 8 most populated microstates utilizing a Markovian lag time of 1.5 million steps. The agreement between residence probabilites predicted by the MSM and those directly obtained from MD simulation serves as the validation for the MSM. The error bars in the ITS plot and the CK test plots are calculated by bootstrapping the MD trajectories 20 times with replacements.

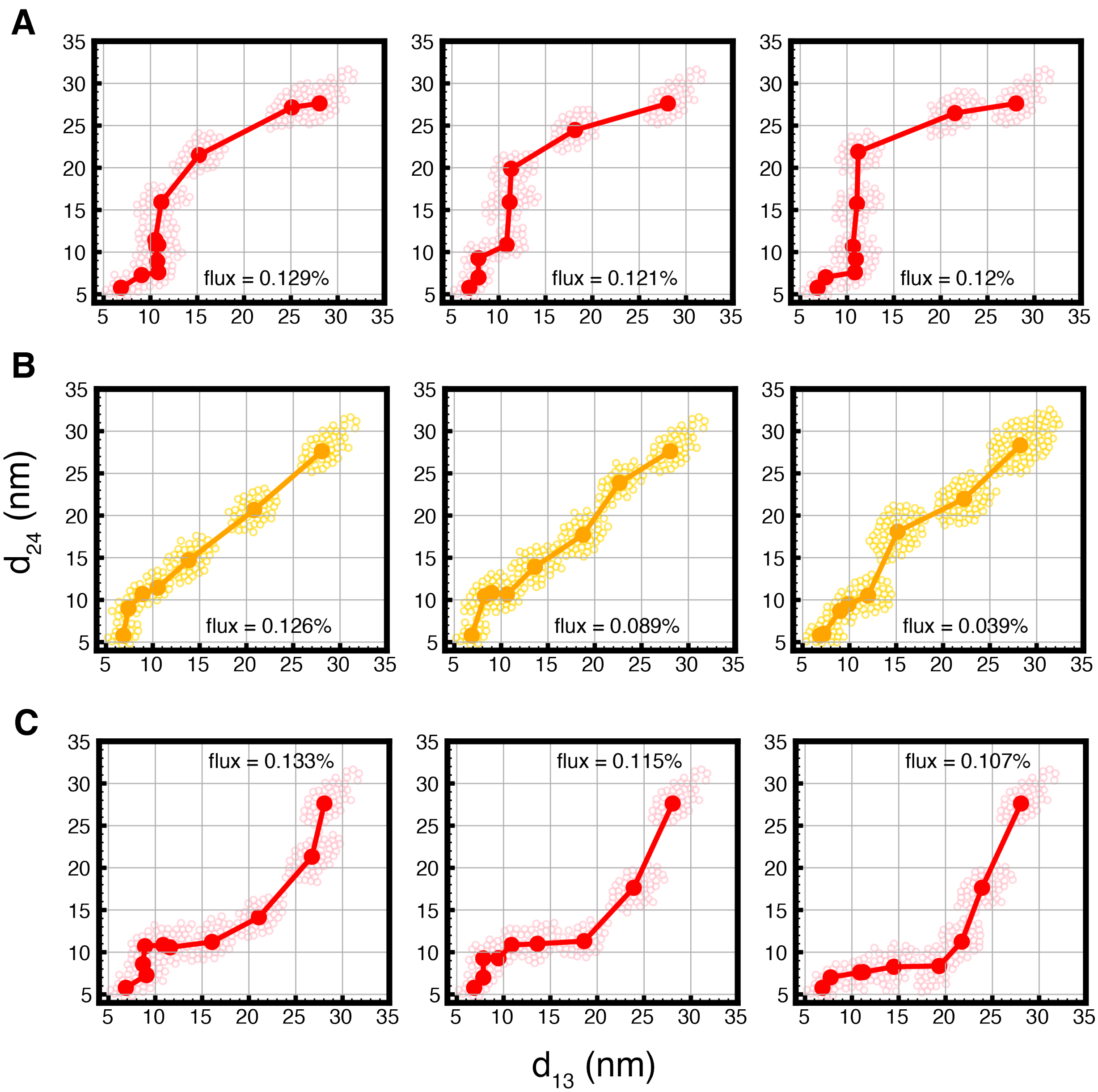


Figure D.11: Visualize the representative kinetic pathways exhibiting the highest fluxes within three distinct path channels for the NRL = 172 system. (A) Up-sequential channel, (B) concerted channel, and (C) down-sequential channel. The solid dots represent the centers of the microstates that the pathways traverse, while the dashed dots indicate the samples belonging to those microstates.

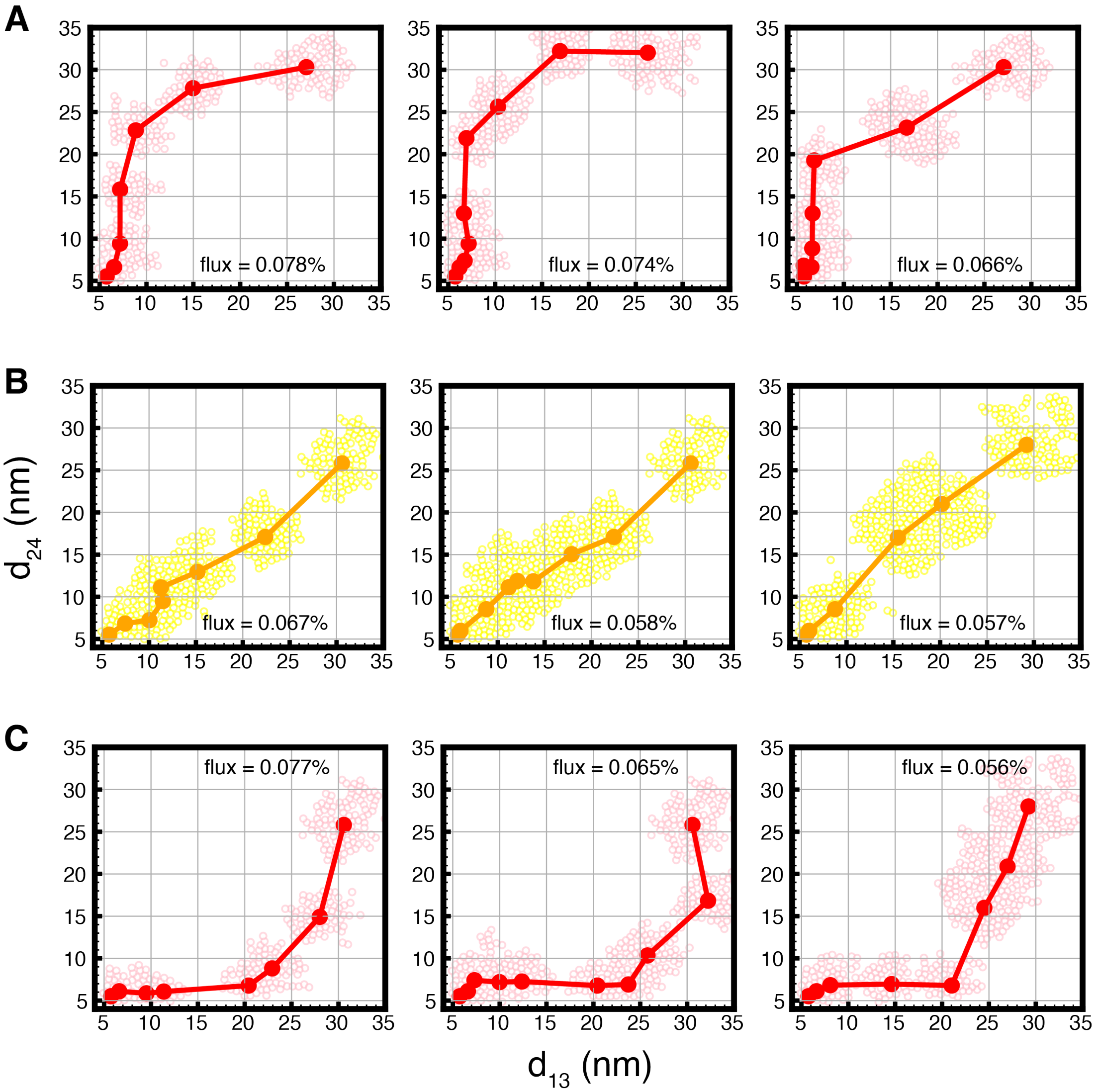


Figure D.12: Visualize the representative kinetic pathways exhibiting the highest fluxes within three distinct path channels for the NRL = 177 system. (A) Up-sequential channel, (B) concerted channel, and (C) down-sequential channel. The solid dots represent the centers of the microstates that the pathways traverse, while the dashed dots indicate the samples belonging to those microstates.

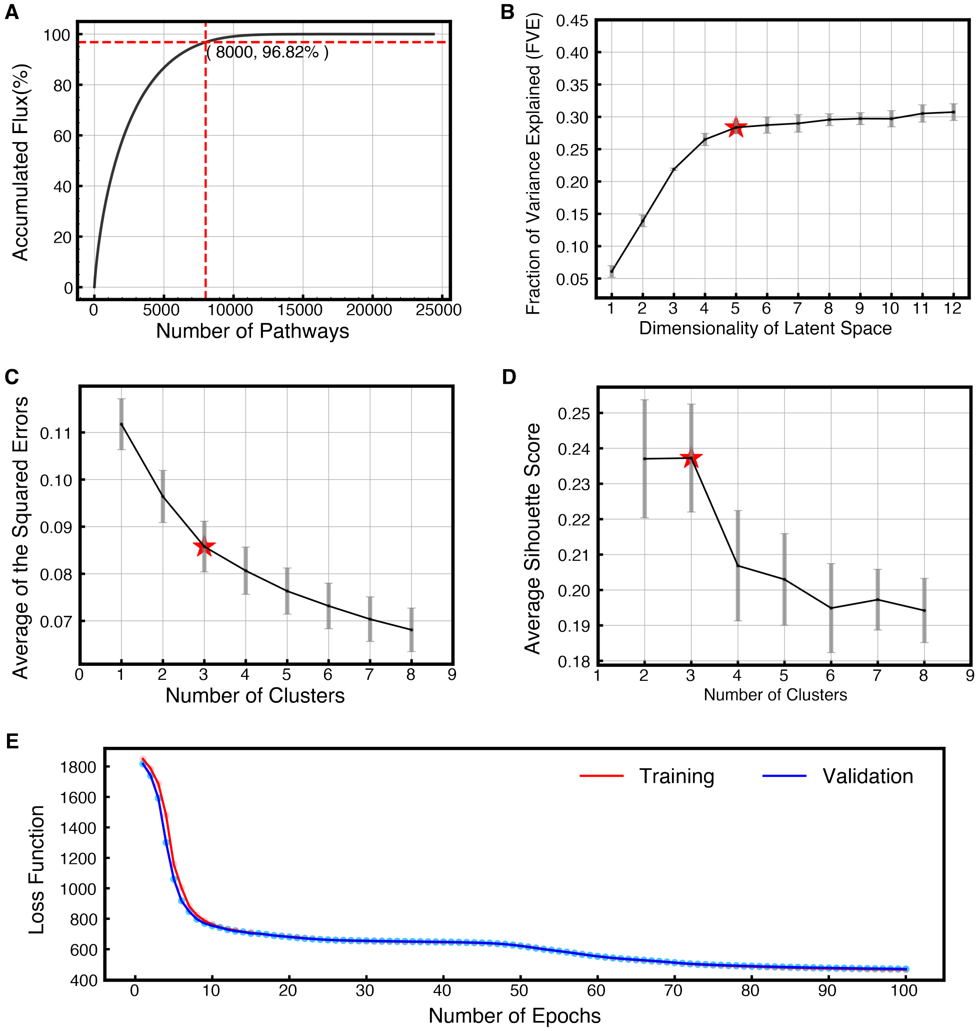


Figure D.13: Cluster the parallel kinetic pathways for the NRL = 167 system by LPC algorithm. (A) Accumulated flux as a function of the number of transition pathways (based on the 530-microstate MSM). Clustering is conducted using the top 8000 pathways, which collectively account for over 95% of the total flux. (B) Fraction of Variance Explained (FVE) as a function of the dimension of the latent space. (C)-(D) Both the average of squared errors and the average Silhouette score, plotted as functions of the number of clusters in the latent space, favor the categorization of 3 clusters. The error bars are estimated by training the VAE models for 10 times with different random seeds. (E) An illustrative example of the VAE training process.

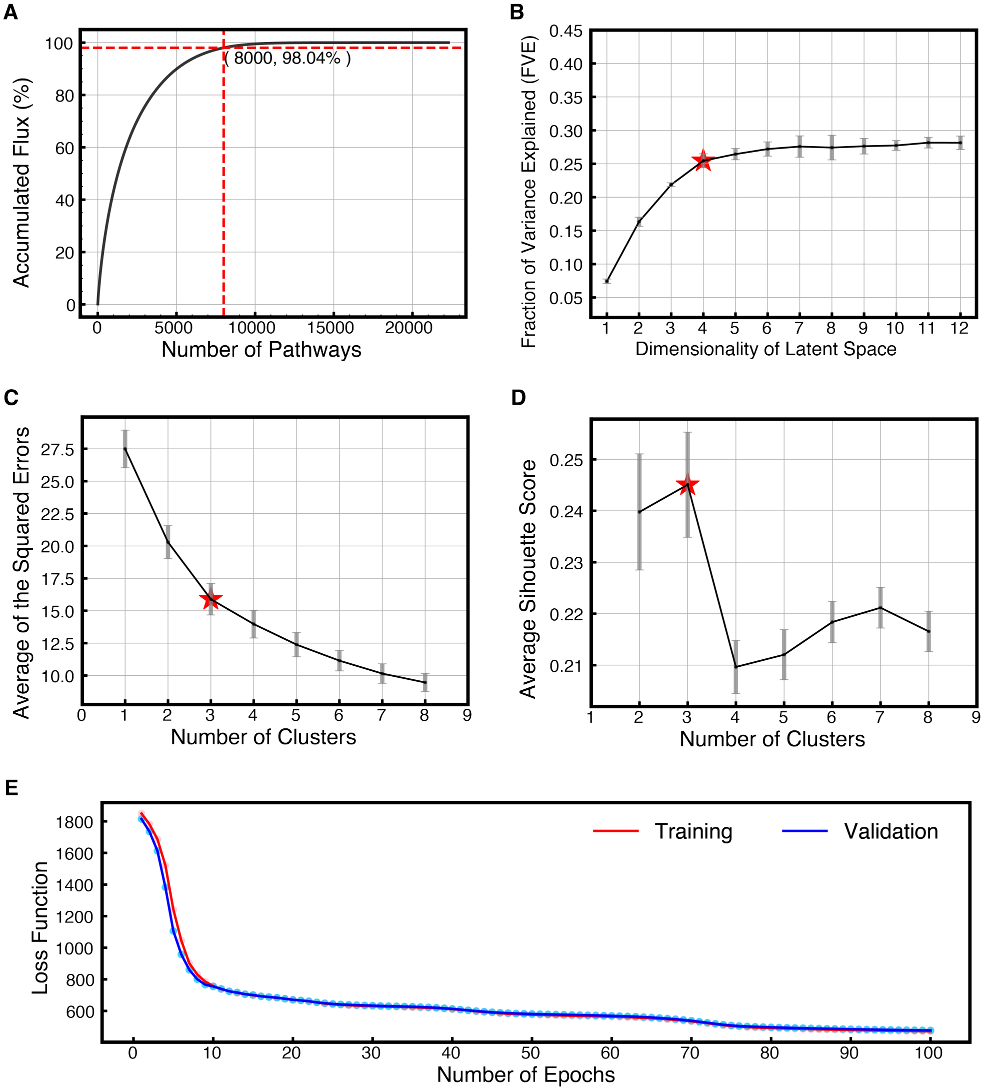


Figure D.14: Cluster the parallel kinetic pathways for the NRL = 172 system by LPC algorithm. (A) Accumulated flux as a function of the number of transition pathways (based on the 500-microstate MSM). Clustering is conducted using the top 8000 pathways, which collectively account for over 98% of the total flux. (B) Fraction of Variance Explained (FVE) as a function of the dimension of the latent space. (C)-(D) Both the average of squared errors and the average Silhouette score, plotted as functions of the number of clusters in the latent space, favor the categorization of 3 clusters. The error bars are estimated by training the VAE models for 10 times with different random seeds. (E) An illustrative example of the VAE training process.

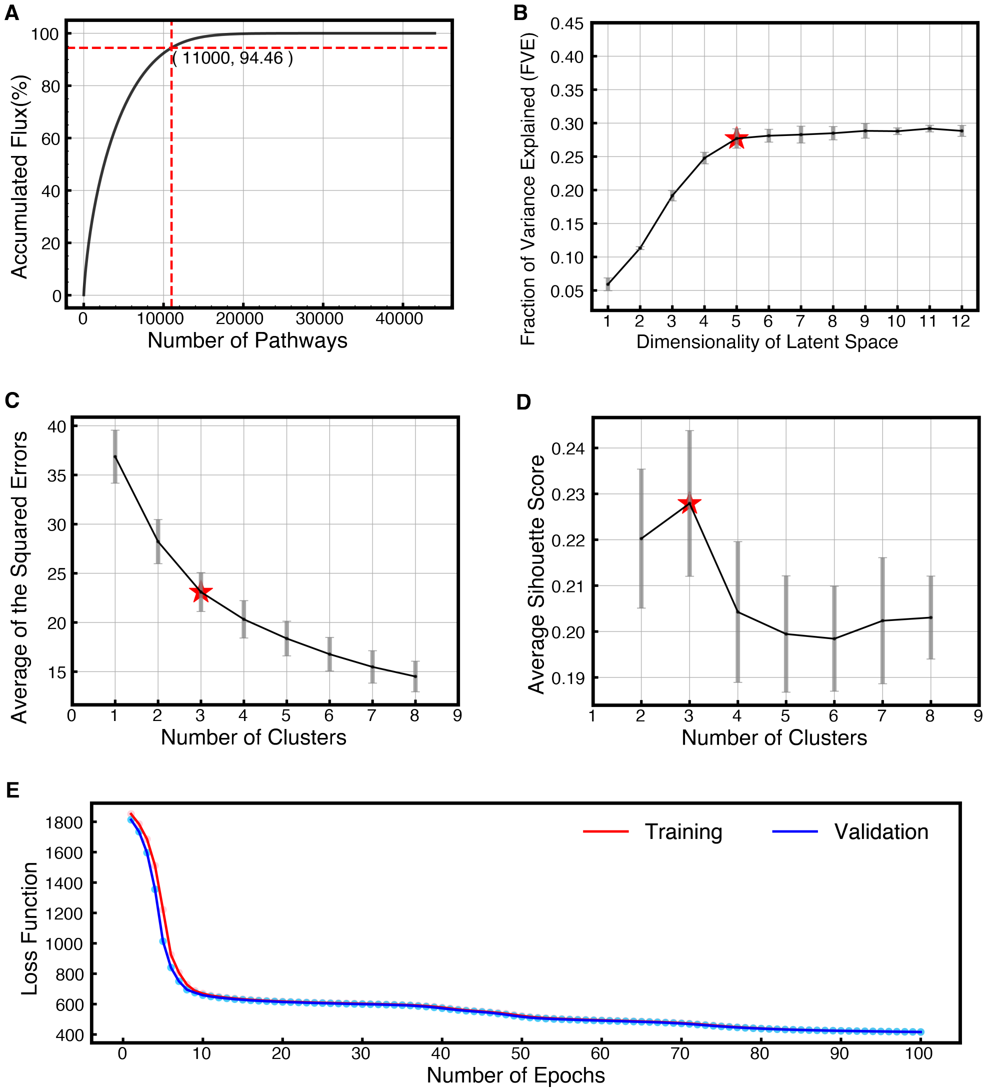


Figure D.15: Cluster the parallel kinetic pathways for the NRL = 177 system by LPC algorithm. (A) Accumulated flux as a function of the number of transition pathways (based on the 1000-microstate MSM). Clustering is conducted using the top 11,000 pathways, which collectively account for around 95% of the total flux. (B) Fraction of Variance Explained (FVE) as a function of the dimension of the latent space. (C)-(D) Both the average of squared errors and the average Silhouette score, plotted as functions of the number of clusters in the latent space, favor the categorization of 3 clusters. The error bars are estimated by training the VAE models for 10 times with different random seeds. (E) An illustrative example of the VAE training process.

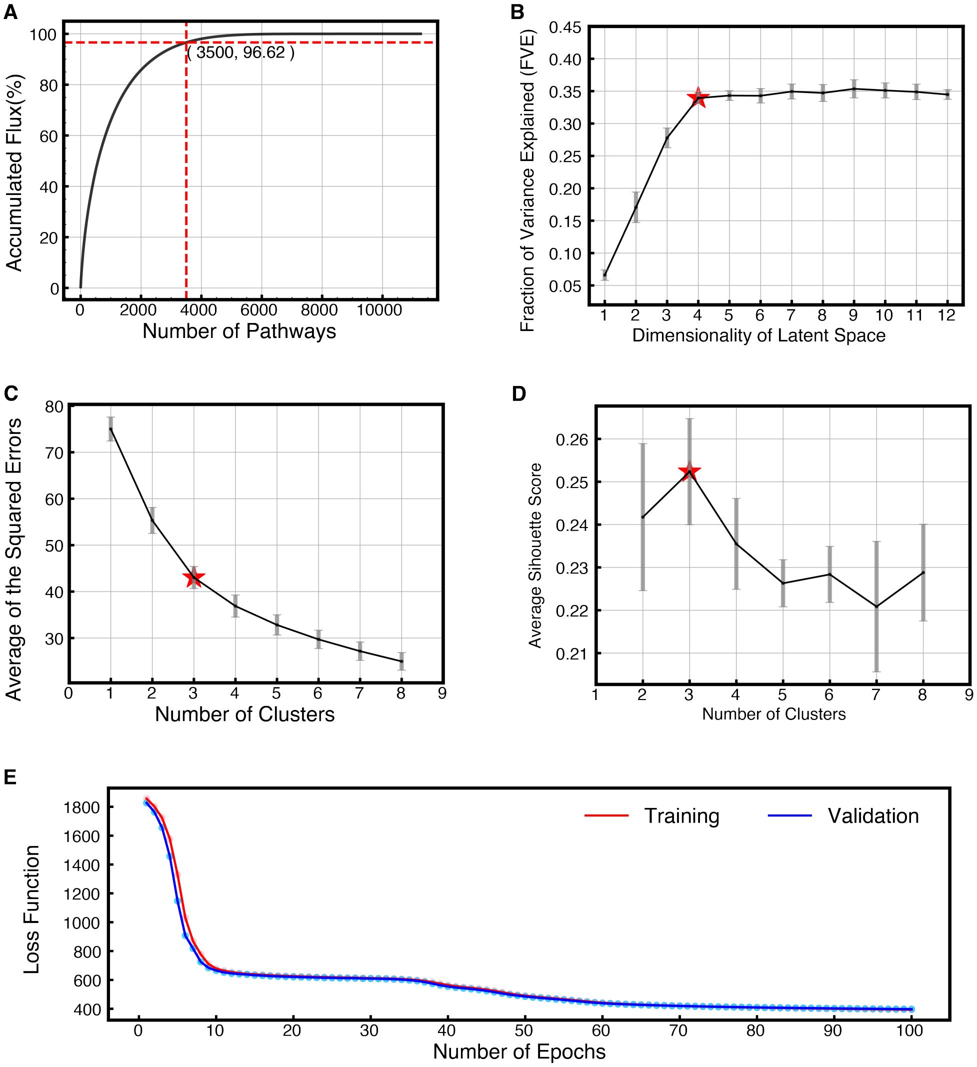


Figure D.16: Cluster the parallel kinetic pathways for the NRL=167 tetra-nucleosome in the sea-of-nucleosome system by LPC algorithm. (A) Accumulated flux as a function of the number of transition pathways (based on the 400-microstate MSM). Clustering is conducted using the top 3500 pathways, which collectively account for more than 95% of the total flux. (B) Fraction of Variance Explained (FVE) as a function of the dimension of the latent space. (C)-(D) Both the average of squared errors and the average Silhouette score, plotted as functions of the number of clusters in the latent space, favor the categorization of 3 clusters. The error bars are estimated by training the VAE models for 10 times with different random seeds. (E) An illustrative example of the VAE training process.

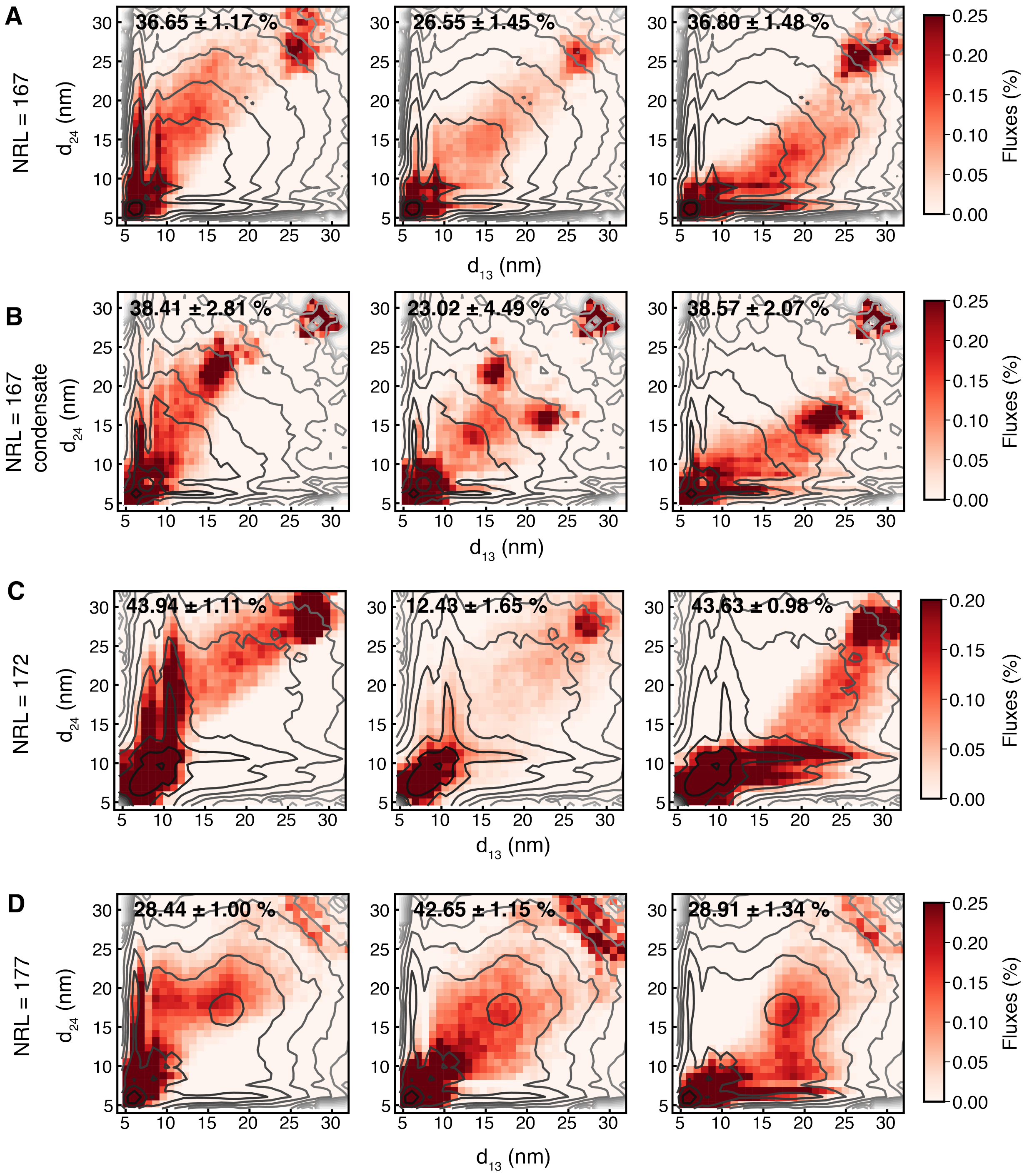
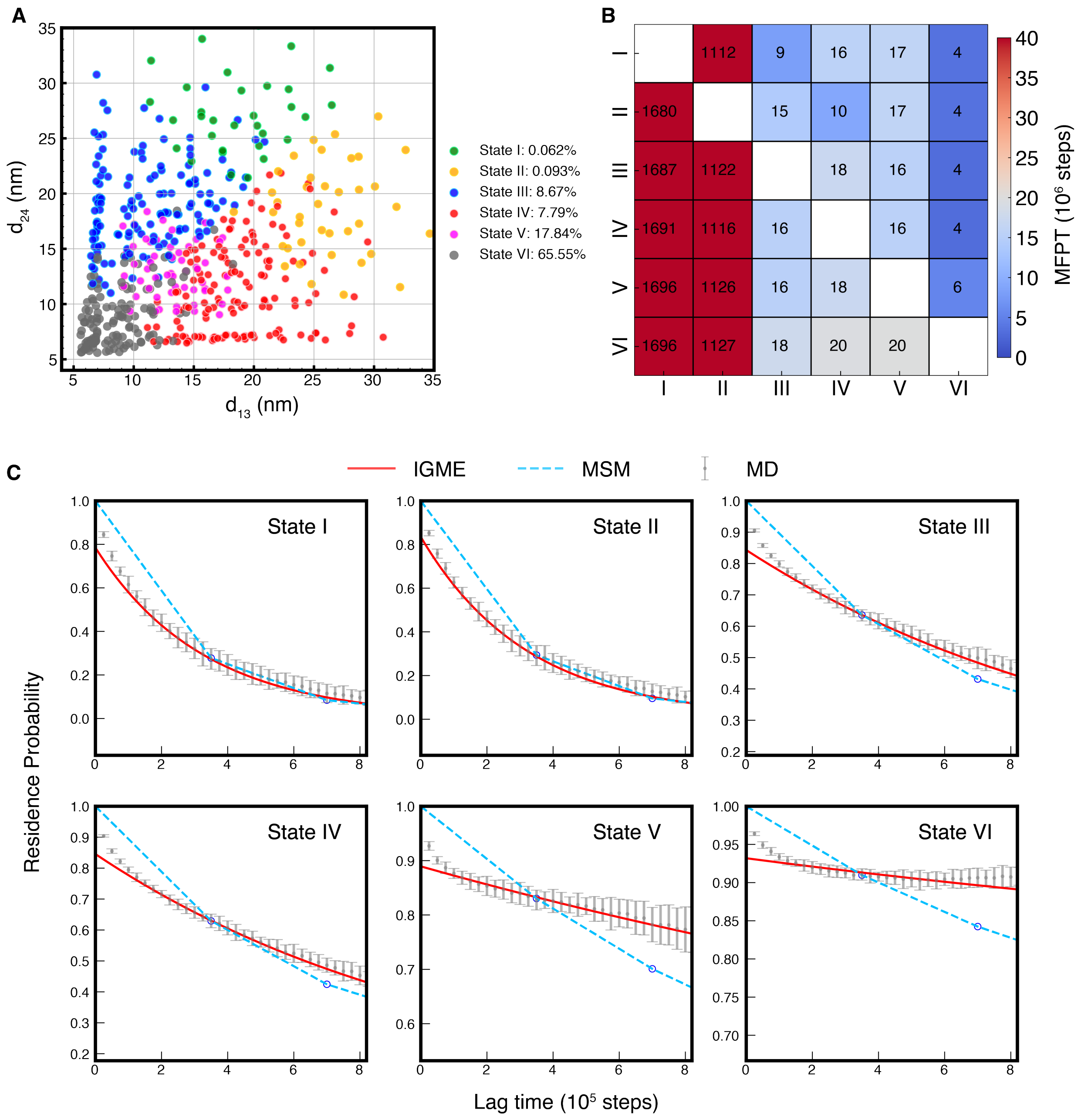


Figure D.17: (A)-(D) Heat maps of all the path channels for the NRL = 167 single, condensate, NRL = 172 single, and NRL = 177 single systems, respectively. The three columns are up-sequential, concerted, and down-sequential pathways. The flux percentages are labeled in each plot.



	I	II	III	IV	V	VI
I		1112	9	16	17	4
II	1680		15	10	17	4
III	1687	1122		18	16	4
IV	1691	1116	16		16	4
V	1696	1126	16	18		6
VI	1696	1127	18	20	20	

Figure D.18: Construct macrostate MSM and predict properties through IGME for the NRL = 167 system. (A) Distribution of macrostates on the d_{13} and d_{24} map along with their corresponding stationary populations. Each dot represents the center position of a microstate, with colors indicating macrostate assignments based on PCCA+. (B) Mean first passage times (MFPTs) of transitions between macrostates predicted by IGME. The numbers indicate the predicted MFPTs in unit 10^6 steps from row index states to column index states. (C) CK test for the 6-macrostate model using 3.5×10^5 steps MSM and IGME. The error bars are estimated by bootstrapping the MD trajectories 20 times.

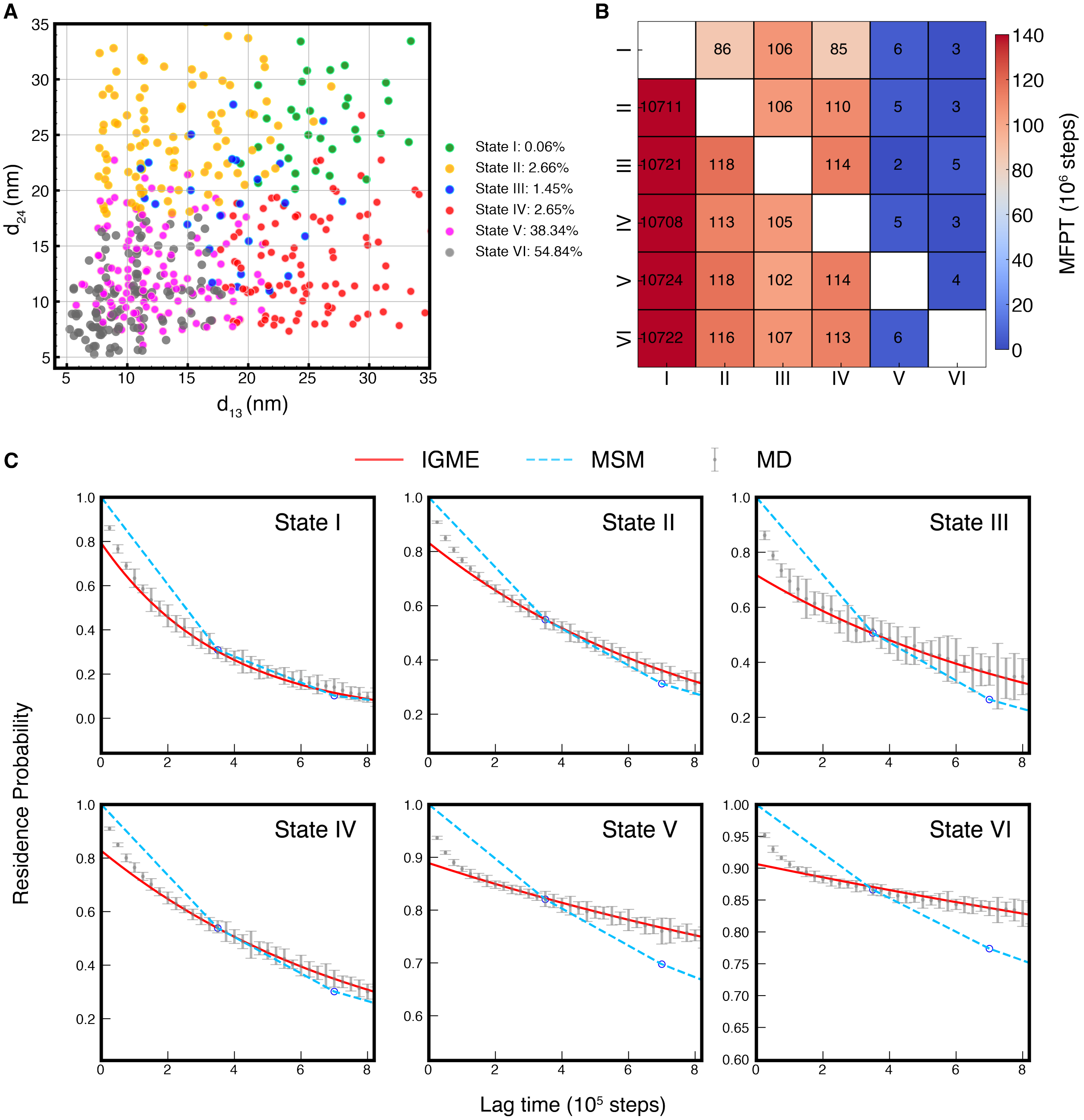
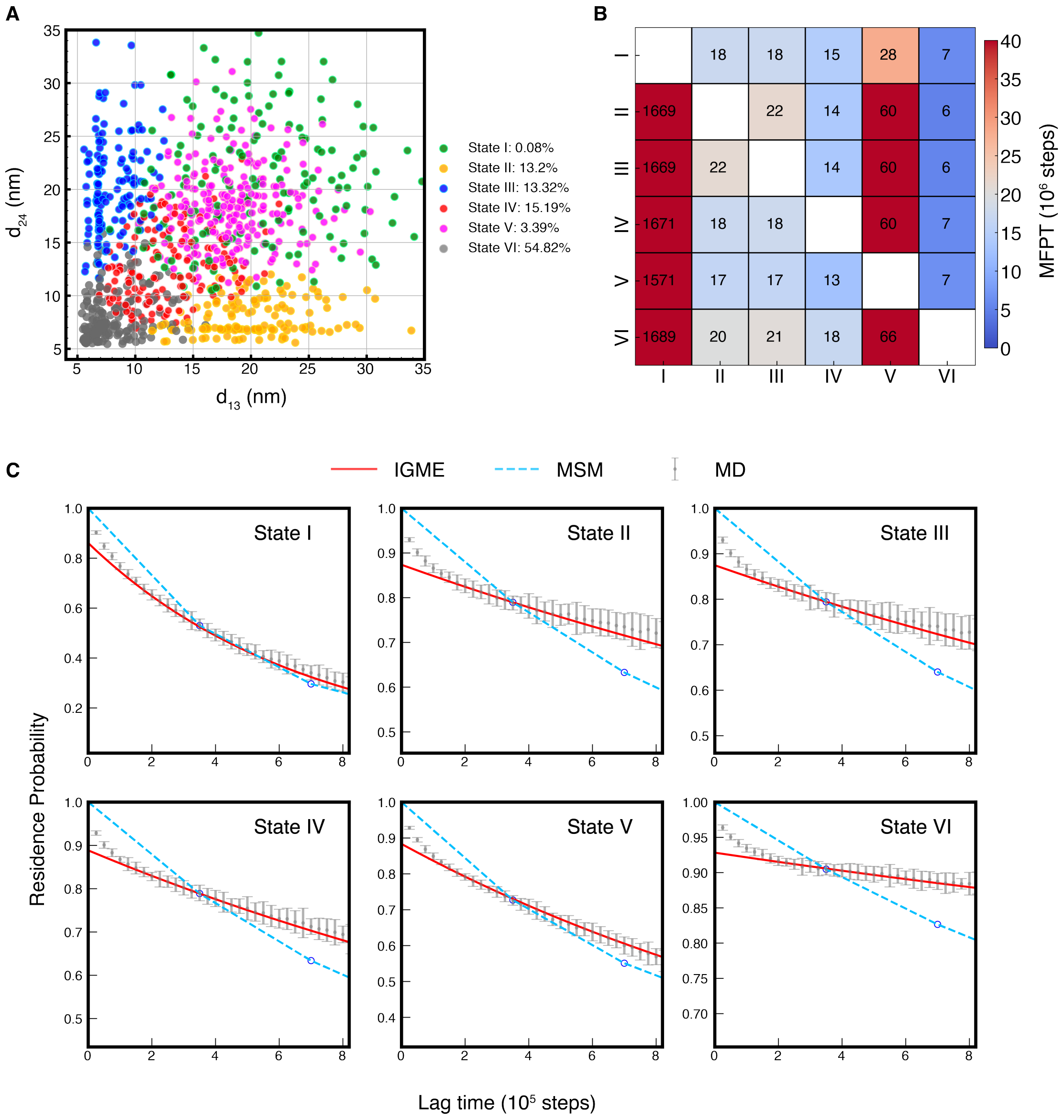
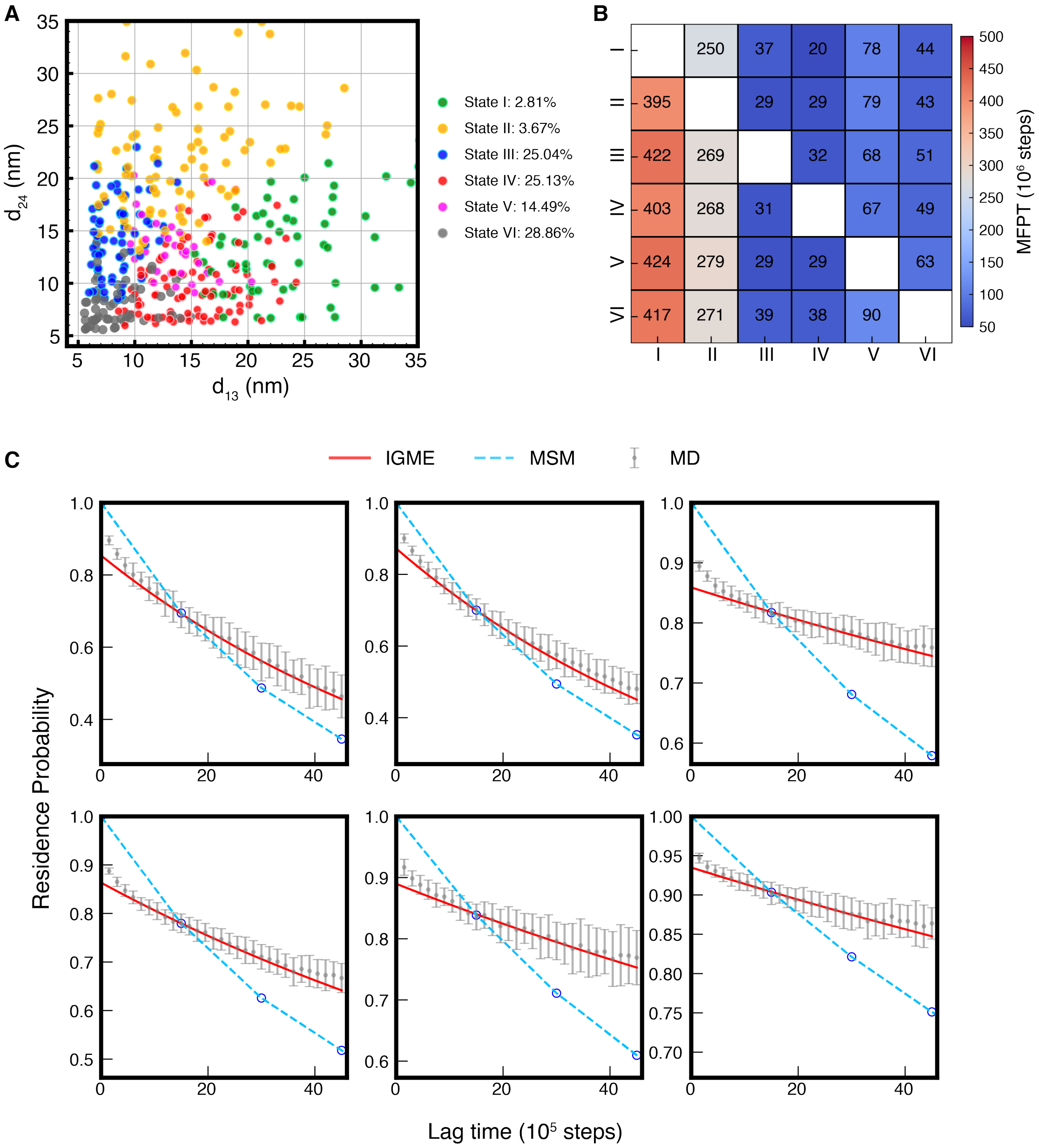


Figure D.19: Construct macrostate MSM and predict properties through IGME for the NRL = 172 system. (A) Distribution of macrostates on the d_{13} and d_{24} map along with their corresponding stationary populations. Each dot represents the center position of a microstate, with colors indicating macrostate assignments based on PCCA+. (B) Mean first passage times (MFPTs) of transitions between macrostates predicted by IGME. The numbers indicate the predicted MFPTs in unit 10^6 steps from row index states to column index states. (C) CK test for the 6-macrostate model using 3.5×10^5 steps MSM and IGME. The error bars are estimated by bootstrapping the MD trajectories 20 times.



	I	II	III	IV	V	VI
I		18	18	15	28	7
II	1669		22	14	60	6
III	1669	22		14	60	6
IV	1671	18	18		60	7
V	1571	17	17	13		7
VI	1689	20	21	18	66	

Figure D.20: Construct macrostate MSM and predict properties through IGME for the NRL = 177 system. (A) Distribution of macrostates on the d_{13} and d_{24} map along with their corresponding stationary populations. Each dot represents the center position of a microstate, with colors indicating macrostate assignments based on PCCA+. (B) Mean first passage times (MFPTs) of transitions between macrostates predicted by IGME. The numbers indicate the predicted MFPTs in unit 10^6 steps from row index states to column index states. (C) CK test for the 6-macrostate model using 3.5×10^5 steps MSM and IGME. The error bars are estimated by bootstrapping the MD trajectories 20 times.



	I	II	III	IV	V	VI
I		250	37	20	78	44
II	395		29	29	79	43
III	422	269		32	68	51
IV	403	268	31		67	49
V	424	279	29	29		63
VI	417	271	39	38	90	

Figure D.21: Construct macrostate MSM and predict properties through IGME for the NRL = 167 condensate system. (A) Distribution of macrostates on the d_{13} and d_{24} map along with their corresponding stationary populations. Each dot represents the center position of a microstate, with colors indicating macrostate assignments based on PCCA+. (B) Mean first passage times (MFPTs) of transitions between macrostates predicted by IGME. The numbers indicate the predicted MFPTs in unit 10^6 steps from row index states to column index states. (C) CK test for the 6-macrostate model using 1.5×10^6 steps MSM and IGME. The error bars are estimated by bootstrapping the MD trajectories 20 times.

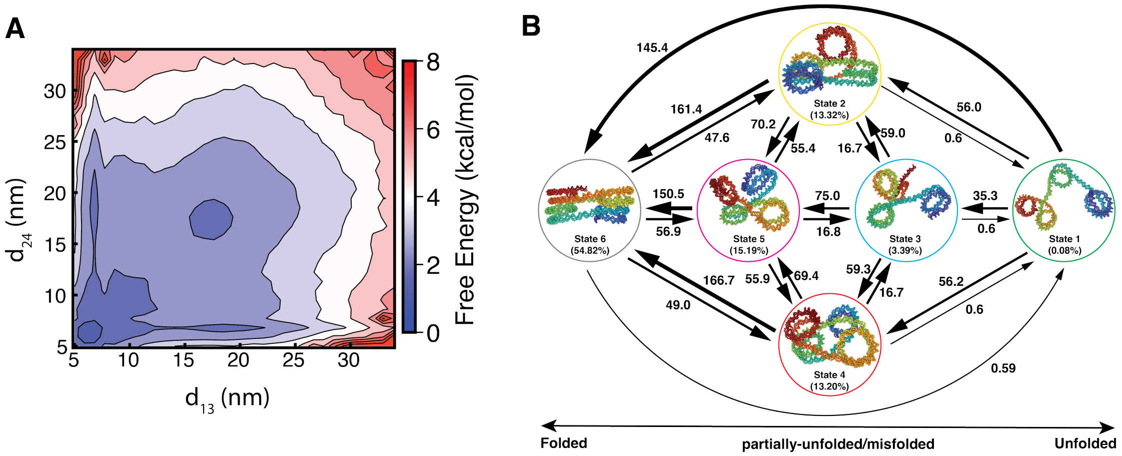


Figure D.22: Analysis results of the NRL = 177 tetra-nucleosome system. (A) The free energy profile along d_{13} and d_{24}. (B) Macrostate non-Markovian dynamics model with inverse MFPT labeled in unit $(10\ \mu s)^{-1}$. Histones are hidden for clarity.

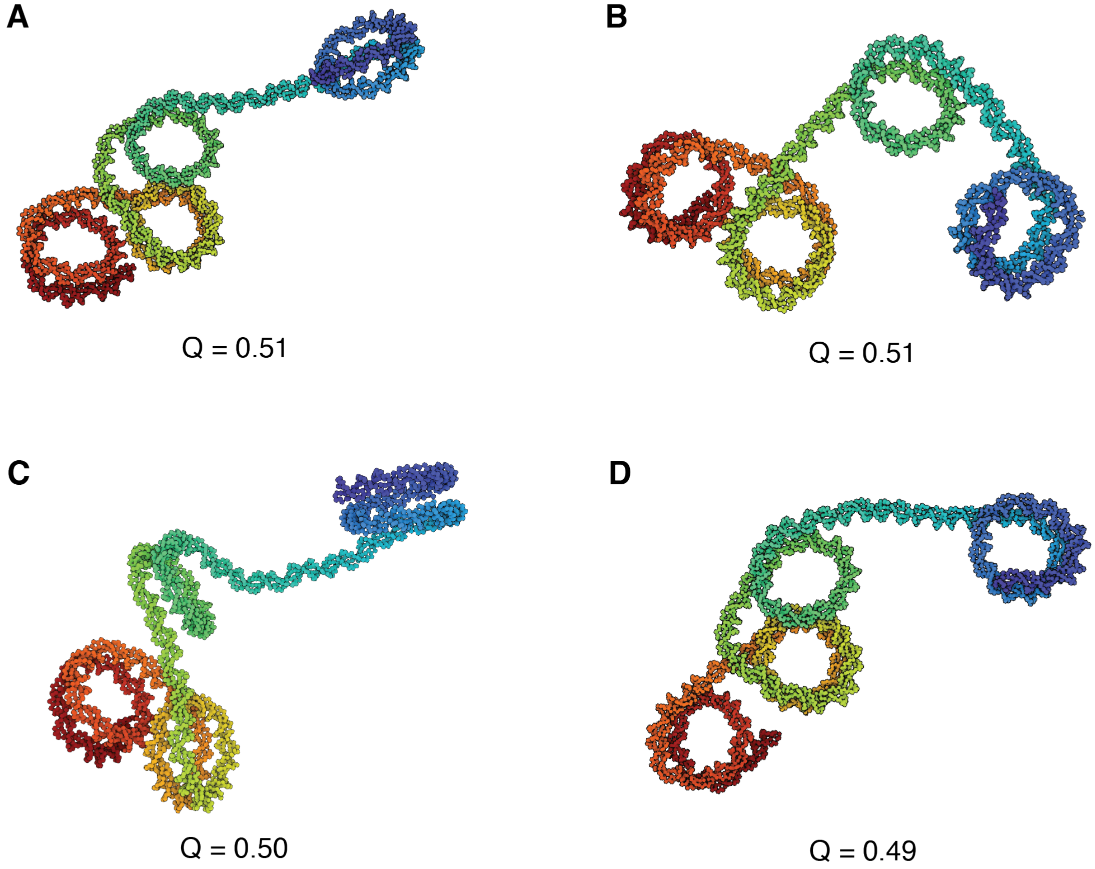


Figure D.23: Representative transition state structures of NRL = 167 singe tetra-nucleosome with committor function close to 0.5. The committor values Q are labeled below each structure.

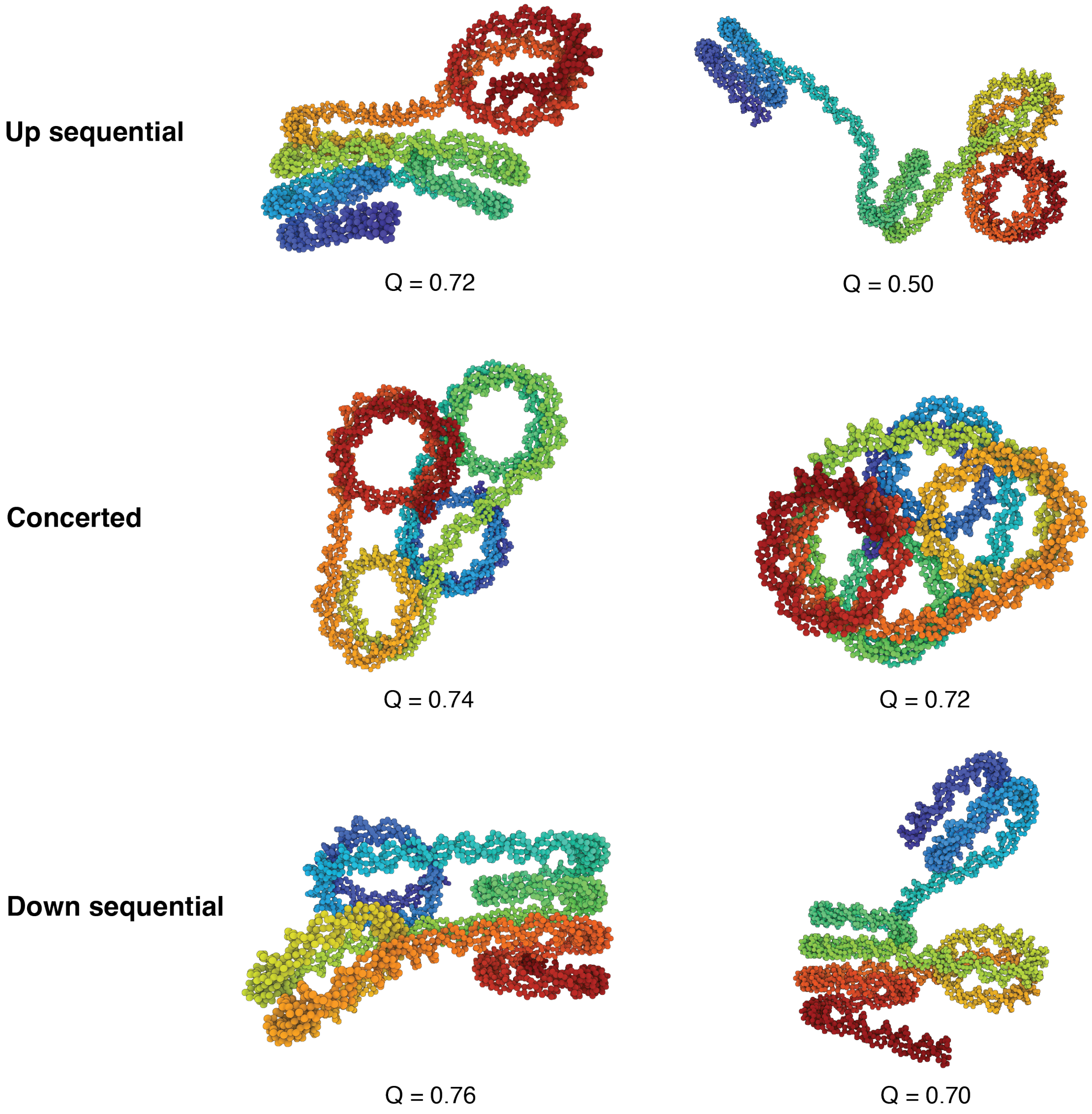


Figure D.24: Representative structures along 3 path channels of NRL = 167 singe tetranucleosome with the highest fluxes. The committor values Q are labeled below each structure.

Table D.1: Compare NRL = 167 tetra-nucleosome OpenMM energy with LAMMPS energy by rerunning a trajectory. The trajectory is a 1-million-step unbiased NVT simulation at 300 K with a timestep of 10 fs performed by LAMMPS. During the rerun, protein dihedrals and native pairs are not included in the comparisons as these potentials are not applied in the simulations. LAMMPS rerun uses the old DNA base step geometry parameters, while OpenMM rerun uses the new ones reported in Open3SPN2 paper [147]. The old and the new base step parameters slightly affect the DNA template structure, thus slightly changing the equilibrium DNA bond, angle, and dihedrals. Meanwhile, cross-stacking interactions are also updated in Open3SPN2 [147]. However, the difference is negligible and should not affect our results. All the energy values are in kcal/mol.

		Protein		DNA							
Frame ID	Software	Bonds	Angles	Bonds	Angles	Stackings	Dihedrals	Base pairs	Cross stackings	Contacts	Electrostatics
1	LAMMPS	11509.91	679.42	12946.48	10229.06	7833.97	-3290.11	-1270.28	-417.01	60.62	-441.75
1	OpenMM	11509.90	679.41	12946.09	10228.26	7833.98	-3290.19	-1270.28	-421.86	60.63	-441.75
2	LAMMPS	11529.70	662.17	12925.30	10195.33	7838.65	-3270.53	-1268.79	-425.52	52.15	-424.23
2	OpenMM	11529.69	662.17	12925.39	10194.42	7838.60	-3270.46	-1268.79	-429.52	52.17	-424.22
3	LAMMPS	11522.15	712.93	12948.77	10205.07	7845.94	-3275.16	-1282.39	-417.87	65.08	-443.99
3	OpenMM	11522.14	712.93	12948.89	10203.98	7845.74	-3275.17	-1282.39	-422.89	65.07	-443.99
4	LAMMPS	11509.97	706.53	12911.58	10191.69	7844.01	-3267.21	-1266.77	-419.71	54.90	-458.91
4	OpenMM	11509.96	706.53	12910.91	10190.58	7843.89	-3267.11	-1266.77	-423.64	54.90	-458.90
5	LAMMPS	11508.79	666.34	12946.52	10222.69	7822.12	-3264.30	-1280.75	-420.09	46.84	-448.44
5	OpenMM	11508.79	666.34	12946.62	10220.81	7822.04	-3264.25	-1280.76	-423.13	46.86	-448.44
6	LAMMPS	11504.78	671.32	12939.06	10184.24	7817.63	-3270.48	-1293.49	-421.79	52.79	-451.66
6	OpenMM	11504.76	671.32	12938.25	10184.10	7817.57	-3270.51	-1293.50	-424.95	52.78	-451.66
7	LAMMPS	11504.54	695.20	12930.58	10217.42	7842.98	-3268.54	-1275.23	-422.52	65.59	-446.10
7	OpenMM	11504.53	695.20	12930.50	10215.45	7842.93	-3268.63	-1275.22	-427.14	65.58	-446.10
8	LAMMPS	11517.74	662.87	12934.04	10190.59	7828.94	-3277.05	-1273.37	-420.83	59.31	-410.87
8	OpenMM	11517.73	662.87	12933.44	10189.83	7828.92	-3277.16	-1273.38	-425.15	59.33	-410.87
9	LAMMPS	11509.69	689.16	12926.25	10248.27	7855.45	-3237.75	-1291.39	-416.42	65.33	-429.43
9	OpenMM	11509.68	689.16	12927.07	10247.74	7855.40	-3237.77	-1291.38	-420.83	65.34	-429.43
10	LAMMPS	11483.97	680.86	12943.83	10224.32	7834.67	-3262.36	-1278.69	-419.54	61.44	-468.65
10	OpenMM	11483.95	680.86	12944.76	10222.46	7834.55	-3262.28	-1278.70	-423.07	61.43	-468.65

Table D.2: Compare two single nucleosome OpenMM energy with LAMMPS energy by rerunning a trajectory. The trajectory is a 0.1-million-step unbiased NVT simulation at 300 K with a timestep of 10 fs performed by LAMMPS. The snapshots were saved every 10,000 steps. Small difference is caused by same reasons mentioned in table D.1 and is negligible. All the energy values are in unit kcal/mol.

		Protein		DNA							
Frame ID	Software	Bonds	Angles	Bonds	Angles	Stackings	Dihedrals	Base pairs	Cross stackings	Contacts	Electrostatics
1	LAMMPS	1650.88	189.64	2143.52	3168.15	2644.02	-1469.24	-393.78	-187.58	55.97	-183.99
1	OpenMM	1651.60	189.66	2230.20	3162.02	2644.01	-1469.78	-393.78	-187.14	55.97	-183.99
2	LAMMPS	1654.88	189.08	2142.63	3193.59	2640.01	-1474.48	-406.79	-188.36	60.63	-192.49
2	OpenMM	1655.60	189.11	2228.84	3186.11	2640.01	-1474.94	-406.79	-188.46	60.64	-192.49
3	LAMMPS	1646.26	198.79	2140.26	3191.23	2646.23	-1468.23	-407.74	-186.77	51.89	-203.44
3	OpenMM	1647.04	198.83	2226.49	3182.68	2646.24	-1469.18	-407.74	-187.02	51.89	-203.44
4	LAMMPS	1644.86	179.53	2158.04	3216.87	2637.28	-1470.98	-429.57	-190.11	51.74	-207.68
4	OpenMM	1645.62	179.57	2244.57	3210.04	2637.29	-1471.70	-429.57	-191.16	51.75	-207.68
5	LAMMPS	1647.16	174.63	2146.99	3193.40	2646.37	-1473.03	-408.23	-185.41	56.07	-218.35
5	OpenMM	1647.92	174.68	2233.76	3186.91	2646.37	-1473.55	-408.23	-185.14	56.08	-218.35
6	LAMMPS	1659.40	201.84	2149.16	3177.91	2661.89	-1468.73	-404.40	-181.06	60.03	-208.02
6	OpenMM	1660.08	201.86	2235.80	3172.52	2661.90	-1469.37	-404.40	-181.55	60.02	-208.03
7	LAMMPS	1640.32	203.55	2139.00	3176.86	2637.54	-1473.43	-419.26	-187.15	51.70	-209.24
7	OpenMM	1640.92	203.58	2225.15	3172.96	2637.53	-1473.88	-419.26	-188.17	51.69	-209.24
8	LAMMPS	1649.89	188.39	2156.81	3200.63	2646.82	-1472.63	-402.73	-184.09	58.28	-189.88
8	OpenMM	1650.75	188.41	2243.95	3194.89	2646.82	-1473.11	-402.73	-184.78	58.29	-189.88
9	LAMMPS	1649.54	179.47	2139.95	3193.25	2647.39	-1468.93	-395.35	-182.68	50.21	-201.40
9	OpenMM	1650.36	179.50	2226.19	3185.96	2647.40	-1469.56	-395.35	-181.72	50.21	-201.40
10	LAMMPS	1651.76	197.59	2150.19	3188.49	2644.81	-1465.87	-412.90	-185.01	50.58	-205.03
10	OpenMM	1652.30	197.58	2236.09	3183.64	2644.81	-1466.44	-412.89	-185.43	50.58	-205.03

References

[1] X. Lin, Y. Qi, A. P. Latham, and B. Zhang, “Multiscale modeling of genome organization with maximum entropy optimization,” *The Journal of chemical physics*, vol. 155, no. 1, 2021.

[2] S. Liu, A. Athreya, Z. Lao, and B. Zhang, “From nucleosomes to compartments: Physicochemical interactions underlying chromatin organization,” *Annual Review of Biophysics*, vol. 53, 2024.

[3] K. Luger, A. W. Mäder, R. K. Richmond, D. F. Sargent, and T. J. Richmond, “Crystal structure of the nucleosome core particle at 2.8 Å resolution,” *Nature*, vol. 389, no. 6648, pp. 251–260, 1997.

[4] C. A. Davey, D. F. Sargent, K. Luger, A. W. Maeder, and T. J. Richmond, “Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 Å resolution,” *Journal of molecular biology*, vol. 319, no. 5, pp. 1097–1113, 2002.

[5] R. K. McGinty and S. Tan, “Nucleosome structure and function,” *Chemical reviews*, vol. 115, no. 6, pp. 2255–2273, 2015.

[6] B. Li, M. Carey, and J. L. Workman, “The role of chromatin during transcription,” *Cell*, vol. 128, no. 4, pp. 707–719, 2007.

[7] S. Eustermann, A. B. Patel, K.-P. Hopfner, Y. He, and P. Korber, “Energy-driven genome regulation by atp-dependent chromatin remodellers,” *Nature Reviews Molecular Cell Biology*, pp. 1–24, 2023.

[8] T. Schalch, S. Duda, D. F. Sargent, and T. J. Richmond, “X-ray structure of a tetranucleosome and its implications for the chromatin fibre,” *Nature*, vol. 436, no. 7047, pp. 138–141, 2005.

[9] F. Song, P. Chen, D. Sun, M. Wang, L. Dong, D. Liang, R.-M. Xu, P. Zhu, and G. Li, “Cryo-em study of the chromatin fiber reveals a double helix twisted by tetranucleosomal units,” *Science*, vol. 344, no. 6182, pp. 376–380, 2014.

[10] M. Eltsov, K. M. MacLellan, K. Maeshima, A. S. Frangakis, and J. Dubochet, “Analysis of cryo-electron microscopy images does not support the existence of 30-nm chromatin fibers in mitotic chromosomes in situ,” *Proceedings of the National Academy of Sciences*, vol. 105, no. 50, pp. 19 732–19 737, 2008.

[11] K. Maeshima, R. Rogge, S. Tamura, Y. Joti, T. Hikima, H. Szerlong, C. Krause, J. Herman, E. Seidel, J. DeLuca, *et al.*, "Nucleosomal arrays self-assemble into supramolecular globular structures lacking 30-nm fibers," *The EMBO journal*, vol. 35, no. 10, pp. 1115–1132, 2016.

[12] K. Maeshima, S. Ide, and M. Babokhov, "Dynamic chromatin organization without the 30-nm fiber," *Current opinion in cell biology*, vol. 58, pp. 95–104, 2019.

[13] H. D. Ou, S. Phan, T. J. Deerinck, A. Thor, M. H. Ellisman, and C. C. O'shea, "Chromemt: Visualizing 3d chromatin structure and compaction in interphase and mitotic cells," *Science*, vol. 357, no. 6349, eaag0025, 2017.

[14] S. F. Banani, H. O. Lee, A. A. Hyman, and M. K. Rosen, "Biomolecular condensates: Organizers of cellular biochemistry," *Nature reviews Molecular cell biology*, vol. 18, no. 5, pp. 285–298, 2017.

[15] A. S. Lyon, W. B. Peeples, and M. K. Rosen, "A framework for understanding the functions of biomolecular condensates across scales," *Nature Reviews Molecular Cell Biology*, vol. 22, no. 3, pp. 215–235, 2021.

[16] C. P. Brangwynne, P. Tompa, and R. V. Pappu, "Polymer physics of intracellular phase transitions," *Nature Physics*, vol. 11, no. 11, pp. 899–904, 2015.

[17] B. A. Gibson, L. K. Doolittle, M. W. Schneider, L. E. Jensen, N. Gamarra, L. Henry, D. W. Gerlich, S. Redding, and M. K. Rosen, "Organization of chromatin by intrinsic and regulated phase separation," *Cell*, vol. 179, no. 2, pp. 470–484, 2019.

[18] B. A. Gibson, C. Blaukopf, T. Lou, L. Chen, L. K. Doolittle, I. Finkelstein, G. J. Narlikar, D. W. Gerlich, and M. K. Rosen, "In diverse conditions, intrinsic chromatin condensates have liquid-like material properties," *Proceedings of the National Academy of Sciences*, vol. 120, no. 18, e2218085120, 2023.

[19] T. Nozaki, S. Shinkai, S. Ide, K. Higashi, S. Tamura, M. A. Shimazoe, M. Nakagawa, Y. Suzuki, Y. Okada, M. Sasai, *et al.*, "Condensed but liquid-like domain organization of active chromatin regions in living human cells," *Science Advances*, vol. 9, no. 14, eadf1488, 2023.

[20] H. Strickfaden, T. O. Tolsma, A. Sharma, D. A. Underhill, J. C. Hansen, and M. J. Hendzel, "Condensed chromatin behaves like a solid on the mesoscale in vitro and in living cells," *Cell*, vol. 183, no. 7, pp. 1772–1784, 2020.

[21] J. C. Hansen, K. Maeshima, and M. J. Hendzel, "The solid and liquid states of chromatin," *Epigenetics & Chromatin*, vol. 14, no. 1, p. 50, 2021.

[22] M. Karplus and J. A. McCammon, "Molecular dynamics simulations of biomolecules," *Nature structural biology*, vol. 9, no. 9, pp. 646–652, 2002.

[23] S. A. Hollingsworth and R. O. Dror, "Molecular dynamics simulation for all," *Neuron*, vol. 99, no. 6, pp. 1129–1143, 2018.

[24] S. Kilic, S. Felekyan, O. Doroshenko, I. Boichenko, M. Dimura, H. Vardanyan, L. C. Bryan, G. Arya, C. A. Seidel, and B. Fierz, "Single-molecule fret reveals multiscale chromatin dynamics modulated by hp1α," *Nature communications*, vol. 9, no. 1, p. 235, 2018.

[25] M. Zhang, C. Díaz-Celis, B. Onoa, C. Cañari-Chumpitaz, K. I. Requejo, J. Liu, M. Vien, E. Nogales, G. Ren, and C. Bustamante, "Molecular organization of the early stages of nucleosome phase separation visualized by cryo-electron tomography," *Molecular cell*, vol. 82, no. 16, pp. 3000–3014, 2022.

[26] D. E. Shaw, P. J. Adams, A. Azaria, J. A. Bank, B. Batson, A. Bell, M. Bergdorf, J. Bhatt, J. A. Butts, T. Correia, *et al.*, "Anton 3: Twenty microseconds of molecular dynamics simulation before lunch," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 2021, pp. 1–11.

[27] S. Kmiecik, D. Gront, M. Kolinski, L. Wieteska, A. E. Dawid, and A. Kolinski, "Coarse-grained protein models and their applications," *Chemical reviews*, vol. 116, no. 14, pp. 7898–7936, 2016.

[28] M. Chakraborty, C. Xu, and A. D. White, "Encoding and selecting coarse-grain mapping operators with hierarchical graphs," *The Journal of Chemical Physics*, vol. 149, no. 13, 2018.

[29] W. Wang and R. Gómez-Bombarelli, "Coarse-graining auto-encoders for molecular dynamics," *npj Computational Materials*, vol. 5, no. 1, p. 125, 2019.

[30] J. Jin, A. J. Pak, A. E. Durumeric, T. D. Loose, and G. A. Voth, "Bottom-up coarse-graining: Principles and perspectives," *Journal of Chemical Theory and Computation*, vol. 18, no. 10, pp. 5759–5791, 2022.

[31] A. Davtyan, N. P. Schafer, W. Zheng, C. Clementi, P. G. Wolynes, and G. A. Papoian, "Awsem-md: Protein structure prediction using coarse-grained physical potentials and bioinformatically based local structure biasing," *The Journal of Physical Chemistry B*, vol. 116, no. 29, pp. 8494–8503, 2012.

[32] I. Zaporozhets and C. Clementi, "Multibody terms in protein coarse-grained models: A top-down perspective," *The Journal of Physical Chemistry B*, vol. 127, no. 31, pp. 6920–6927, 2023.

[33] J. S. Smith, O. Isayev, and A. E. Roitberg, "Ani-1: An extensible neural network potential with dft accuracy at force field computational cost," *Chemical science*, vol. 8, no. 4, pp. 3192–3203, 2017.

[34] L. Zhang, J. Han, H. Wang, R. Car, and E. Weinan, "Deep potential molecular dynamics: A scalable model with the accuracy of quantum mechanics," *Physical review letters*, vol. 120, no. 14, p. 143 001, 2018.

[35] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, "Schnet–a deep learning architecture for molecules and materials," *The Journal of Chemical Physics*, vol. 148, no. 24, 2018.

[36] J. Wang, S. Olsson, C. Wehmeyer, A. Pérez, N. E. Charron, G. De Fabritiis, F. Noé, and C. Clementi, "Machine learning of coarse-grained molecular dynamics force fields," *ACS central science*, vol. 5, no. 5, pp. 755–767, 2019.

[37] F. Noé, A. Tkatchenko, K.-R. Müller, and C. Clementi, "Machine learning for molecular simulation," *Annual review of physical chemistry*, vol. 71, pp. 361–390, 2020.

[38] J. Behler, "Four generations of high-dimensional neural network potentials," *Chemical Reviews*, vol. 121, no. 16, pp. 10 037–10 072, 2021.

[39] S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt, and B. Kozinsky, "E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials," *Nature communications*, vol. 13, no. 1, p. 2453, 2022.

[40] J. Airas, X. Ding, and B. Zhang, "Transferable implicit solvation via contrastive learning of graph neural networks," *ACS Central Science*, vol. 9, no. 12, pp. 2286–2297, 2023.

[41] H. I. Ingólfsson, C. A. Lopez, J. J. Uusitalo, D. H. de Jong, S. M. Gopal, X. Periole, and S. J. Marrink, "The power of coarse graining in biomolecular simulations," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 4, no. 3, pp. 225–248, 2014.

[42] W. G. Noid, J.-W. Chu, G. S. Ayton, V. Krishna, S. Izvekov, G. A. Voth, A. Das, and H. C. Andersen, "The multiscale coarse-graining method. i. a rigorous bridge between atomistic and coarse-grained models," *The Journal of chemical physics*, vol. 128, no. 24, 2008.

[43] M. S. Shell, "The relative entropy is fundamental to multiscale and inverse thermodynamic problems," *The Journal of chemical physics*, vol. 129, no. 14, 2008.

[44] M. Gutmann and A. Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, JMLR Workshop and Conference Proceedings, 2010, pp. 297–304.

[45] M. U. Gutmann and A. Hyvärinen, "Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics.," *Journal of machine learning research*, vol. 13, no. 2, 2012.

[46] X. Ding, X. Lin, and B. Zhang, "Stability and folding pathways of tetra-nucleosome from six-dimensional free energy surface," *Nature communications*, vol. 12, no. 1, p. 1091, 2021.

[47] J. Moller and J. J. de Pablo, "Bottom-up meets top-down: The crossroads of multiscale chromatin modeling," *Biophysical Journal*, vol. 118, no. 9, pp. 2057–2065, 2020.

[48] S. E. Farr, E. J. Woods, J. A. Joseph, A. Garaizar, and R. Collepardo-Guevara, "Nucleosome plasticity is a critical element of chromatin liquid–liquid phase separation and multivalent nucleosome interactions," *Nature communications*, vol. 12, no. 1, p. 2883, 2021.

[49] S. Liu, X. Lin, and B. Zhang, "Chromatin fiber breaks into clutches under tension and crowding," *Nucleic Acids Research*, vol. 50, no. 17, pp. 9738–9747, 2022.

[50] R. Leicher, E. J. Ge, X. Lin, M. J. Reynolds, W. Xie, T. Walz, B. Zhang, T. W. Muir, and S. Liu, "Single-molecule and in silico dissection of the interaction between polycomb repressive complex 2 and chromatin," *Proceedings of the National Academy of Sciences*, vol. 117, no. 48, pp. 30 465–30 475, 2020.

[51] C. Tan and S. Takada, “Nucleosome allostery in pioneer transcription factor binding,” *Proceedings of the National Academy of Sciences*, vol. 117, no. 34, pp. 20 586–20 596, 2020.

[52] A. Kumar and H. Kono, “Heterochromatin protein 1 (hp1): Interactions with itself and chromatin components,” *Biophysical reviews*, vol. 12, no. 2, pp. 387–400, 2020.

[53] D. V. Fyodorov, B.-R. Zhou, A. I. Skoultchi, and Y. Bai, “Emerging roles of linker histones in regulating chromatin structure and function,” *Nature reviews Molecular cell biology*, vol. 19, no. 3, pp. 192–206, 2018.

[54] K. Cermakova and H. C. Hodges, “Interaction modules that impart specificity to disordered protein,” *Trends in biochemical sciences*, vol. 48, no. 5, pp. 477–490, 2023.

[55] A. P. Latham and B. Zhang, “Consistent force field captures homologue-resolved hp1 phase separation,” *Journal of chemical theory and computation*, vol. 17, no. 5, pp. 3134–3144, 2021.

[56] G. L. Dignon, W. Zheng, Y. C. Kim, R. B. Best, and J. Mittal, “Sequence determinants of protein phase behavior from a coarse-grained model,” *PLoS computational biology*, vol. 14, no. 1, e1005941, 2018.

[57] R. M. Regy, J. Thompson, Y. C. Kim, and J. Mittal, “Improved coarse-grained model for studying sequence dependent phase separation of disordered proteins,” *Protein Science*, vol. 30, no. 7, pp. 1371–1379, 2021.

[58] J. A. Joseph, A. Reinhardt, A. Aguirre, P. Y. Chew, K. O. Russell, J. R. Espinosa, A. Garaizar, and R. Collepardo-Guevara, “Physics-driven coarse-grained model for biomolecular phase separation with near-quantitative accuracy,” *Nature Computational Science*, vol. 1, no. 11, pp. 732–743, 2021.

[59] G. Tesei, T. K. Schulze, R. Crehuet, and K. Lindorff-Larsen, “Accurate model of liquid–liquid phase behavior of intrinsically disordered proteins from optimization of single-chain properties,” *Proceedings of the National Academy of Sciences*, vol. 118, no. 44, e2111696118, 2021.

[60] S. Hihara, C.-G. Pack, K. Kaizu, T. Tani, T. Hanafusa, T. Nozaki, S. Takemoto, T. Yoshimi, H. Yokota, N. Imamoto, *et al.*, “Local nucleosome dynamics facilitate chromatin accessibility in living mammalian cells,” *Cell reports*, vol. 2, no. 6, pp. 1645–1656, 2012.

[61] S. Liu, C. Wang, A. P. Latham, X. Ding, and B. Zhang, “Openabc enables flexible, simplified, and efficient gpu accelerated simulations of biomolecular condensates,” *PLOS Computational Biology*, vol. 19, no. 9, e1011442, 2023.

[62] P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, *et al.*, “Openmm 7: Rapid development of high performance algorithms for molecular dynamics,” *PLoS computational biology*, vol. 13, no. 7, e1005659, 2017.

[63] S. Cao, Y. Qiu, M. L. Kalin, and X. Huang, “Integrative generalized master equation: A method to study long-timescale biomolecular dynamics via the integrals of memory kernels,” *The Journal of Chemical Physics*, vol. 159, no. 13, 2023.

[64] E. Oberbeckmann, K. Quililan, P. Cramer, and A. M. Oudelaar, “In vitro reconstitution of chromatin domains shows a role for nucleosome positioning in 3d genome organization,” *Nature Genetics*, pp. 1–10, 2024.

[65] Y. I. Yang, Q. Shao, J. Zhang, L. Yang, and Y. Q. Gao, “Enhanced sampling in molecular dynamics,” *The Journal of chemical physics*, vol. 151, no. 7, 2019.

[66] Y. Song and D. P. Kingma, “How to train your energy-based models,” *arXiv preprint arXiv:2101.03288*, 2021.

[67] G. E. Hinton, “Training products of experts by minimizing contrastive divergence,” *Neural computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[68] A. Hyvärinen and P. Dayan, “Estimation of non-normalized statistical models by score matching.,” *Journal of Machine Learning Research*, vol. 6, no. 4, 2005.

[69] X. Ding and B. Zhang, “Contrastive learning of coarse-grained force fields,” *Journal of chemical theory and computation*, vol. 18, no. 10, pp. 6334–6344, 2022.

[70] X. Ding, “Optimizing force fields with experimental data using ensemble reweighting and potential contrasting,” 2024.

[71] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw, “How fast-folding proteins fold,” *Science*, vol. 334, no. 6055, pp. 517–520, 2011.

[72] P. Robustelli, S. Piana, and D. E. Shaw, “Developing a molecular dynamics force field for both folded and disordered protein states,” *Proceedings of the National Academy of Sciences*, vol. 115, no. 21, E4758–E4766, 2018.

[73] S. Piana, P. Robustelli, D. Tan, S. Chen, and D. E. Shaw, “Development of a force field for the simulation of single-chain proteins and protein–protein complexes,” *Journal of chemical theory and computation*, vol. 16, no. 4, pp. 2494–2507, 2020.

[74] H. S. Ashbaugh and H. W. Hatch, “Natively unfolded protein stability as a coil-to-globule transition in charge/hydropathy space,” *Journal of the American Chemical Society*, vol. 130, no. 29, pp. 9536–9542, 2008.

[75] G. M. Torrie and J. P. Valleau, “Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling,” *Journal of computational physics*, vol. 23, no. 2, pp. 187–199, 1977.

[76] X. Ding, J. Z. Vilseck, and C. L. Brooks III, “Fast solver for large scale multistate bennett acceptance ratio equations,” *Journal of chemical theory and computation*, vol. 15, no. 2, pp. 799–802, 2019.

[77] K. Sarthak, D. Winogradoff, Y. Ge, S. Myong, and A. Aksimentiev, “Benchmarking molecular dynamics force fields for all-atom simulations of biological condensates,” *Journal of Chemical Theory and Computation*, vol. 19, no. 12, pp. 3721–3740, 2023.

[78] T. Zarin, B. Strome, A. N. Nguyen Ba, S. Alberti, J. D. Forman-Kay, and A. M. Moses, “Proteome-wide signatures of function in highly diverged intrinsically disordered regions,” *Elife*, vol. 8, e46883, 2019.

[79] S. Miyazawa and R. L. Jernigan, “Residue–residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading,” *Journal of molecular biology*, vol. 256, no. 3, pp. 623–644, 1996.

[80] O. T. Unke, M. Stöhr, S. Ganscha, T. Unterthiner, H. Maennel, S. Kashubin, D. Ahlin, M. Gastegger, L. Medrano Sandonas, J. T. Berryman, *et al.*, “Biomolecular dynamics with machine-learned quantum-mechanical force fields trained on diverse chemical fragments,” *Science Advances*, vol. 10, no. 14, eadn4397, 2024.

[81] C. P. Brangwynne, T. J. Mitchison, and A. A. Hyman, “Active liquid-like behavior of nucleoli determines their size and shape in xenopus laevis oocytes,” *Proceedings of the National Academy of Sciences*, vol. 108, no. 11, pp. 4334–4339, 2011.

[82] A. Boija, I. A. Klein, and R. A. Young, “Biomolecular condensates and cancer,” *Cancer cell*, vol. 39, no. 2, pp. 174–192, 2021.

[83] W. Borcherds, A. Bremer, M. B. Borgia, and T. Mittag, “How do intrinsically disordered protein regions encode a driving force for liquid–liquid phase separation?” *Current opinion in structural biology*, vol. 67, pp. 41–50, 2021.

[84] A. S. Belmont, “Nuclear compartments: An incomplete primer to nuclear compartments, bodies, and genome organization relative to nuclear architecture,” *Cold Spring Harbor Perspectives in Biology*, vol. 14, no. 7, a041268, 2022.

[85] A. P. Latham and B. Zhang, “Molecular determinants for the layering and coarsening of biological condensates,” *Aggregate*, e306, 2022.

[86] R. V. Pappu, S. R. Cohen, F. Dar, M. Farag, and M. Kar, “Phase transitions of associative biomacromolecules,” *Chemical Reviews*, 2023.

[87] C. P. Brangwynne, C. R. Eckmann, D. S. Courson, A. Rybarska, C. Hoege, J. Gharakhani, F. Jülicher, and A. A. Hyman, “Germline p granules are liquid droplets that localize by controlled dissolution/condensation,” *Science*, vol. 324, no. 5935, pp. 1729–1732, 2009.

[88] Y. Shin and C. P. Brangwynne, “Liquid phase condensation in cell physiology and disease,” *Science*, vol. 357, no. 6357, eaaf4382, 2017.

[89] G. J. Narlikar, “Phase-separation in chromatin organization,” *Journal of biosciences*, vol. 45, pp. 1–5, 2020.

[90] J.-M. Choi, A. S. Holehouse, and R. V. Pappu, “Physical principles underlying the complex biology of intracellular phase transitions,” *Annual review of biophysics*, vol. 49, pp. 107–133, 2020.

[91] B. R. Sabari, A. Dall’Agnese, and R. A. Young, “Biomolecular condensates in the nucleus,” *Trends in biochemical sciences*, vol. 45, no. 11, pp. 961–977, 2020.

[92] P. Bhat, D. Honson, and M. Guttman, “Nuclear compartmentalization as a mechanism of quantitative control of gene expression,” *Nature Reviews Molecular Cell Biology*, vol. 22, no. 10, pp. 653–670, 2021.

[93] N. Hori and S. Takada, “Coarse-grained structure-based model for rna-protein complexes developed by fluctuation matching,” *Journal of Chemical Theory and Computation*, vol. 8, no. 9, pp. 3384–3394, 2012.

[94] B. E. Husic, N. E. Charron, D. Lemm, J. Wang, A. Pérez, M. Majewski, A. Krämer, Y. Chen, S. Olsson, G. de Fabritiis, *et al.*, "Coarse graining molecular dynamics with graph neural networks," *The Journal of chemical physics*, vol. 153, no. 19, p. 194 101, 2020.

[95] R. M. Regy, G. L. Dignon, W. Zheng, Y. C. Kim, and J. Mittal, "Sequence dependent phase separation of protein-polynucleotide mixtures elucidated using molecular simulations," *Nucleic acids research*, vol. 48, no. 22, pp. 12 593–12 603, 2020.

[96] Y. Qi and B. Zhang, "Chromatin network retards nucleoli coalescence," *Nature Communications*, vol. 12, no. 1, p. 6824, 2021.

[97] K. Kamat, Z. Lao, Y. Qi, Y. Wang, J. Ma, and B. Zhang, "Compartmentalization with nuclear landmarks yields random, yet precise, genome organization," *Biophysical Journal*, vol. 122, no. 7, pp. 1376–1389, 2023.

[98] P. C. Souza, R. Alessandri, J. Barnoud, S. Thallmair, I. Faustino, F. Grünewald, I. Patmanidis, H. Abdizadeh, B. M. Bruininks, T. A. Wassenaar, *et al.*, "Martini 3: A general purpose force field for coarse-grained molecular dynamics," *Nature methods*, vol. 18, no. 4, pp. 382–388, 2021.

[99] T. Dannenhoffer-Lafage and R. B. Best, "A data-driven hydrophobicity scale for predicting liquid–liquid phase separation of proteins," *The Journal of Physical Chemistry B*, vol. 125, no. 16, pp. 4046–4056, 2021.

[100] W. Li, W. Wang, and S. Takada, "Energy landscape views for interplays among folding, binding, and allostery of calmodulin domains," *Proceedings of the National Academy of Sciences*, vol. 111, no. 29, pp. 10 550–10 555, 2014.

[101] J. Wessén, T. Pal, S. Das, Y.-H. Lin, and H. S. Chan, "A simple explicit-solvent model of polyampholyte phase behaviors and its ramifications for dielectric effects in biomolecular condensates," *The Journal of Physical Chemistry B*, vol. 125, no. 17, pp. 4337–4358, 2021.

[102] X. Lin, R. Leicher, S. Liu, and B. Zhang, "Cooperative dna looping by prc2 complexes," *Nucleic Acids Research*, vol. 49, no. 11, pp. 6238–6248, 2021.

[103] Y. Chen, A. Krämer, N. E. Charron, B. E. Husic, C. Clementi, and F. Noé, "Machine learning implicit solvation for molecular dynamics," *The Journal of Chemical Physics*, vol. 155, no. 8, p. 084 101, 2021.

[104] H. T. Nguyen, N. Hori, and D. Thirumalai, "Condensates in rna repeat sequences are heterogeneously organized and exhibit reptation dynamics," *Nature chemistry*, vol. 14, no. 7, pp. 775–785, 2022.

[105] C. Tan, J. Jung, C. Kobayashi, D. U. L. Torre, S. Takada, and Y. Sugita, "Implementation of residue-level coarse-grained models in genesis for large-scale molecular dynamics simulations," *PLOS Computational Biology*, vol. 18, no. 4, e1009578, 2022.

[106] M. Feric, N. Vaidya, T. S. Harmon, D. M. Mitrea, L. Zhu, T. M. Richardson, R. W. Kriwacki, R. V. Pappu, and C. P. Brangwynne, "Coexisting liquid phases underlie nucleolar subcompartments," *Cell*, vol. 165, no. 7, pp. 1686–1697, 2016.

[107] T. S. Harmon, A. S. Holehouse, M. K. Rosen, and R. V. Pappu, "Intrinsically disordered linkers determine the interplay between phase separation and gelation in multivalent proteins," *elife*, vol. 6, e30294, 2017.

[108] G. Shi, L. Liu, C. Hyeon, and D. Thirumalai, "Interphase human chromosome exhibits out of equilibrium glassy dynamics," *Nature communications*, vol. 9, no. 1, p. 3161, 2018.

[109] J.-M. Choi, F. Dar, and R. V. Pappu, "Lassi: A lattice model for simulating phase transitions of multivalent proteins," *PLoS computational biology*, vol. 15, no. 10, e1007028, 2019.

[110] G. L. Dignon, R. B. Best, and J. Mittal, "Biomolecular phase separation: From molecular driving forces to macroscopic properties," *Annual review of physical chemistry*, vol. 71, p. 53, 2020.

[111] U. Baul, D. Chakraborty, M. L. Mugnai, J. E. Straub, and D. Thirumalai, "Sequence effects on size, shape, and structural heterogeneity in intrinsically disordered proteins," *The Journal of Physical Chemistry B*, vol. 123, no. 16, pp. 3462–3474, 2019.

[112] M. Di Pierro, D. A. Potoyan, P. G. Wolynes, and J. N. Onuchic, "Anomalous diffusion, spatial coherence, and viscoelasticity from the energy landscape of human chromosomes," *Proceedings of the National Academy of Sciences*, vol. 115, no. 30, pp. 7753–7758, 2018.

[113] A. Kluber, T. A. Burt, and C. Clementi, "Size and topology modulate the effects of frustration in protein folding," *Proceedings of the National Academy of Sciences*, vol. 115, no. 37, pp. 9234–9239, 2018.

[114] X. Chen, M. Chen, N. P. Schafer, and P. G. Wolynes, "Exploring the interplay between fibrillization and amorphous aggregation channels on the energy landscapes of tau repeat isoforms," *Proceedings of the National Academy of Sciences*, vol. 117, no. 8, pp. 4125–4130, 2020.

[115] H. Wu, Y. Dalal, and G. A. Papoian, "Binding dynamics of disordered linker histone h1 with a nucleosomal particle," *Journal of molecular biology*, vol. 433, no. 6, p. 166 881, 2021.

[116] A. P. Latham and B. Zhang, "On the stability and layered organization of protein-dna condensates," *Biophysical Journal*, vol. 121, no. 9, pp. 1727–1737, 2022.

[117] A. P. Latham, L. Zhu, D. A. Sharon, S. Ye, A. P. Willard, X. Zhang, and B. Zhang, "Frustrated microphase separation produces interfacial environment within biological condensates," *bioRxiv*, pp. 2023–03, 2023.

[118] W. Zheng, G. L. Dignon, N. Jovic, X. Xu, R. M. Regy, N. L. Fawzi, Y. C. Kim, R. B. Best, and J. Mittal, "Molecular details of protein condensates probed by microsecond long atomistic simulations," *The Journal of Physical Chemistry B*, vol. 124, no. 51, pp. 11 671–11 679, 2020.

[119] S. A. Thody, H. D. Clements, H. Baniasadi, A. S. Lyon, M. S. Sigman, and M. K. Rosen, "Small molecule properties define partitioning into biomolecular condensates," *bioRxiv*, pp. 2022–12, 2022.

[120] N. Galvanetto, M. T. Ivanović, A. Chowdhury, A. Sottini, M. Nüesch, D. Nettels, R. Best, and B. Schuler, "Ultrafast molecular dynamics observed within a dense protein condensate," *bioRxiv*, pp. 2022–12, 2022.

[121] I. A. Klein, A. Boija, L. K. Afeyan, S. W. Hawken, M. Fan, A. Dall'Agnese, O. Oksuz, J. E. Henninger, K. Shrinivas, B. R. Sabari, *et al.*, "Partitioning of cancer therapeutics in nuclear condensates," *Science*, vol. 368, no. 6497, pp. 1386–1392, 2020.

[122] H. J. Berendsen, D. van der Spoel, and R. van Drunen, "Gromacs: A message-passing parallel molecular dynamics implementation," *Computer physics communications*, vol. 91, no. 1-3, pp. 43–56, 1995.

[123] S. Jo, T. Kim, V. G. Iyer, and W. Im, "Charmm-gui: A web-based graphical user interface for charmm," *Journal of computational chemistry*, vol. 29, no. 11, pp. 1859–1865, 2008.

[124] R. Salomon-Ferrer, D. A. Case, and R. C. Walker, "An overview of the amber biomolecular simulation package," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 3, no. 2, pp. 198–210, 2013.

[125] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl, "Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers," *SoftwareX*, vol. 1, pp. 19–25, 2015.

[126] J. C. Phillips, D. J. Hardy, J. D. Maia, J. E. Stone, J. V. Ribeiro, R. C. Bernardi, R. Buch, G. Fiorin, J. Hénin, W. Jiang, *et al.*, "Scalable molecular dynamics on cpu and gpu architectures with namd," *The Journal of chemical physics*, vol. 153, no. 4, p. 044 130, 2020.

[127] A. P. Thompson, H. M. Aktulga, R. Berger, D. S. Bolintineanu, W. M. Brown, P. S. Crozier, P. J. in't Veld, A. Kohlmeyer, S. G. Moore, T. D. Nguyen, *et al.*, "Lammps-a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales," *Computer Physics Communications*, vol. 271, p. 108 171, 2022.

[128] Y. Sugita and Y. Okamoto, "Replica-exchange molecular dynamics method for protein folding," *Chemical physics letters*, vol. 314, no. 1-2, pp. 141–151, 1999.

[129] M. Källberg, H. Wang, S. Wang, J. Peng, Z. Wang, H. Lu, and J. Xu, "Template-based protein structure modeling using the raptorx web server," *Nature protocols*, vol. 7, no. 8, pp. 1511–1522, 2012.

[130] Z. Zhang, X. Liu, K. Yan, M. E. Tuckerman, and J. Liu, "Unified efficient thermostat scheme for the canonical ensemble with holonomic or isokinetic constraints via molecular dynamics," *The Journal of Physical Chemistry A*, vol. 123, no. 28, pp. 6056–6079, 2019.

[131] D. W. Urry, D. C. Gowda, T. M. Parker, C.-H. Luan, M. C. Reid, C. M. Harris, A. Pattanaik, and R. D. Harris, "Hydrophobicity scale for proteins based on inverse temperature transitions," *Biopolymers: Original Research on Biomolecules*, vol. 32, no. 9, pp. 1243–1250, 1992.

[132] J. S. Rowlinson and B. Widom, *Molecular theory of capillarity*. Courier Corporation, 2013.

[133] J. K. Noel, M. Levi, M. Raghunathan, H. Lammert, R. L. Hayes, J. N. Onuchic, and P. C. Whitford, "Smog 2: A versatile software package for generating structure-based models," *PLoS computational biology*, vol. 12, no. 3, e1004794, 2016.

[134] A. C. Murthy, G. L. Dignon, Y. Kan, G. H. Zerze, S. H. Parekh, J. Mittal, and N. L. Fawzi, "Molecular interactions underlying liquid-liquid phase separation of the fus low-complexity domain," *Nature structural & molecular biology*, vol. 26, no. 7, pp. 637–648, 2019.

[135] B. S. Schuster, G. L. Dignon, W. S. Tang, F. M. Kelley, A. K. Ranganath, C. N. Jahnke, A. G. Simpkins, R. M. Regy, D. A. Hammer, M. C. Good, *et al.*, "Identifying sequence perturbations to an intrinsically disordered protein that determine its phase-separation behavior," *Proceedings of the National Academy of Sciences*, vol. 117, no. 21, pp. 11 421–11 431, 2020.

[136] A. E. Conicella, G. L. Dignon, G. H. Zerze, H. B. Schmidt, A. M. D'Ordine, Y. C. Kim, R. Rohatgi, Y. M. Ayala, J. Mittal, and N. L. Fawzi, "Tdp-43 α-helical structure tunes liquid–liquid phase separation and function," *Proceedings of the National Academy of Sciences*, vol. 117, no. 11, pp. 5883–5894, 2020.

[137] G. Krainer, T. J. Welsh, J. A. Joseph, J. R. Espinosa, S. Wittmann, E. de Csilléry, A. Sridhar, Z. Toprakcioglu, G. Gudiškytė, M. A. Czekalska, *et al.*, "Reentrant liquid condensate phase of proteins is stabilized by hydrophobic and non-ionic interactions," *Nature communications*, vol. 12, no. 1, p. 1085, 2021.

[138] J. W. Pitera and J. D. Chodera, "On the use of experimental observations to bias simulated ensembles," *Journal of chemical theory and computation*, vol. 8, no. 10, pp. 3445–3451, 2012.

[139] B. Roux and J. Weare, "On the statistical equivalence of restrained-ensemble simulations with the maximum entropy method," *The Journal of chemical physics*, vol. 138, no. 8, 2013.

[140] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, "Funnels, pathways, and the energy landscape of protein folding: A synthesis," *Proteins: Structure, Function, and Bioinformatics*, vol. 21, no. 3, pp. 167–195, 1995.

[141] A. P. Latham and B. Zhang, "Maximum entropy optimized force field for intrinsically disordered proteins," *Journal of chemical theory and computation*, vol. 16, no. 1, pp. 773–781, 2019.

[142] A. P. Latham and B. Zhang, "Improving coarse-grained protein force fields with small-angle x-ray scattering data," *The Journal of Physical Chemistry B*, vol. 123, no. 5, pp. 1026–1034, 2019.

[143] A. P. Latham and B. Zhang, "Unifying coarse-grained force fields for folded and disordered proteins," *Current opinion in structural biology*, vol. 72, pp. 63–70, 2022.

[144] R. Regmi, S. Srinivasan, A. P. Latham, V. Kukshal, W. Cui, B. Zhang, R. Bose, and G. S. Schlau-Cohen, "Phosphorylation-dependent conformations of the disordered carboxyl-terminus domain in the epidermal growth factor receptor," *The journal of physical chemistry letters*, vol. 11, no. 23, pp. 10 037–10 044, 2020.

[145] A. Savelyev and G. A. Papoian, “Chemically accurate coarse graining of double-stranded dna,” *Proceedings of the National Academy of Sciences*, vol. 107, no. 47, pp. 20 340–20 345, 2010.

[146] D. M. Hinckley, G. S. Freeman, J. K. Whitmer, and J. J. De Pablo, “An experimentally-informed coarse-grained 3-site-per-nucleotide model of dna: Structure, thermodynamics, and dynamics of hybridization,” *The Journal of chemical physics*, vol. 139, no. 14, 10B604_1, 2013.

[147] W. Lu, C. Bueno, N. P. Schafer, J. Moller, S. Jin, X. Chen, M. Chen, X. Gu, A. Davtyan, J. J. de Pablo, *et al.*, “Openawsem with open3spn2: A fast, flexible, and accessible framework for large-scale coarse-grained biomolecular simulations,” *PLoS computational biology*, vol. 17, no. 2, e1008308, 2021.

[148] J. Lequieu, A. Córdoba, D. C. Schwartz, and J. J. de Pablo, “Tension-dependent free energies of nucleosome unwrapping,” *ACS central science*, vol. 2, no. 9, pp. 660–666, 2016.

[149] C. Tan, T. Terakawa, and S. Takada, “Dynamic coupling among protein binding, sliding, and dna bending revealed by molecular dynamics,” *Journal of the American Chemical Society*, vol. 138, no. 27, pp. 8512–8522, 2016.

[150] T. Parsons and B. Zhang, “Critical role of histone tail entropy in nucleosome unwinding,” *The Journal of chemical physics*, vol. 150, no. 18, 2019.

[151] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, “Pytorch: An imperative style, high-performance deep learning library,” *Advances in neural information processing systems*, vol. 32, 2019.

[152] Y. Li and Y. Zhang, “Remo: A new protocol to refine full atomic protein models from c-alpha traces by optimizing hydrogen-bonding networks,” *Proteins: Structure, Function, and Bioinformatics*, vol. 76, no. 3, pp. 665–676, 2009.

[153] G. L. Dignon, W. Zheng, R. B. Best, Y. C. Kim, and J. Mittal, “Relation between single-molecule properties and phase behavior of intrinsically disordered proteins,” *Proceedings of the National Academy of Sciences*, vol. 115, no. 40, pp. 9929–9934, 2018.

[154] R. M. Regy, W. Zheng, and J. Mittal, “Using a sequence-specific coarse-grained model for studying protein liquid–liquid phase separation,” in *Methods in enzymology*, vol. 646, Elsevier, 2021, pp. 1–17.

[155] J. A. Anderson, J. Glaser, and S. C. Glotzer, “Hoomd-blue: A python package for high-performance molecular dynamics and hard particle monte carlo simulations,” *Computational Materials Science*, vol. 173, p. 109 363, 2020.

[156] D. Sehnal, S. Bittrich, M. Deshpande, R. Svobodová, K. Berka, V. Bazgier, S. Velankar, S. K. Burley, J. Koča, and A. S. Rose, “Mol* viewer: Modern web app for 3d visualization and analysis of large biomolecular structures,” *Nucleic Acids Research*, vol. 49, no. W1, W431–W437, 2021.

[157] C. Maison and G. Almouzni, “Hp1 and the dynamics of heterochromatin maintenance,” *Nature reviews Molecular cell biology*, vol. 5, no. 4, pp. 296–305, 2004.

[158] T. J. Nott, E. Petsalaki, P. Farber, D. Jervis, E. Fussner, A. Plochowietz, T. D. Craggs, D. P. Bazett-Jones, T. Pawson, J. D. Forman-Kay, *et al.*, “Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles,” *Molecular cell*, vol. 57, no. 5, pp. 936–947, 2015.

[159] K. A. Burke, A. M. Janke, C. L. Rhine, and N. L. Fawzi, “Residue-by-residue view of in vitro fus granules that bind the c-terminal domain of rna polymerase ii,” *Molecular cell*, vol. 60, no. 2, pp. 231–241, 2015.

[160] J. Huang, S. Rauscher, G. Nawrocki, T. Ran, M. Feig, B. L. De Groot, H. Grubmüller, and A. D. MacKerell Jr, “Charmm36m: An improved force field for folded and intrinsically disordered proteins,” *Nature methods*, vol. 14, no. 1, pp. 71–73, 2017.

[161] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, “Comparison of simple potential functions for simulating liquid water,” *The Journal of chemical physics*, vol. 79, no. 2, pp. 926–935, 1983.

[162] Y. Itoh, E. J. Woods, K. Minami, K. Maeshima, and R. Collepardo-Guevara, “Liquid-like chromatin in the cell: What can we learn from imaging and computational modeling?” *Current Opinion in Structural Biology*, vol. 71, pp. 123–135, 2021.

[163] K. Luger, M. L. Dechassa, and D. J. Tremethick, “New insights into nucleosome and chromatin structure: An ordered state or a disordered affair?” *Nature reviews Molecular cell biology*, vol. 13, no. 7, pp. 436–447, 2012.

[164] T. Schlick, J. Hayes, and S. Grigoryev, “Toward convergence of experimental studies and theoretical modeling of the chromatin fiber,” *Journal of Biological Chemistry*, vol. 287, no. 8, pp. 5183–5191, 2012.

[165] J. D. McGhee, J. M. Nickol, G. Felsenfeld, and D. C. Rau, “Higher order structure of chromatin: Orientation of nucleosomes within the 30 nm chromatin solenoid is independent of species and spacer length,” *Cell*, vol. 33, no. 3, pp. 831–841, 1983.

[166] C. Woodcock, L.-L. Frado, and J. Rattner, “The higher-order structure of chromatin: Evidence for a helical ribbon arrangement.,” *The Journal of cell biology*, vol. 99, no. 1, pp. 42–52, 1984.

[167] J. Widom and A. Klug, “Structure of the 3000Å chromatin filament: X-ray diffraction from oriented samples,” *Cell*, vol. 43, no. 1, pp. 207–213, 1985.

[168] S. Williams, B. Athey, L. Muglia, R. Schappe, A. Gough, and J. Langmore, “Chromatin fibers are left-handed double helices with diameter and mass per unit length that depend on linker length,” *Biophysical journal*, vol. 49, no. 1, pp. 233–248, 1986.

[169] P. Lowary and J. Widom, “New dna sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning,” *Journal of molecular biology*, vol. 276, no. 1, pp. 19–42, 1998.

[170] P. J. Robinson, L. Fairall, V. A. Huynh, and D. Rhodes, "Em measurements define the dimensions of the "30-nm" chromatin fiber: Evidence for a compact, interdigitated structure," *Proceedings of the National Academy of Sciences*, vol. 103, no. 17, pp. 6506–6511, 2006.

[171] A. Routh, S. Sandin, and D. Rhodes, "Nucleosome repeat length and linker histone stoichiometry determine chromatin fiber structure," *Proceedings of the National Academy of Sciences*, vol. 105, no. 26, pp. 8872–8877, 2008.

[172] Y. Takizawa, C.-H. Ho, H. Tachiwana, H. Matsunami, W. Kobayashi, M. Suzuki, Y. Arimura, T. Hori, T. Fukagawa, M. D. Ohi, *et al.*, "Cryo-em structures of centromeric tri-nucleosomes containing a central cenp-a nucleosome," *Structure*, vol. 28, no. 1, pp. 44–53, 2020.

[173] K. Zhou, M. Gebala, D. Woods, K. Sundararajan, G. Edwards, D. Krzizike, J. Wereszczynski, A. F. Straight, and K. Luger, "Cenp-n promotes the compaction of centromeric chromatin," *Nature Structural & Molecular Biology*, vol. 29, no. 4, pp. 403–413, 2022.

[174] T.-H. S. Hsieh, A. Weiner, B. Lajoie, J. Dekker, N. Friedman, and O. J. Rando, "Mapping nucleosome resolution chromosome folding in yeast by micro-c," *Cell*, vol. 162, no. 1, pp. 108–119, 2015.

[175] M. E. Fisher and A. B. Kolomeisky, "The force exerted by a molecular motor," *Proceedings of the National Academy of Sciences*, vol. 96, no. 12, pp. 6597–6602, 1999.

[176] Z. Jiang and B. Zhang, "Theory of active chromatin remodeling," *Physical review letters*, vol. 123, no. 20, p. 208 102, 2019.

[177] C. Uhler and G. Shivashankar, "Regulation of genome organization and gene expression by nuclear mechanotransduction," *Nature reviews Molecular cell biology*, vol. 18, no. 12, pp. 717–727, 2017.

[178] A. Tajik, Y. Zhang, F. Wei, J. Sun, Q. Jia, W. Zhou, R. Singh, N. Khanna, A. S. Belmont, and N. Wang, "Transcription upregulation via force-induced direct stretching of chromatin," *Nature materials*, vol. 15, no. 12, pp. 1287–1296, 2016.

[179] C. Y. Zhou, S. L. Johnson, N. I. Gamarra, and G. J. Narlikar, "Mechanisms of atp-dependent chromatin remodeling motors," *Annual review of biophysics*, vol. 45, pp. 153–181, 2016.

[180] B. Van Steensel and A. S. Belmont, "Lamina-associated domains: Links with chromosome architecture, heterochromatin, and gene repression," *Cell*, vol. 169, no. 5, pp. 780–791, 2017.

[181] Y. Chen, Y. Zhang, Y. Wang, L. Zhang, E. K. Brinkman, S. A. Adam, R. Goldman, B. Van Steensel, J. Ma, and A. S. Belmont, "Mapping 3d genome organization relative to nuclear compartments using tsa-seq as a cytological ruler," *Journal of Cell Biology*, vol. 217, no. 11, pp. 4025–4048, 2018.

[182] S. A. Quinodoz, N. Ollikainen, B. Tabak, A. Palla, J. M. Schmidt, E. Detmar, M. M. Lai, A. A. Shishkin, P. Bhat, Y. Takei, *et al.*, "Higher-order inter-chromosomal hubs shape 3d genome organization in the nucleus," *Cell*, vol. 174, no. 3, pp. 744–757, 2018.

[183] S. M. Schreiner, P. K. Koo, Y. Zhao, S. G. Mochrie, and M. C. King, “The tethering of chromatin to the nuclear envelope supports nuclear mechanics,” *Nature communications*, vol. 6, no. 1, p. 7159, 2015.

[184] Y. Shin, Y.-C. Chang, D. S. Lee, J. Berry, D. W. Sanders, P. Ronceray, N. S. Wingreen, M. Haataja, and C. P. Brangwynne, “Liquid nuclear condensates mechanically sense and restructure the genome,” *Cell*, vol. 175, no. 6, pp. 1481–1491, 2018.

[185] B. Fierz and M. G. Poirier, “Biophysics of chromatin dynamics,” *Annual review of biophysics*, vol. 48, pp. 321–345, 2019.

[186] S. Baldi, P. Korber, and P. B. Becker, “Beads on a string-nucleosome array arrangements and folding of the chromatin fiber,” *Nature structural & molecular biology*, vol. 27, no. 2, pp. 109–118, 2020.

[187] Y. Cui and C. Bustamante, “Pulling a single chromatin fiber reveals the forces that maintain its higher-order structure,” *Proceedings of the National Academy of Sciences*, vol. 97, no. 1, pp. 127–132, 2000.

[188] M. Kruithof, F.-T. Chien, A. Routh, C. Logie, D. Rhodes, and J. Van Noort, “Single-molecule force spectroscopy reveals a highly compliant helical folding for the 30-nm chromatin fiber,” *Nature structural & molecular biology*, vol. 16, no. 5, pp. 534–540, 2009.

[189] W. Li, P. Chen, J. Yu, L. Dong, D. Liang, J. Feng, J. Yan, P.-Y. Wang, Q. Li, Z. Zhang, *et al.*, “Fact remodels the tetranucleosomal unit of chromatin fibers for gene transcription,” *Molecular cell*, vol. 64, no. 1, pp. 120–133, 2016.

[190] H. Meng, K. Andresen, and J. Van Noort, “Quantitative analysis of single-molecule force spectroscopy on folded chromatin fibers,” *Nucleic acids research*, vol. 43, no. 7, pp. 3578–3590, 2015.

[191] T. Ha, T. Enderle, D. Ogletree, D. S. Chemla, P. R. Selvin, and S. Weiss, “Probing the interaction between two single molecules: Fluorescence resonance energy transfer between a single donor and a single acceptor.,” *Proceedings of the National Academy of Sciences*, vol. 93, no. 13, pp. 6264–6268, 1996.

[192] R. Roy, S. Hohng, and T. Ha, “A practical guide to single-molecule fret,” *Nature methods*, vol. 5, no. 6, pp. 507–516, 2008.

[193] M. G. Poirier, E. Oh, H. S. Tims, and J. Widom, “Dynamics and function of compact nucleosome arrays,” *Nature structural & molecular biology*, vol. 16, no. 9, pp. 938–944, 2009.

[194] J. M. Victor, J. Zlatanova, M. Barbi, and J. Mozziconacci, “Pulling chromatin apart: Unstacking or unwrapping?” *BMC biophysics*, vol. 5, pp. 1–6, 2012.

[195] B. E. de Jong, T. B. Brouwer, A. Kaczmarczyk, B. Visscher, and J. van Noort, “Rigid basepair monte carlo simulations of one-start and two-start chromatin fiber unfolding by force,” *Biophysical Journal*, vol. 115, no. 10, pp. 1848–1859, 2018.

[196] D. Norouzi and V. B. Zhurkin, “Dynamics of chromatin fibers: Comparison of monte carlo simulations with force spectroscopy,” *Biophysical journal*, vol. 115, no. 9, pp. 1644–1655, 2018.

[197] N. Kepper, R. Ettig, R. Stehr, S. Marnach, G. Wedemann, and K. Rippe, "Force spectroscopy of chromatin fibers: Extracting energetics and structural information from monte carlo simulations," *Biopolymers*, vol. 95, no. 7, pp. 435–447, 2011.

[198] O. Perišić and T. Schlick, "Computational strategies to address chromatin structure problems," *Physical biology*, vol. 13, no. 3, p. 035 006, 2016.

[199] W. Alvarado, J. Moller, A. L. Ferguson, and J. J. de Pablo, "Tetranucleosome interactions drive chromatin folding," *ACS Central Science*, vol. 7, no. 6, pp. 1019–1027, 2021.

[200] R. Collepardo-Guevara and T. Schlick, "The effect of linker histone's nucleosome binding affinity on chromatin unfolding mechanisms," *Biophysical journal*, vol. 101, no. 7, pp. 1670–1680, 2011.

[201] E. F. Koslover, C. J. Fuller, A. F. Straight, and A. J. Spakowitz, "Local geometry and elasticity in compact chromatin structure," *Biophysical journal*, vol. 99, no. 12, pp. 3941–3950, 2010.

[202] A. Kaczmarczyk, H. Meng, O. Ordu, J. v. Noort, and N. H. Dekker, "Chromatin fibers stabilize nucleosomes under torsional stress," *Nature communications*, vol. 11, no. 1, p. 126, 2020.

[203] M. A. Ricci, C. Manzo, M. F. García-Parajo, M. Lakadamyali, and M. P. Cosma, "Chromatin fibers are formed by heterogeneous groups of nucleosomes in vivo," *Cell*, vol. 160, no. 6, pp. 1145–1158, 2015.

[204] C. Clementi, H. Nymeyer, and J. N. Onuchic, "Topological and energetic factors: What determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? an investigation for small globular proteins," *Journal of molecular biology*, vol. 298, no. 5, pp. 937–953, 2000.

[205] S. Miyazawa and R. L. Jernigan, "Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation," *Macromolecules*, vol. 18, no. 3, pp. 534–552, 1985.

[206] G. S. Freeman, D. M. Hinckley, J. P. Lequieu, J. K. Whitmer, and J. J. De Pablo, "Coarse-grained modeling of dna curvature," *The Journal of chemical physics*, vol. 141, no. 16, 2014.

[207] J. Moller, J. Lequieu, and J. J. de Pablo, "The free energy landscape of internucleosome interactions and its relation to chromatin fiber structure," *ACS central science*, vol. 5, no. 2, pp. 341–348, 2019.

[208] S. Plimpton, "Fast parallel algorithms for short-range molecular dynamics," *Journal of computational physics*, vol. 117, no. 1, pp. 1–19, 1995.

[209] W. Shinoda, M. Shiga, and M. Mikami, "Rapid estimation of elastic constants by molecular dynamics simulation under constant stress," *Physical Review B*, vol. 69, no. 13, p. 134 103, 2004.

[210] V. Böhm, A. R. Hieb, A. J. Andrews, A. Gansen, A. Rocker, K. Tóth, K. Luger, and J. Langowski, "Nucleosome accessibility governed by the dimer/tetramer interface," *Nucleic acids research*, vol. 39, no. 8, pp. 3093–3102, 2011.

[211] B. Zhang, W. Zheng, G. A. Papoian, and P. G. Wolynes, "Exploring the free energy landscape of nucleosomes," *Journal of the American Chemical Society*, vol. 138, no. 26, pp. 8126–8133, 2016.

[212] H. Kenzaki and S. Takada, "Linker dna length is a key to tri-nucleosome folding," *Journal of Molecular Biology*, vol. 433, no. 6, p. 166 792, 2021.

[213] B. D. Brower-Toland, C. L. Smith, R. C. Yeh, J. T. Lis, C. L. Peterson, and M. D. Wang, "Mechanical disruption of individual nucleosomes reveals a reversible multistage release of dna," *Proceedings of the National Academy of Sciences*, vol. 99, no. 4, pp. 1960–1965, 2002.

[214] M. Ohno, T. Ando, D. G. Priest, V. Kumar, Y. Yoshida, and Y. Taniguchi, "Subnucleosomal genome structure reveals distinct nucleosome folding motifs," *Cell*, vol. 176, no. 3, pp. 520–534, 2019.

[215] S. Cai, D. Böck, M. Pilhofer, and L. Gan, "The in situ structures of mono-, di-, and trinucleosomes in human heterochromatin," *Molecular biology of the cell*, vol. 29, no. 20, pp. 2450–2457, 2018.

[216] R. Collepardo-Guevara and T. Schlick, "Crucial role of dynamic linker histone binding and divalent ions for DNA accessibility and gene regulation revealed by mesoscale modeling of oligonucleosomes," *Nucleic Acids Res.*, vol. 40, no. 18, pp. 8803–8817, 2012.

[217] N. Krietenstein, S. Abraham, S. V. Venev, N. Abdennur, J. Gibcus, T.-H. S. Hsieh, K. M. Parsi, L. Yang, R. Maehr, L. A. Mirny, *et al.*, "Ultrastructural details of mammalian chromosome architecture," *Molecular cell*, vol. 78, no. 3, pp. 554–565, 2020.

[218] G. J. Narlikar, R. Sundaramoorthy, and T. Owen-Hughes, "Mechanisms and functions of ATP-dependent chromatin-remodeling enzymes," *Cell*, vol. 154, no. 3, pp. 490–503, 2013.

[219] E. A. Galburt, S. W. Grill, A. Wiedmann, L. Lubkowska, J. Choy, E. Nogales, M. Kashlev, and C. Bustamante, "Backtracking determines the force sensitivity of RNAP II in a factor-dependent manner," *Nature*, vol. 446, no. 7137, pp. 820–823, 2007.

[220] E. A. Galburt, S. W. Grill, and C. Bustamante, "Single molecule transcription elongation," *Methods*, vol. 48, no. 4, pp. 323–332, 2009.

[221] G. Lia, E. Praly, H. Ferreira, C. Stockdale, Y. C. Tse-Dinh, D. Dunlap, V. Croquette, D. Bensimon, and T. Owen-Hughes, "Direct observation of DNA distortion by the RSC complex," *Mol. Cell*, vol. 21, no. 3, pp. 417–425, 2006.

[222] J. Widom, "Physicochemical studies of the folding of the 100 Å nucleosome filament into the 300 Å filament: Cation dependence," *J. Mol. Biol.*, vol. 190, no. 3, pp. 411–424, 1986.

[223] P. Giannasca, R. Horowitz, and C. Woodcock, "Transitions between in situ and isolated chromatin," *J. Cell Sci.*, vol. 105, no. 2, pp. 551–561, 1993.

[224] R. A. Horowitz, D. Agard, J. Sedat, and C. Woodcock, "The three-dimensional architecture of chromatin in situ: Electron tomography reveals fibers composed of a continuously variable zig-zag nucleosomal ribbon.," *J. Cell Biol.*, vol. 125, no. 1, pp. 1–10, 1994.

[225] J. Zlatanova, S. H. Leuba, G. Yang, C. Bustamante, and K. Van Holde, "Linker DNA accessibility in chromatin fibers of different conformations: A reevaluation," *Proc. Natl. Acad. Sci.*, vol. 91, no. 12, pp. 5277–5280, 1994.

[226] C. Woodcock and R. Horowitz, "Chromatin organization re-viewed," *Trends Cell Biol.*, vol. 5, no. 7, pp. 272–277, 1995.

[227] S. A. Grigoryev, J. Bednar, and C. L. Woodcock, "MENT, a heterochromatin protein that mediates higher order chromatin folding, is a new serpin family member," *J. Biol. Chem.*, vol. 274, no. 9, pp. 5626–5636, 1999.

[228] S. A. Grigoryev, "Keeping fingers crossed: Heterochromatin spreading through interdigitation of nucleosome arrays," *FEBS Lett.*, vol. 564, no. 1-2, pp. 4–8, 2004.

[229] E. Zaccarelli, "Colloidal gels: Equilibrium and non-equilibrium routes," *Journal of Physics: Condensed Matter*, vol. 19, no. 32, p. 323 101, 2007.

[230] N. Khanna, Y. Zhang, J. S. Lucas, O. K. Dudko, and C. Murre, "Chromosome dynamics near the sol-gel phase transition dictate the timing of remote genomic interactions," *Nature communications*, vol. 10, no. 1, p. 2771, 2019.

[231] I. Eshghi, J. A. Eaton, and A. Zidovska, "Interphase chromatin undergoes a local sol-gel transition upon cell differentiation," *Physical review letters*, vol. 126, no. 22, p. 228 101, 2021.

[232] M. Shogren-Knaak, H. Ishii, J.-M. Sun, M. J. Pazin, J. R. Davie, and C. L. Peterson, "Histone h4-k16 acetylation controls chromatin structure and protein interactions," *Science*, vol. 311, no. 5762, pp. 844–847, 2006.

[233] P. J. Robinson, W. An, A. Routh, F. Martino, L. Chapman, R. G. Roeder, and D. Rhodes, "30 nm chromatin fibre decompaction requires both h4-k16 acetylation and linker histone eviction," *Journal of molecular biology*, vol. 381, no. 4, pp. 816–825, 2008.

[234] D. A. Potoyan and G. A. Papoian, "Regulation of the h4 tail binding and folding landscapes via lys-16 acetylation," *Proceedings of the National Academy of Sciences*, vol. 109, no. 44, pp. 17 857–17 862, 2012.

[235] G. D. Bascom, C. G. Myers, and T. Schlick, "Mesoscale modeling reveals formation of an epigenetically driven HOXC gene hub," *Proc. Natl. Acad. Sci.*, vol. 116, no. 11, pp. 4955–4962, 2019.

[236] S. G. Swygert, D. Lin, S. Portillo-Ledesma, P.-Y. Lin, D. R. Hunt, C.-F. Kao, T. Schlick, W. S. Noble, and T. Tsukiyama, "Local chromatin fiber folding represses transcription and loop extrusion in quiescent cells," *Elife*, vol. 10, e72062, 2021.

[237] A. Luque, G. Ozer, and T. Schlick, "Correlation among dna linker length, linker histone concentration, and histone tails in chromatin," *Biophysical journal*, vol. 110, no. 11, pp. 2309–2319, 2016.

[238] T. Nikitina, D. Norouzi, S. A. Grigoryev, and V. B. Zhurkin, “Dna topology in chromatin is defined by nucleosome spacing,” *Science advances*, vol. 3, no. 10, e1700957, 2017.

[239] D. Norouzi and V. B. Zhurkin, “Topological polymorphism of the two-start chromatin fiber,” *Biophysical journal*, vol. 108, no. 10, pp. 2591–2600, 2015.

[240] S. Portillo-Ledesma, L. H. Tsao, M. Wagley, M. Lakadamyali, M. P. Cosma, and T. Schlick, “Nucleosome clutches are regulated by chromatin internal parameters,” *J. Mol. Biol.*, vol. 433, no. 6, p. 166 701, 2021.

[241] K. Struhl and E. Segal, “Determinants of nucleosome positioning,” *Nat. Struct. Mol. Biol.*, vol. 20, no. 3, pp. 267–273, 2013.

[242] B.-R. Zhou, J. Jiang, H. Feng, R. Ghirlando, T. S. Xiao, and Y. Bai, “Structural mechanisms of nucleosome recognition by linker histones,” *Mol. Cell*, vol. 59, no. 4, pp. 628–638, 2015.

[243] B.-R. Zhou, H. Feng, H. Kato, L. Dai, Y. Yang, Y. Zhou, and Y. Bai, “Structural insights into the histone H1-nucleosome complex,” *Proc. Natl. Acad. Sci.*, vol. 110, no. 48, pp. 19 390–19 395, 2013.

[244] I. Garcia-Saez, H. Menoni, R. Boopathi, M. S. Shukla, L. Soueidan, M. Noirclerc-Savoye, A. Le Roy, D. A. Skoufias, J. Bednar, A. Hamiche, *et al.*, “Structure of an H1-bound 6-nucleosome array reveals an untwisted two-start chromatin fiber conformation,” *Mol. Cell*, vol. 72, no. 5, pp. 902–915, 2018.

[245] P. A. Gómez-García, S. Portillo-Ledesma, M. V. Neguembor, M. Pesaresi, W. Oweis, T. Rohrlich, S. Wieser, E. Meshorer, T. Schlick, M. P. Cosma, *et al.*, “Mesoscale modeling and single-nucleosome tracking reveal remodeling of clutch folding and dynamics in stem cell differentiation,” *Cell Rep.*, vol. 34, no. 2, p. 108 614, 2021.

[246] M. J. Rowley and V. G. Corces, “Organizational principles of 3d genome architecture,” *Nature Reviews Genetics*, vol. 19, no. 12, pp. 789–800, 2018.

[247] L. Mirny and J. Dekker, “Mechanisms of chromosome folding and nuclear organization: Their interplay and open questions,” *Cold Spring Harbor perspectives in biology*, vol. 14, no. 7, a040147, 2022.

[248] K. Maeshima, S. Hihara, and M. Eltsov, “Chromatin structure: Does the 30-nm fibre exist in vivo?” *Current opinion in cell biology*, vol. 22, no. 3, pp. 291–297, 2010.

[249] Y. Nishino, M. Eltsov, Y. Joti, K. Ito, H. Takata, Y. Takahashi, S. Hihara, A. S. Frangakis, N. Imamoto, T. Ishikawa, *et al.*, “Human mitotic chromosomes consist predominantly of irregularly folded nucleosome fibres without a 30-nm chromatin structure,” *The EMBO journal*, vol. 31, no. 7, pp. 1644–1653, 2012.

[250] W. Alvarado, V. Agrawal, W. S. Li, V. P. Dravid, V. Backman, J. J. de Pablo, and A. L. Ferguson, “Denoising autoencoder trained on simulation-derived structures for noise reduction in chromatin scanning transmission electron microscopy,” *ACS Central Science*, 2023.

[251] K. Rippe, “Liquid–liquid phase separation in chromatin,” *Cold Spring Harbor perspectives in biology*, vol. 14, no. 2, a040683, 2022.

[252] A. G. Larson, D. Elnatan, M. M. Keenen, M. J. Trnka, J. B. Johnston, A. L. Burlingame, D. A. Agard, S. Redding, and G. J. Narlikar, "Liquid droplet formation by hp1α suggests a role for phase separation in heterochromatin," *Nature*, vol. 547, no. 7662, pp. 236–240, 2017.

[253] S. Sanulli, M. Trnka, V. Dharmarajan, R. Tibble, B. Pascal, A. Burlingame, P. Griffin, J. Gross, and G. Narlikar, "Hp1 reshapes nucleosome core to promote phase separation of heterochromatin," *Nature*, vol. 575, no. 7782, pp. 390–394, 2019.

[254] J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren, "Topological domains in mammalian genomes identified by analysis of chromatin interactions," *Nature*, vol. 485, no. 7398, pp. 376–380, 2012.

[255] S. S. Rao, M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, *et al.*, "A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping," *Cell*, vol. 159, no. 7, pp. 1665–1680, 2014.

[256] W. Schwarzer, N. Abdennur, A. Goloborodko, A. Pekowska, G. Fudenberg, Y. Loe-Mie, N. A. Fonseca, W. Huber, C. H. Haering, L. Mirny, *et al.*, "Two independent modes of chromatin organization revealed by cohesin removal," *Nature*, vol. 551, no. 7678, pp. 51–56, 2017.

[257] L. Hilbert, Y. Sato, K. Kuznetsova, T. Bianucci, H. Kimura, F. Jülicher, A. Honigmann, V. Zaburdaev, and N. L. Vastenhouw, "Transcription organizes euchromatin via microphase separation," *Nature communications*, vol. 12, no. 1, p. 1360, 2021.

[258] S. Park, M. Mitchener, H. Dao, T. Muir, and T. Ha, "Biophysical driving forces of heterochromatin organization," *Biophysical Journal*, vol. 121, no. 3, 159a, 2022.

[259] W. Li, P. G. Wolynes, and S. Takada, "Frustration, specific sequence dependence, and nonlinearity in large-amplitude fluctuations of allosteric proteins," *Proceedings of the National Academy of Sciences*, vol. 108, no. 9, pp. 3504–3509, 2011.

[260] F. Noé, C. Schütte, E. Vanden-Eijnden, L. Reich, and T. R. Weikl, "Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations," *Proceedings of the National Academy of Sciences*, vol. 106, no. 45, pp. 19 011–19 016, 2009.

[261] E. Vanden-Eijnden *et al.*, "Towards a theory of transition paths," *Journal of statistical physics*, vol. 123, no. 3, pp. 503–523, 2006.

[262] V. S. Pande, K. Beauchamp, and G. R. Bowman, "Everything you wanted to know about markov state models but were afraid to ask," *Methods*, vol. 52, no. 1, pp. 99–105, 2010.

[263] B. E. Husic and V. S. Pande, "Markov state models: From an art to a science," *Journal of the American Chemical Society*, vol. 140, no. 7, pp. 2386–2396, 2018.

[264] K. A. Konovalov, I. C. Unarta, S. Cao, E. C. Goonetilleke, and X. Huang, "Markov state models to study the functional dynamics of proteins in the wake of machine learning," *JACS Au*, vol. 1, no. 9, pp. 1330–1341, 2021.

[265] W. Wang, S. Cao, L. Zhu, and X. Huang, "Constructing markov state models to elucidate the functional conformational changes of complex biomolecules," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 8, no. 1, e1343, 2018.

[266] J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F. Noé, "Markov models of molecular kinetics: Generation and validation," *The Journal of chemical physics*, vol. 134, no. 17, 2011.

[267] Y. Qiu, M. S. O'Connor, M. Xue, B. Liu, and X. Huang, "An efficient path classification algorithm based on variational autoencoder to identify metastable path channels for complex conformational changes," *Journal of Chemical Theory and Computation*, vol. 19, no. 14, pp. 4728–4742, 2023.

[268] J. D. Chodera and F. Noé, "Markov state models of biomolecular conformational dynamics," *Current opinion in structural biology*, vol. 25, pp. 135–144, 2014.

[269] V. A. Voelz, G. R. Bowman, K. Beauchamp, and V. S. Pande, "Molecular simulation of ab initio protein folding for a millisecond folder ntl9 (1- 39)," *Journal of the American Chemical Society*, vol. 132, no. 5, pp. 1526–1528, 2010.

[270] D.-A. Silva, G. R. Bowman, A. Sosa-Peinado, and X. Huang, "A role for both conformational selection and induced fit in ligand binding by the lao protein," *PLoS computational biology*, vol. 7, no. 5, e1002054, 2011.

[271] D. Shukla, Y. Meng, B. Roux, and V. S. Pande, "Activation pathway of src kinase reveals intermediate states as targets for drug design," *Nature communications*, vol. 5, no. 1, p. 3397, 2014.

[272] B. Liu, Y. Qiu, E. C. Goonetilleke, and X. Huang, "Kinetic network models to study molecular self-assembly in the wake of machine learning," *MRS Bulletin*, vol. 47, no. 9, pp. 958–966, 2022.

[273] X. Zeng, B. Li, Q. Qiao, L. Zhu, Z.-Y. Lu, and X. Huang, "Elucidating dominant pathways of the nano-particle self-assembly process," *Physical Chemistry Chemical Physics*, vol. 18, no. 34, pp. 23 494–23 499, 2016.

[274] P. Metzner, C. Schütte, and E. Vanden-Eijnden, "Transition path theory for markov jump processes," *Multiscale Modeling & Simulation*, vol. 7, no. 3, pp. 1192–1219, 2009.

[275] G. Pérez-Hernández, F. Paul, T. Giorgino, G. De Fabritiis, and F. Noé, "Identification of slow molecular order parameters for markov model construction," *The Journal of chemical physics*, vol. 139, no. 1, 07B604_1, 2013.

[276] F. Noé and C. Clementi, "Kinetic distance and kinetic maps from molecular dynamics simulation," *Journal of chemical theory and computation*, vol. 11, no. 10, pp. 5002–5011, 2015.

[277] Y. Naritomi and S. Fuchigami, "Slow dynamics of a protein backbone in molecular dynamics simulation revealed by time-structure based independent component analysis," *The Journal of Chemical Physics*, vol. 139, no. 21, 12B605_1, 2013.

[278] C. R. Schwantes and V. S. Pande, "Improvements in markov state model construction reveal many non-native interactions in the folding of ntl9," *Journal of chemical theory and computation*, vol. 9, no. 4, pp. 2000–2009, 2013.

[279] R. T. McGibbon and V. S. Pande, "Variational cross-validation of slow dynamical modes in molecular kinetics," *The Journal of chemical physics*, vol. 142, no. 12, 03B621_1, 2015.

[280] A. K.-h. Yik, Y. Qiu, I. C. Unarta, S. Cao, and X. Huang, "A step-by-step guide on how to construct quasi-markov state models to study functional conformational changes of biological macromolecules," in *A Practical Guide to Recent Advances in Multiscale Modeling and Simulation of Biomolecules*, AIP Publishing LLC Melville, New York, 2023, pp. 10–1.

[281] M. Chen, M. A. Cuendet, and M. E. Tuckerman, "Heating and flooding: A unified approach for rapid generation of free energy surfaces," *The Journal of Chemical Physics*, vol. 137, no. 2, p. 024 102, 2012.

[282] L. Meng, F. K. Sheong, X. Zeng, L. Zhu, and X. Huang, "Path lumping: An efficient algorithm to identify metastable path channels for conformational dynamics of multibody systems," *The Journal of Chemical Physics*, vol. 147, no. 4, p. 044 112, 2017.

[283] T. S. Lewis, V. Sokolova, H. Jung, H. Ng, and D. Tan, "Structural basis of chromatin regulation by histone variant h2a. z," *Nucleic Acids Research*, vol. 49, no. 19, pp. 11 379–11 391, 2021.

[284] L. N. Voong, L. Xi, A. C. Sebeson, B. Xiong, J.-P. Wang, and X. Wang, "Insights into nucleosome organization in mouse embryonic stem cells through chemical mapping," *Cell*, vol. 167, no. 6, pp. 1555–1570, 2016.

[285] S. J. Correll, M. H. Schubert, and S. A. Grigoryev, "Short nucleosome repeats impose rotational modulations on chromatin fibre folding," *The EMBO journal*, vol. 31, no. 10, pp. 2416–2426, 2012.

[286] M. Dombrowski, M. Engeholm, C. Dienemann, S. Dodonova, and P. Cramer, "Histone h1 binding to nucleosome arrays depends on linker dna length and trajectory," *Nature structural & molecular biology*, vol. 29, no. 5, pp. 493–501, 2022.

[287] C. R. Clapier, J. Iwasa, B. R. Cairns, and C. L. Peterson, "Mechanisms of action and regulation of atp-dependent chromatin-remodelling complexes," *Nature reviews Molecular cell biology*, vol. 18, no. 7, pp. 407–422, 2017.

[288] J. E. Henninger, O. Oksuz, K. Shrinivas, I. Sagi, G. LeRoy, M. M. Zheng, J. O. Andrews, A. V. Zamudio, C. Lazaris, N. M. Hannett, *et al.*, "Rna-mediated feedback control of transcriptional condensates," *Cell*, vol. 184, no. 1, pp. 207–225, 2021.

[289] N. Galvanetto, M. T. Ivanović, A. Chowdhury, A. Sottini, M. F. Nüesch, D. Nettels, R. B. Best, and B. Schuler, "Extreme dynamics in a biomolecular condensate," *Nature*, vol. 619, no. 7971, pp. 876–883, 2023.

[290] M. Doi, *Soft matter physics.* Oxford University Press, USA, 2013.

[291] T. Busby III and T. Misteli, "Building chromatin from the ground up," *Nature Genetics*, pp. 1–2, 2024.

[292] T. Sun, V. Minhas, A. Mirzoev, N. Korolev, A. P. Lyubartsev, and L. Nordenskiöld, "A bottom-up coarse-grained model for nucleosome–nucleosome interactions with explicit ions," *Journal of Chemical Theory and Computation*, vol. 18, no. 6, pp. 3948–3960, 2022.

[293] T. Sun, N. Korolev, V. Minhas, A. Mirzoev, A. P. Lyubartsev, and L. Nordenskiöld, "Multiscale modeling reveals the ion-mediated phase separation of nucleosome core particles," *Biophysical Journal*, 2023.

[294] X. Lin and B. Zhang, "Explicit ion modeling predicts physicochemical interactions for chromatin organization," *Elife*, vol. 12, RP90073, 2024.

[295] M. R. Shirts and J. D. Chodera, "Statistically optimal analysis of samples from multiple equilibrium states," *The Journal of chemical physics*, vol. 129, no. 12, 2008.

[296] R. W. Zwanzig, "High-temperature equation of state by a perturbation method. i. nonpolar gases," *The Journal of Chemical Physics*, vol. 22, no. 8, pp. 1420–1426, 1954.

[297] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers: Original Research on Biomolecules*, vol. 22, no. 12, pp. 2577–2637, 1983.

[298] R. T. McGibbon, K. A. Beauchamp, M. P. Harrigan, C. Klein, J. M. Swails, C. X. Hernández, C. R. Schwantes, L.-P. Wang, T. J. Lane, and V. S. Pande, "Mdtraj: A modern open library for the analysis of molecular dynamics trajectories," *Biophysical journal*, vol. 109, no. 8, pp. 1528–1532, 2015.

[299] J. K. Noel, P. C. Whitford, and J. N. Onuchic, "The shadow map: A general contact definition for capturing the dynamics of biomolecular folding and function," *The journal of physical chemistry B*, vol. 116, no. 29, pp. 8692–8702, 2012.

[300] M. Mirdita, K. Schütze, Y. Moriwaki, L. Heo, S. Ovchinnikov, and M. Steinegger, "Colabfold: Making protein folding accessible to all," *Nature methods*, vol. 19, no. 6, pp. 679–682, 2022.

[301] C. J. Wilson, W.-Y. Choy, and M. Karttunen, "Alphafold2: A role for disordered protein/region prediction?" *International journal of molecular sciences*, vol. 23, no. 9, p. 4591, 2022.

[302] C. W. Hopkins, S. Le Grand, R. C. Walker, and A. E. Roitberg, "Long-time-step molecular dynamics through hydrogen mass repartitioning," *Journal of chemical theory and computation*, vol. 11, no. 4, pp. 1864–1874, 2015.

[303] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, *et al.*, "Scipy 1.0: Fundamental algorithms for scientific computing in python," *Nature methods*, vol. 17, no. 3, pp. 261–272, 2020.

[304] L. H. Kapcha and P. J. Rossky, "A simple atomic-level hydrophobicity scale reveals protein interfacial structure," *Journal of molecular biology*, vol. 426, no. 2, pp. 484–498, 2014.

[305] X. Wang, S. Ramírez-Hinestrosa, J. Dobnikar, and D. Frenkel, "The lennard-jones potential: When (not) to use it," *Physical Chemistry Chemical Physics*, vol. 22, no. 19, pp. 10 624–10 633, 2020.

[306] N. Michaud-Agrawal, E. J. Denning, T. B. Woolf, and O. Beckstein, "Mdanalysis: A toolkit for the analysis of molecular dynamics simulations," *Journal of computational chemistry*, vol. 32, no. 10, pp. 2319–2327, 2011.

[307] R. J. Gowers, M. Linke, J. Barnoud, T. J. E. Reddy, M. N. Melo, S. L. Seyler, J. Domanski, D. L. Dotson, S. Buchoux, I. M. Kenney, *et al.*, "Mdanalysis: A python package for the rapid analysis of molecular dynamics simulations," Los Alamos National lab.(LANL), Los Alamos, NM (United States), Tech. Rep., 2019.

[308] G. Bussi, D. Donadio, and M. Parrinello, "Canonical sampling through velocity rescaling," *The Journal of chemical physics*, vol. 126, no. 1, p. 014 101, 2007.

[309] "Promoting transparency and reproducibility in enhanced molecular simulations," *Nature methods*, vol. 16, no. 8, pp. 670–673, 2019.

[310] M. Bonomi, D. Branduardi, G. Bussi, C. Camilloni, D. Provasi, P. Raiteri, D. Donadio, F. Marinelli, F. Pietrucci, R. A. Broglia, *et al.*, "Plumed: A portable plugin for free-energy calculations with molecular dynamics," *Computer Physics Communications*, vol. 180, no. 10, pp. 1961–1972, 2009.

[311] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman, "The weighted histogram analysis method for free-energy calculations on biomolecules. i. the method," *Journal of computational chemistry*, vol. 13, no. 8, pp. 1011–1021, 1992.

[312] X.-J. Lu and W. K. Olson, "3dna: A software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures," *Nucleic acids research*, vol. 31, no. 17, pp. 5108–5121, 2003.

[313] D. L. Theobald, "Rapid calculation of rmsds using a quaternion-based characteristic polynomial," *Acta Crystallographica Section A: Foundations of Crystallography*, vol. 61, no. 4, pp. 478–480, 2005.

[314] P. Liu, D. K. Agrafiotis, and D. L. Theobald, "Fast determination of the optimal rotational matrix for macromolecular superpositions," *Journal of computational chemistry*, vol. 31, no. 7, pp. 1561–1563, 2010.

[315] D. J. Wales, *Energy Landscapes* (Cambridge Molecular Science). Cambridge, UK ; New York: Cambridge University Press, 2003, ISBN: 978-0-521-81415-7.

[316] G. S. Freeman, J. P. Lequieu, D. M. Hinckley, J. K. Whitmer, and J. J. De Pablo, "Dna shape dominates sequence affinity in nucleosome formation," *Physical Review Letters*, vol. 113, no. 16, p. 168 101, 2014.

[317] J. J. Funke, P. Ketterer, C. Lieleg, S. Schunter, P. Korber, and H. Dietz, "Uncovering the forces between nucleosomes using dna origami," *Science advances*, vol. 2, no. 11, e1600974, 2016.

[318] R. Buning, W. Kropff, K. Martens, and J. van Noort, "Spfret reveals changes in nucleosome breathing by neighboring nucleosomes," *Journal of Physics: Condensed Matter*, vol. 27, no. 6, p. 064 103, 2015.

[319] M. R. Machado, E. E. Barrera, F. Klein, M. Sóñora, S. Silva, and S. Pantano, "The sirah 2.0 force field: Altius, fortius, citius," *Journal of Chemical Theory and Computation*, vol. 15, no. 4, pp. 2719–2733, 2019.

[320] D. C. Woods, F. Rodríguez-Ropero, and J. Wereszczynski, "The dynamic influence of linker histone saturation within the poly-nucleosome array," *Journal of molecular biology*, vol. 433, no. 10, p. 166 902, 2021.

[321] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[322] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, *et al.*, "Highly accurate protein structure prediction with alphafold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021.

[323] S. Mihardja, A. J. Spakowitz, Y. Zhang, and C. Bustamante, "Effect of force on mononucleosomal dynamics," *Proceedings of the National Academy of Sciences*, vol. 103, no. 43, pp. 15 871–15 876, 2006.

[324] S. Wei, S. J. Falk, B. E. Black, and T.-H. Lee, "A novel hybrid single molecule approach reveals spontaneous dna motion in the nucleosome," *Nucleic acids research*, vol. 43, no. 17, e111–e111, 2015.

[325] F.-T. Chien and T. Van Der Heijden, "Characterization of nucleosome unwrapping within chromatin fibers using magnetic tweezers," *Biophysical journal*, vol. 107, no. 2, pp. 373–383, 2014.

[326] G. J. Gemmen, R. Sim, K. A. Haushalter, P. C. Ke, J. T. Kadonaga, and D. E. Smith, "Forced unraveling of nucleosomes assembled on heterogeneous dna using core histones, nap-1, and acf," *Journal of molecular biology*, vol. 351, no. 1, pp. 89–99, 2005.

[327] L. Maragliano and E. Vanden-Eijnden, "A temperature accelerated method for sampling free energy and determining reaction pathways in rare events simulations," *Chemical physics letters*, vol. 426, no. 1-3, pp. 168–175, 2006.

[328] A. Laio and M. Parrinello, "Escaping free-energy minima," *Proceedings of the national academy of sciences*, vol. 99, no. 20, pp. 12 562–12 566, 2002.

[329] H. Kamberaj, R. Low, and M. Neal, "Time reversible and symplectic integrators for molecular dynamics simulations of rigid molecules," *The Journal of chemical physics*, vol. 122, no. 22, p. 224 114, 2005.

[330] A. B. de Oliveira Jr, V. G. Contessoto, A. Hassan, S. Byju, A. Wang, Y. Wang, E. Dodero-Rojas, U. Mohanty, J. K. Noel, J. N. Onuchic, *et al.*, "Smog 2 and opensmog: Extending the limits of structure-based models," *Protein Science*, vol. 31, no. 1, pp. 158–172, 2022.

[331] S. Röblitz and M. Weber, "Fuzzy spectral clustering by pcca+: Application to markov state models and data classification," *Advances in Data Analysis and Classification*, vol. 7, pp. 147–179, 2013.

[332] P. Deuflhard and M. Weber, "Robust perron cluster analysis in conformation dynamics," *Linear algebra and its applications*, vol. 398, pp. 161–184, 2005.

[333] K. A. Beauchamp, G. R. Bowman, T. J. Lane, L. Maibaum, I. S. Haque, and V. S. Pande, "Msmbuilder2: Modeling conformational dynamics on the picosecond to millisecond scale," *Journal of chemical theory and computation*, vol. 7, no. 10, pp. 3412–3419, 2011.

[334] F. Litzinger, L. Boninsegna, H. Wu, F. Nüske, R. Patel, R. Baraniuk, F. Noé, and C. Clementi, "Rapid calculation of molecular kinetics using compressed sensing," *Journal of Chemical Theory and Computation*, vol. 14, no. 5, pp. 2771–2783, 2018.

[335] E. W. Dijkstra, "A note on two problems in connexion with graphs," in *Edsger Wybe Dijkstra: His Life, Work, and Legacy*, 2022, pp. 287–290.

[336] E. Boattini, S. Marín-Aguilar, S. Mitra, G. Foffi, F. Smallenburg, and L. Filion, "Autonomously revealing hidden local structures in supercooled liquids," *Nature communications*, vol. 11, no. 1, p. 5479, 2020.

[337] R. Nainggolan, R. Perangin-angin, E. Simarmata, and A. F. Tarigan, "Improved the performance of the k-means cluster using the sum of squared error (sse) optimized by using the elbow method," in *Journal of Physics: Conference Series*, IOP Publishing, vol. 1361, 2019, p. 012 015.

[338] H. Belyadi and A. Haghighat, "Unsupervised machine learning: Clustering algorithms," *Machine Learning Guide for Oil and Gas Using Python*, pp. 125–168, 2021.

[339] R. Zwanzig, *Nonequilibrium statistical mechanics*. Oxford university press, 2001.

[340] G. Hummer and A. Szabo, "Optimal dimensionality reduction of multistate kinetic and markov-state models," *The Journal of Physical Chemistry B*, vol. 119, no. 29, pp. 9029–9037, 2015.

[341] S. Cao, A. Montoya-Castillo, W. Wang, T. E. Markland, and X. Huang, "On the advantages of exploiting memory in markov state models for biomolecular dynamics," *The Journal of Chemical Physics*, vol. 153, no. 1, 2020.