THE UNIVERSITY OF CHICAGO

UNDERSTANDING CHROMATIN REMODELING THROUGH PHYSICS-BASED MACHINE LEARNING APPROACHES

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
AND
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN BIOPHYSICAL SCIENCES

BY
WALTER ALVARADO

CHICAGO, ILLINOIS
AUGUST 2023

I dedicate this work to my first grade teachers, Donna and Mr. Grantham.

"A jack of all trades is a master of none, but oftentimes better than a master of one."

TABLE OF CONTENTS

LIST OF FIGURES vii

LIST OF TABLES xiii

ACKNOWLEDGMENTS xiv

ABSTRACT xv

1 INTRODUCTION 1

2 TETRANUCLEOSOME INTERACTIONS DRIVE CHROMATIN FOLDING . . 6
- 2.1 Author contributions 6
- 2.2 Abstract 6
- 2.3 Introduction 7
- 2.4 Results and Discussion 10
  - 2.4.1 NRL 157 11
  - 2.4.2 NRL 187 15
  - 2.4.3 NRL 197 18
- 2.5 Conclusion 20
- 2.6 Materials and Methods 21
  - 2.6.1 Simulating the Chromatin Fiber 21
  - 2.6.2 Diffusion Maps 22
- 2.7 Acknowledgements 24
- 2.8 Supporting Figures 25

3 DENOISING AUTOENCODER TRAINED ON SIMULATION-DERIVED STRUCTURES FOR NOISE REDUCTION IN CHROMATIN SCANNING TRANSMISSION ELECTRON MICROSCOPY 27
- 3.1 Author contributions 27
- 3.2 Abstract 27
- 3.3 Introduction 28
- 3.4 Methods 31
  - 3.4.1 Coarse-Grained Molecular Dynamics Simulations and Generation of Synthetic STEM Data 31
  - 3.4.2 Denoising Autoencoder (DAE) Architecture 36
  - 3.4.3 Denoising Performance 39
  - 3.4.4 ChromSTEM Sample Preparation, Imaging, and Reconstruction for A549 Cell Nucleus 41
- 3.5 Results and Discussion 43
  - 3.5.1 Testing on Synthetic Data 44
  - 3.5.2 Application to Experimental Data 47

3.5.3 Identifying Packing Domains and Their Statistical Properties From Denoised ChromSTEM Stack . . . 50
3.6 Conclusions . . . 55
3.7 Acknowledgements . . . 56
3.8 Supporting Figures . . . 57

4 CHARACTERIZING PTM-MODIFIED CHROMATIN FIBERS . . . 60
4.1 Introduction . . . 60
4.2 Methods . . . 62
4.3 Results and Discussion . . . 65
4.3.1 Lower interaction strength induces compaction dynamics in tetranucleosome building blocks. . . . 66
4.3.2 Longer NRLs seems to mediate expansion . . . 68
4.4 Conclusion . . . 73

5 MOLECULAR CHARACTERIZATION OF COVID-19 THERAPEUTICS: LUTEOLIN AS AN ALLOSTERIC MODULATOR OF THE SPIKE PROTEIN OF SARS-COV-2 . . . 74
5.1 Author contributions . . . 74
5.2 Abstract . . . 74
5.3 Design, System, Application . . . 75
5.4 Introduction . . . 76
5.5 Methods . . . 78
5.5.1 Molecular Docking . . . 78
5.5.2 Molecular Dynamics . . . 79
5.5.3 MM/GBSA Calculations . . . 79
5.5.4 Thermodynamic Integration Calculations . . . 80
5.5.5 Contact Maps and RMSF Calculations . . . 80
5.5.6 Strain Analysis . . . 80
5.5.7 Principal Component Analysis . . . 82
5.6 Results and discussion . . . 82
5.6.1 Calculating Binding Free Energies . . . 82
5.6.2 Structural Analysis of Luteolin Binding . . . 86
5.7 Conclusion . . . 91
5.8 Acknowledgments . . . 93

6 CONCLUSION . . . 94
6.1 Summary of contributions . . . 94
6.2 Future directions . . . 95

REFERENCES . . . 96

LIST OF FIGURES

2.1 Recent experimental studies have suggested the existence of two basic secondary structural motifs in chromatin involved in epigenetic regulatory function: the α-tetrahedron (A,C) and β-rhombus (B,D) [89]. α-tetrahedron motifs have been suggested to contribute to local chromatin compaction. β-rhombus conformations promote an open ladder-like chromatin structure that forms linear elongated aggregates. **A,B)** Representative PDB structures from cryo-EM and x-ray crystallography of tri- and tetranucleosomes [122, 27]. **B,C)** Representative structures from 1CPN simulations of tetranucleosome fibers. 9

2.2 Two tetranucleosome motifs, α-tetrahedron and β-rhombus, induce local chromatin compaction or form elongated aggregates at short nucleosome repeat lengths of 157. **A,D)** In 4-nucleosome fibers we observe a global free energy minimum at $(\psi_1 \approx -1.8, \psi_2 \approx -0.7)$ and a metastable minimum at $(\psi_1 \approx 1.3, \psi_2 \approx -1.5)$ that lies $\sim$2.21 kcal/mol higher in free energy. The global minimum contains β-rhombus-like structures, while the local minimum contains compact α-tetrahedron-like packings. **B,E)** The 8-nucleosome fiber possesses two metastable states with a large basin containing the global free energy minimum residing at $(\psi_1 \approx 0.2, \psi_2 \approx -0.8)$ and a weak local minimum at $(\psi_1 \approx -1.5, \psi_2 \approx 2.8)$. The more compact structures reside in the local minimum and contain predominantly α-tetrahedron-like packings, whereas the more extended in the global free energy minimum are β-rhombus-like. **C,F)** The 16-nucleosome fiber possesses a global minimum at $(\psi_1 \approx -0.99, \psi_2 \approx -0.23)$ containing the β-rhombus-like fiber that is $\sim$0.83 kcal/mol more stable than the local minimum at $(\psi_1 \approx 2.44, \psi_2 \approx -0.01)$ containing the α-tetrahedron-like fiber. 13

2.3 Chromatin fibers of 187 NRL show a high propensity for small α-tetrahedron and β-rhombus clusters. Nucleosomes engage in short-lived stacking interactions that form distinct tetranucleosome motifs. **A,D)** In 4-nucleosome fibers, we observe two β-rhombus clusters corresponding to the two distinct ways in which nucleosomes can arrange themselves to form the motif. The leading diffusion CV correlated with α-structure preference ($\rho^{(4)}_{\alpha,\psi_1} = 0.95$) and the second leading CV correlated with β-preference ($\rho^{(4)}_{\beta,\psi_2} = 0.97$). **B,E)** In the 8-nucleosome fibers, α-tetrahedron motifs contribute to local chromatin compaction, while β-rhombus structures resemble the more ladder-like chromatin structure. β-preference is highly correlated with the leading diffusion CV ($\rho^{(8)}_{\beta,\psi_1} = 0.91$). **C,F)** A 16-nucleosome fiber exhibits structural heterogeneity comprised of several α-tetrahedron and β-rhombus motifs. The second CV is moderately correlated with the end-to-end distance of the fiber ($\rho_{R_{end}}, \psi^{(16)}_2 = 0.56$). 17

2.4 Chromatin fibers at NRL 197 are highly irregular and flexible and resemble a "sea of nucleosomes" model. An increase in fiber length is accompanied with an increase in structural irregularity and flexibility. **A,D)** In 4-nucleosome fibers, the increase in nucleosome repeat length allows for more complex nucleosome arrangements of α/β-structures. The leading diffusion map CV is strongly correlated with end-to-end distance of the fiber R_{end} ($\rho^{(4)}_{R_{end},\psi_1} = 0.88$). **B,E)** In 8-nucleosome fibers, the α-tetrahedron motifs contributes to local chromatin compaction while β-rhombus structures resemble the more ladder-like chromatin structure. The second CV is moderately correlated with end-to-end distance ($\rho^{(8)}_{R_{end},\psi_2} = 0.42$). **C,F)** In 16-nucleosome fibers, local chromatin motion is isotropic and largely driven by thermal fluctuations. As for the 8-nucleosome case, the second CV is moderately correlated with end-to-end distance ($\rho^{(16)}_{R_{end},\psi_2} = 0.69$). 19
2.5 Analysis of the 4, 8, and 16-nucleosome systems at varying NRLs using diffusion maps reveal a leading gap in the eigenvalue spectra after the 1st non-trivial mode. 25
2.6 Two representative structures of the 8-nucleosome system at NRL 187 shows a high propensity for small α-tetrahedron (A) and β-rhombus clusters (B). 26
2.7 A fiber of 16-nucleosomes at an NRL of 197 transitioning between an open (left) to closed state (right). The α-tetrahedron and β-rhombus motifs can be considered as the folding units of the chromatin fiber. 26

3.1 A denoising autoencoder (DAE) is constructed and trained on simulations of the chromatin fiber. We simulate nucleosome arrangements using the 1CPN model of chromatin and use the resulting trajectories to generate synthethic STEM images by superimposing crystal structures of the nucleosome (PDB: 1KX5) and DNA snippets. Noise commonly found in angle annular dark field (HAADF) STEM experiments is applied to the images and the DAE trained to remove this noise and preserve the underlying signal. 34
3.2 A denoising autoencoder (DAE) comprises an encoder that compresses the noisy image into a low-dimensional latent space embedding and a decoder that decompresses this embedding into a denoised image. The latent space presents an information bottleneck that the trained DAE model uses to reject noise and preserve signal, enabling reconstruction of denoised images. The DAE is trained on noise-free images for which the ground truth is known and which are artificially corrupted by noise under a noise model representative of the intended application domain for the trained DAE. The image illustrates a DAE that performs an encoding of a 28×28 pixel greyscale (i.e., single channel) image into a 64-channel 8×8 latent space embedding under three convolution plus max pooling layers, followed by decoding under three convolutional plus upsampling layers to generate a denoised 28×28 pixel image [53] 37

3.3 Resolution in dense chromatin regions is obstructed by the intrinsic noise of STEM imaging. a) The α-tetrahedron and β-rhombus tetranucleosome motifs have been proposed to play a regulatory and epigenetic role in the accessibility of DNA to external cellular machinery. The α-tetrahedron promotes DNA compaction whereas the β-rhombus results in elongated chomatic structures. Histone proteins are colored in red and DNA is colored in blue. b) In this work we employ high-resolution ChromSTEM tomograms comprised of 33 slices at 1.23 μm $\times$ 1.23 μm $\times$ 100 nm. The structural resolution accessible to experimental ChromSTEM tomograms is limited by the conformational variability of chromatin within chromatin-rich regions, Poisson noise, and the ability of image segmentation approaches to differentiate background and chromatin signal by voxel intensity. . 45
3.4 Illustrative example of DAE denoising performance to one selected synthetic ChromSTEM test image harvested from the 1CPN MD simulations. a) The selected snapshot was harvested from 1CPN MD simulations of chromatin fibers varying from 150-200 nucleosome repeat length (NRL) and comprised of 4-16 nucleosomes. b) The noise-free synthetic ChromSTEM image I was constructed from the MD snapshot using Eqn. 3.1. This constitutes the ground truth image against which we evaluate denoising performance. c) The noisy image $\tilde{I}$ was generated by adding artificial noise representative of that found in angle annular dark field (HAADF) STEM experiments to the noise-free image using Eqn. 3.2. The denoised image $\hat{I}$ produced from the noisy test image by d) non-local means (NLM), e) block-matching and 3D filtering (BM3D), and f) the DAE. The DAE outperforms NLM and BM3D along all three performance metrics (low MSE, high PSNR, high SSIM) for this particular image and over all 3000 test images (cf. Table 3.1). 47
3.5 Application of the DAE to denoise the experimental tomogram of an imaged A549 cell. The a) original experimental image and b) the image generated after passage through the trained DAE. To improve visual clarity and better highlight features of the images, the pixel intensities are normalized to a [0,1] scale and colored by a pseudo-color gradient indicated by the colorbar as opposed to a single greyscale channel. The denoised image achieves improved resolution of nucleosome-level features within chromatin-rich regions of the experimental image. A subsection comparison between the original c) and denoised experiment e) shows the reduction of noise and results in a smoother 3D reconstruction of the chromatin fiber from the denoised image f) compared to the original d). g) Comparison of the power spectral density (PSD), $P(k)$ between the raw and denoised images shows the denoised image to preserve the large-scale, low-frequency energy density at small wavenumbers k corresponding to the morphological structure of the chromatin fiber, and attenuate the small-scale, high-frequency components at high k that can be primarily attributed to noise. 49

3.6 Denoised ChromSTEM images reveal tetranucleosomes motifs within a dense chromatin cluster. a) Analysis of nucleosome clusters extracted from chromatin-rich regions of the a) raw experimental tomogram and after passing through the DAE. The denoised image clearly shows the presence of α-tetrahedron motifs that are difficult to discern in the raw image. b) Using Chimera, we construct a prototypical tetranucleosome motif (PDB:1KX5) within the extracted volume of our denoised tomogram and find an optimal fit with an average high correlation score of 0.87 [97]. The construction of the 3D interpolation from the 2D imaging slices is computationally expensive but can, in principle, be extended to large sections of chromatin using high performance computing resources. 51
3.7 Denoised ChromSTEM images reveal tetranucleosomes motifs within dense chromatin clusters. Analysis of nucleosome clusters extracted from chromatin-rich regions within a 200 $\times$ 200 nm^2 section of the A) raw experimental tomogram and B) after passing through the DAE. The denoised image clearly shows the presence of C) α-tetrahedron motifs that are difficult to discern in the raw image. We find no evidence for β-rhombus motifs or for the 30-nm fiber. 52
3.8 Structural analysis of chromatin-rich packing domains from the DAE-denoised A549 3D ChromSTEM tomogram. a) A 3D conformation of a packing domain identified from the denoised ChromSTEM tomogram (Figure 4.5b). Statistical distribution of b) domain size R_f, c) packing scaling exponent D, and d) cluster volume concentration CVC, over the 85 chromatin-rich packing domains identified from the denoised ChromSTEM tomogram. Denoising enables identification of $\sim$12% more domains and domains more closely associated in space relative to analysis of the raw 3D ChromSTEM tomograms. 54
3.9 Mass scaling and density analysis originating from the domain centers. A) Mass and radial chromatin density are evaluated starting from the center of a domain (white circle with cyan outline) in concentric circles with increasing distance, r. B) Mass scaling of an individual domain in the log-log scale. We performed linear regression on the mass scaling curve and obtained a slope, $D < 3$ for r up to 68 nm (blue dashed line). Beyond the red asterisk, a more significant divergence (>5% error) in the mass scaling behavior is observed. Further, as r increases, there is a sharp transition to the supra-domain regime with D approaching 3. C) Radial chromatin density of an individual domain in the log-log scale. Radial chromatin density of a domain initially is almost constantly high, roughly near the center of the domain. The density then decreases rapidly at moderate distances from the domain center. After a given large distance shown as a red asterisk at 81 nm, radial density increases again. This increase is potentially due to the end of one domain boundary and the interactions with neighboring domains. 57
3.10 Characterization of morphological properties of original higher-noise tomograph of A549 cells. Statistical distribution of chromatin packing scaling D, cluster volume concentration CVC, and domain size R_f. 58

3.11 Denoising can resolve domains that are closer in space. Left: Domain centers were estimated from denoised tomograms. Right: Five representative regions of the raw and the denoised tomograms show that more domains were identified in the denoised tomogram. Centers are indicated in cyan. 59

4.1 PTM induced modifications are modeled as modification to the well depth, e_0, of the Zewdie Potential. A) Two previous motifs, α-tetrahedron and β-rhombus have been suggested to induce chromatin fiber elongation or compaction. B) Adjusts to the well depth, e_0, are assumed to model possible modifications to the nucleosome interaction strength when two nucleosomes are in a stacked configuration C). . . 64
4.2 Qualitative analysis of tetranucleosomes reveals compaction as intranucleosomal interaction decreases. A) Angle nucleosome angle was calculated to measure folding and compaction. B) Heat capacity of tetranucleosomes models suggest increase structural stability as interaction strength decreases. C) Rouse mode analysis shows interactions with neighboring nucleosomes decreases and interchromatin interaction increases. 66
4.3 Diffusion maps of all conditions reveals two distinct folding motifs accessible by decreasing nucleosome interaction strength. A) A diffusion map colored by identifying motifs with at least a single nucleosome at a distance a deviation away from the mean. Alpha motifs are colored in red and beta motifs are colored in blue. B) Decreasing interaction strength below 0.7 allows access to alpha motif region. C and D) H4 and control nucleosomes can form alpha motifs. E) High interaction strength locks tetranucleosomes in beta state. 69
4.4 An increase in nucleosome repeat length allows chromatin to elongate at low nucleosome interaction strength. The average radius of gyration for all conditions were calculated for fibers of four, eight, and 16 nucleosomes at NRLs of A) 157, B) 187, and C) 197. 70
4.5 Diffusion maps of the 157 NRL system shows formation of systems that resemble a 30 nm fiber and globular system induced by kinks in the fiber from the formation of alpha structure. A) Diffusion map embedding with all conditions considered and colored by radius of gyration display compaction and elongation of the fiber. B) At high interaction strength, compacted structures form an are energetically stable. C) For the control model, beta motifs seem to be the preferred state. D) For acetylated models, both compacted and decompacted conformations are accessible. 71
4.6 Chromatin fiber begins to behave like a liquid polymer. A) First diffusion mode (colored by radius of gyration) of the diffusion map embedding captures compaction and elongation of the fiber. B) At high interaction strength, a single global minimum is observed. As the interaction strength is decreases C,D), the fiber can explore a wider free energy landscape. 72

5.1 The RBD domain (tan) of SARS-CoV-2 recognizes ACE2 (white) as its receptor. We identify and characterize the interactions at three potential small-drug binding sites located at the binding interface between the RBD and the ACE2 receptor, inside the ACE2 protein, and a new previously unidentified distal site to which the drug Luteolin has a high binding affinity (red). Nitrofurantoin is shown in blue and Sapropterin is shown in green. 84
5.2 The interaction diagrams for equilibrated configurations of ligands at the interface (left, Nitrofurantoin), in the RBD domain (middle, Luteolin) and in ACE2 region (right, Sapropterin) are shown. Hydrogen bonds act as the dominating interactions responsible for stabilizing Nitrofurantoin which are formed between the carbonyl oxygens of Nitrofurantoin and the Lys353 residue of ACE2 and RBD residue Gln493. Luteolin is stabilized by a hydrogen bond between its carbonyl oxygen and Tyr369 and pi-alkyl stacking between its aromatic group and Phe377 of the RBD domain. Sapropterin, which is found buried in the ACE2 cavity is stabilized by hydrogen bonds. 87
5.3 Relative protein-protein non-native contact maps in the presence of A) Eriodictyol, B) Nitrofurantoin, C) Sapropterin, and D) Luteolin. The relative non-native contact maps measure the change in contacts relative to the complex with no ligands (red more contacts, blue less contacts). From the graphs we see that the first three panels have identical contact profiles compared to D. 88
5.4 Luteolin induces large allosteric strain when RBD domain is bound to ACE2. A) Root Mean Squared Fluctuations (RMSF) between the RBD/ACE2 complex with top scoring ligand. B) Shear strains mapped. For shear train calculations, only Cα atoms are included. Strain analysis suggest a strong allosteric strain to the ACE2 binding region of the RBD domain when RBD is in complex with ACE2 and Luteolin. 89
5.5 Bound Luteolin induces large strain at RBD/ACE binding region. A) Difference in estimated Root Mean Squared Fluctuations (RMSF) between the RBD domain and Luteolin in complex are shown. (B) Shear strains mapped onto RBD/LUT complex. Regions flanking disulfide bonds have the highest atomic fluctuations that contribute the deformation energy of the ACE2 binding region. C,D) Distal site binding disrupts intramolecular RBD interactions inducing conformational changes at ACE2 binding interface. Visualization of residue-residue cross correlations. Blue lines indicate anti-correlation motions with values between -0.4 and -0.6. Higher correlations between distal sites *sans* ligand (C) and in the presence of Luteolin (D). 90
5.6 Distal binding by Luteolin induces conformational changes at ACE2 binding site. Induced conformational changes at loop binding regions are visualized as captured by the second dominant principal component. Images of important conformational changes are superimposed to emphasize conformations changes introduced after Luteolin binding to distal binding site. 91

LIST OF TABLES

3.1 The mean and standard deviation for 3000 synthetic ChromSTEM test images was calculated to compare the denoising performance of our DAE against non-local means (NLM) and block-matching and 3D filtering (BM3D). Snapshots were harvested from 1CPN MD simulations of chromatin fibers varying from 150-200 nucleosome repeat length (NRL) and comprised of 4-16 nucleosomes and converted into noise-free images I and noisy images $\tilde{I}$ using Eqns. 3.1 and 3.2. Denoising performance is compared using the mean square error (MSE), structural similarity index (SSIM), and peak signal-to-noise ratio (PSNR) metrics. The DAE outperforms non-local means and BM3D along all three performance metrics (low MSE, high PSNR, high SSIM). 46

5.1 Ligand Binding Energies for Simple Glycan Model 83
5.2 Ligand Binding Energies for "Abundant" Glycan Model 84

ACKNOWLEDGMENTS

First and foremost, I would like to express my heartfelt gratitude to my advisors Profs. Juan de Pablo and Andrew Ferguson. Their wisdom, patience, and unending support have been invaluable to my growth and has shaped the course of my academic and personal journey. They've challenged me to exceed my own expectations and inspired me in countless ways. I am truly grateful for their impactful presence in my life.

To my parents, thank you for instilling in me the value of hard work and for the sacrifices you've made to ensure a better life for us. I also owe a debt of gratitude to my siblings, Stephanie and Rene, who taught me that laughter is an excellent way of coping with stress and adversity.

And of course, I want to thank Michelle, my partner, my best friend, and my confidante, whose love and support has given me an immeasurable amount of courage. Thank you for embarking on this journey with me.

ABSTRACT

The packing of nucleosomes regulates gene expression through genome condensation and expansion, but the specific structures and their thermodynamic stabilities remain unresolved. In this work, we employ the use of a meso-scale model of chromatin, referred to as "1-Cylinder-per-Nucleosome," or 1CPN, in combination with nonlinear manifold learning to identify and characterize the structure and free energy of metastable states of short chromatin segments. Our results reveal the intrinsic formation of two previously characterized tetranucleosomal conformations, the "α-tetrahedron" and the "β-rhombus," which have been suggested to play a role in inducing chromatin compaction or elongation, respectively. Building upon these findings, we leverage convolutional neural networks and molecular dynamics simulations, to design a deep convolutional denoising autoencoder (DAE) capable of providing nucleosome-level resolution of scanning transmission electron microscopy images of chromatin (ChromSTEM). Our DAE is trained on synthetic images generated from simulations of the chromatin fiber using the 1CPN model of chromatin, thereby learning structural features driven by the physics of chromatin folding. We find that our DAE outperforms other well-known denoising algorithms without degradation of structural features and allows for the resolution of individual nucleosomes and organized domains within chromatin-dense regions. Lastly, we investigate how post-translational modifications (PTMs), modeled as modifications to the nucleosome interaction potential, affect the construction of these motifs and, consequently, the chromatin fiber as a whole. Our study provides important insight into chromatin folding and highlights the value of interdisciplinary approaches in this field.

CHAPTER 1

INTRODUCTION

Chromatin, the complex of DNA, RNA, and proteins found in the nuclei of eukaryotic cells, has a dynamic, multi-scale structure that regulates transcription, replication, and DNA repair [126, 39]. Its basic building block, the nucleosome, is a disk-like DNA-protein complex comprising approximately 146 basepairs (bp) of DNA wrapped around a protein complex known as the histone octomer [39]. These small positively charged proteins (H2A, H2B, H3, and H4) bind tightly to the negatively charged DNA. [54]. Adjacent nucleosomes are connected by a segment of linker DNA. When the length of this linker DNA is combined with the length of the core DNA, it is referred to as the nucleosome repeat length (NRL), which has a distribution centered around 180 base pairs, depending on the organism, cell type, or specific loci within a cell type [126].

The long-proposed idea that nucleosomes condense into a 30-nm thick chromatin fiber, progressively folding to form mitotic chromosomes and interphase nuclei, continues to generate debate, as does the character and properties of higher-order chromatin structures *in vivo* [125]. Experimental techniques, such as cryo-electron microscopy, X-ray scattering, scanning transmission electron microscopy, chromosome conformation capture (Hi-C), and super-resolution STORM (stochastic optical reconstruction microscopy) have indicated that chromatin in living cells is predominantly comprised of highly irregular structures [28, 49, 62, 87, 105, 108]. This aligns with the emerging view that chromatin organization is dominated by intermediate scale assemblies ($\sim$3-10 nucleosomes), as revealed by chromosome conformation capture techniques and electron cryotomography [16, 49, 83, 89].

Building on the importance of these smaller nucleosome clusters, tetranucleosomes have been suggested to play a crucial role in gene expression regulation as structural and functional units [49, 89, 118]. They have been successfully crystallized at nucleosome repeat lengths (NRLs) of 157 and 167 bp and detected in lengthier chromatin fibers of 12 tandem repeats

of 177 and 187 bp through cryo-EM [27, 109]. Additionally, cryo-EM has provided evidence of nucleosome stacking in human cell lines, revealing two-start tetranucleosomes stacked one atop the other [29]. Recently, two tetranucleosome motifs, the α-tetrahedron and β-rhombus, have been proposed to serve regulatory functions [89]. Through the integration of nucleosome-resolved Hi-C measurements and simulated annealing-molecular dynamics (SA-MD), nucleosome pair orientations were examined, focusing on which nucleosome ends were connected in the yeast genome [89]. Located near regulatory regions, these specific folding motifs might perform a regulatory function and align with existing knowledge of chromatin organization, such as nucleosomal "clutches" and "topologically associated domains" (TADs) [89]. Nevertheless, it remaines uncertain whether these motifs emerge due to a regulated set of interactions steered by protein partners or if they're intrinsically stable folding states.

Though imaging techniques have advanced considerably, it is remains difficult to capture these smaller folding motifs *in vivo*. Chemical fixation with glutaraldehyde and methanol, which are commonly used in electron microscopy and super-resolution light microscopy, has advanced chromatin research. However, these techniques have the unintended consequence of generating artificial structures that are not normally present under biological conditions [69]. Hydrophobic interactions increase under these conditions, causing free molecules to adhere and form long, continuous artificial structures [69]. In addition, the long duration required for these imaging techniques may obscure essential short-term nucleosome dynamics and misrepresent the true nature of small nucleosome domains [69].

Here, molecular dynamics (MD) can provide a temporal and spatial resolution that is frequently challenging to achieve experimentally. It permits the controlled and efficient study of short-term dynamics and smaller scale structures. Moreover, MD simulations can help distinguish between structures that emerge as a result of biological interactions and structures that may emerge as a result of experimental conditions. By combining these simulations with experimental observations, a more thorough and precise understanding of

the structure, stability, and function of tetranucleosome motifs could be attained. This is crucial for comprehending chromatin packaging, DNA accessibility, and the field of epigenetics as a whole.

In this dissertation, we delve deeper into the investigation of complex chromatin structures, with a particular emphasis on the intrinsic arrangements of nucleosomes. Through a combination of simulation techniques and exhaustive data analysis methods, our aim is to decipher the relationships between nucleosomal organization and chromatin structure in order to shed new light on their roles in gene regulation. Here we present our methodology, key findings, and future directions in the pursuit of elucidating the unique characteristics of chromatin organization and its implications.

In Chapter 2, we investigate the complex multiscale organization of chromatin, a structure crucial for DNA processes such as transcription, replication, and repair. Using the 1CPN mesoscale model of chromatin and diffusion maps, a non-linear manifold learning technique, we analyze the structure and energy dynamics of short chromatin segments consisting of four to 16 nucleosomes. Our study confirmed the formation of two known tetranucleosomal structures: the compact 'α-tetrahedron' and the elongated 'β-rhombus'. These structures have been suggested to play a significant roles in DNA accessibility and gene expression regulation. We noted that as the nucleosome repeat length increased, the structure exhibited greater irregularity and flexibility, leading to a dynamic, liquid-like behavior facilitating structural reorganization. These findings underscore the inherent stability of tetranucleosome motifs, suggesting that local internucleosomal interactions significantly influence chromatin packing, dynamics, and accessibility through emergent local mesoscale structures.

In Chapter 3 we illustrate a general model for linking experimental imaging and theoretical modeling via machine-learning techniques to facilitate high-resolution exploration of the structural organization of biological systems. Using a combination of molecular dynamics and machine-learning techniques, we devised and trained a denoising autoencoder (DAE)

to effectively remove noise typically observed in scanning transmission electron microscopy tomography with ChromEM staining (ChromSTEM) imaging. Our model was trained using physics-based coarse-grained molecular dynamics simulations employing the 1CPN model and made it capable of differentiating between the signal from chromatin structures and artificial noise. When tested on synthetic ChromSTEM images generated from molecular simulations, where the ground truth is precisely known, our model outperformed conventional denoising techniques, achieving a 57% improvement in the mean squared error over block-matching and 3D filtering and a 72% improvement over nonlocal means. On experimental tomographs, the denoised images allowed for the detection of approximately 12% more chromatin-rich packing domains obscured by noise within the raw images.

In Chapter 4, our results reveal that alterations in the intensity of local nucleosome-nucleosome interactions can significantly impact higher-order chromatin folding. Crucially, even as interaction strength diminished, local compaction persisted, thus resolving previously reported counterintuitive findings. Our study also illuminated the balance between different folding motifs and the role of DNA linker length in these dynamics. Altogether, our results offer insights into the mechanisms underpinning the complex interplay of histone modifications on chromatin structure.

With the emergence of the global COVID-19 pandemic, an unexpected need for the application of these molecular modeling techniques surfaced. The pandemic dramatically highlighted the importance of understanding protein structure and dynamics in disease pathogenesis and treatment. In Chapter 5 we present our study on the spike protein of the novel coronavirus, SARS-CoV-2, which was the primary target of immunological interventions. Our research focus shifted to investigate the molecular dynamics of the SARS-CoV-2 spike protein to provide insights into this pressing issue. Notably, the fundamental understanding of protein folding, structure, and dynamics - principles at the heart of biophysics - is the link between my initial research on chromatin fiber and my later endeavor on the viral

spike protein. Incorporating this latter project into my dissertation, therefore, not only reflects the unpredictable nature of science, but also highlights the adaptability and extensive applicability of these computational methods.

CHAPTER 2

TETRANUCLEOSOME INTERACTIONS DRIVE CHROMATIN FOLDING

Reprinted with permission from **Alvarado, W.**; Moller, J.; Ferguson, A. L.; de Pablo, J. J. Tetranucleosome Interactions Drive Chromatin Folding. ACS Cent. Sci. 2021, 7, 6, 1019–1027. DOI: 10.1021/acscentsci.1c00085.

2.1 Author contributions

W.A. and J.M. conceptualized the study and interpreted findings. W.A. conducted simulations and developed analysis software. W.A., J.D.P., and, A.F. wrote the paper. A.F. and J.D.P. supervised the project. All authors participated in reviewing and commenting on the study drafts.

2.2 Abstract

The multi-scale organizational structure of chromatin in eukaryotic cells is instrumental to DNA transcription, replication, and repair. At mesoscopic length scales, nucleosomes pack in a manner that serves to regulate gene expression through condensation and expansion of the genome. The particular structures that arise and their respective thermodynamic stabilities, however, have yet to be fully resolved. In this study, we combine molecular modeling using the 1CPN meso-scale model of chromatin with nonlinear manifold learning to identify and characterize the structure and free energy of metastable states of short chromatin segments comprising between four to 16 nucleosomes. Our results reveal the intrinsic formation of two previously characterized tetranucleosomal conformations, the "α-tetrahedron" and the "β-rhombus", which have been suggested to play an important role in the accessibility of DNA and, respectively, induce local chromatin compaction or elongation.

The spontaneous formation of these motifs is potentially responsible for the slow nucleosome dynamics observed in experimental studies. Increases of the nucleosome repeat length (NRL) are accompanied by more pronounced structural irregularity and flexibility and, ultimately, a dynamic liquid-like behavior that allows for frequent structural reorganization. Our findings indicate that tetranucleosome motifs are intrinsically stable structural states, driven by local internucleosomal interactions, and support a mechanistic picture of chromatin packing, dynamics, and accessibility that is strongly influenced by emergent local mesoscale structure.

2.3 Introduction

Chromatin is the complex of DNA, RNA, and proteins found in eukaryotic cell nuclei. Chromatin's dynamic, multi-scale structure is central to the regulation of transcription, replication, and DNA repair [126, 39]. The basic building block of eukaryotic chromatin is the nucleosome, a disk-like DNA-protein complex of approximately 146 basepairs (bp) of DNA wrapped around a protein complex known as the histone octomer [39]. These small and positively charged proteins bind strongly to the negatively charged DNA. Each nucleosome contains four core histone proteins (H2A, H2B, H3, and H4) which are found in equal proportions in cells [54]. Nucleosomes connect to adjacent nucleosomes through a segment of linker DNA whose combined length with core DNA is referred to as a nucleosome repeat length (NRL). Nucleosome repeat lengths exhibit a distribution centered around 180 bp, depending on organism, cell type, or loci in a given cell type [126].

It has long been proposed and debated that nucleosomes condense into a 30-nm thick chromatin fiber that progressively folds to form mitotic chromosomes and interphase nuclei [125]. The character and properties of higher-order chromatin structures *in vivo* also continue to be the source of debate [125]. Experimental techniques such as cryo-electron microscopy, X-ray scattering, scanning transmission electron microscopy, chromosome conformation capture (Hi-C), and super-resolution STORM (stochastic optical reconstruction microscopy) indicate

that chromatin in living cells is predominantly comprised of highly irregular structures [28, 49, 62, 87, 105, 108]. The emerging view that chromatin organization is dominated by intermediate scale assemblies (∼3-10 nucleosomes) is supported by results from various chromosome conformation capture techniques and electron cryotomography, which have revealed the existence of clusters comprising only a few nucleosomes that may play a role in chromatin biology [16, 49, 83, 89].

Tetranucleosomes are proposed to be functional and structural units that regulate gene expression [49, 89, 118]. They have been crystallizable at NRLs of 157 and 167 bp and observed in longer chromatin fibers of 12 tandem repeats of 177 and 187 bp from cryo-EM [27, 109]. Cryo-EM experimental evidence of nucleosome stacking, with two-start tetranucleosomes stacked on top of each other, has also been reported in human cell lines [29]. Two tetranucleosome motifs have recently been proposed to serve regulatory functions: the α-tetrahedron and β-rhombus [89]. The all-atom structure of the α-tetrahedron and β-rhombus motifs are shown in Figure 2.1A and 2.1B respectively [122, 27]. In Figure 2.1C-D, we provide a representation of these two motifs that relies on the 1CPN model. By combining nucleosome-resolved Hi-C measurements with simulated annealing-molecular dynamics (SA-MD), the orientation of nucleosome pairs were analyzed by focusing on which nucleosome ends were ligated to one another in yeast genome [89]. Given their location near regulatory regions, within genes (α-tetrahedron) and at gene ends (β-rhombus), these specific folding motifs may serve a regulatory function and seem consistent with what is already established about chromatin organization, such as nucleosomal "clutches" and "topologically associated domains" (TADs) [89]. It is unclear, however, whether these motifs arise due to an orchestrated set of interactions governed by protein partners, or if they are intrinsically stable structural states of the nucleosome. Chemical fixation with compounds such as glutaraldehyde and methanol, which are employed in EM and super-resolution light microscopy, have benefited the study of chromatin but introduce the formation of artificial

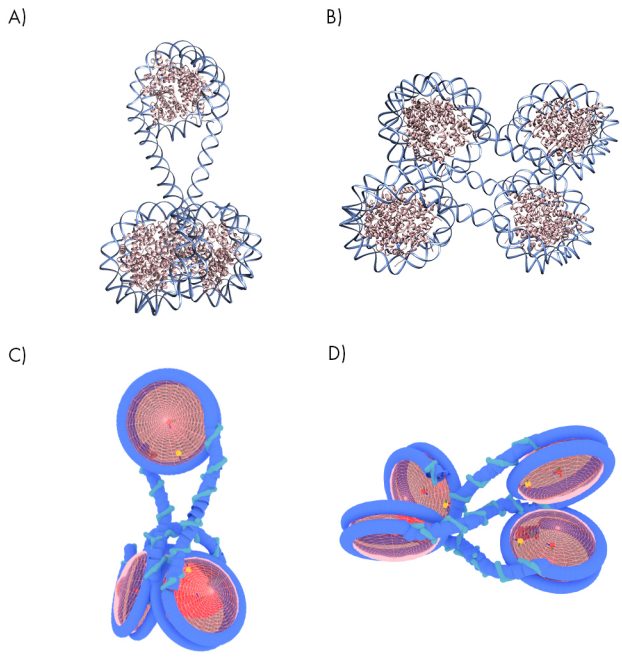

Figure 2.1: Recent experimental studies have suggested the existence of two basic secondary structural motifs in chromatin involved in epigenetic regulatory function: the α-tetrahedron (A,C) and β-rhombus (B,D) [89]. α-tetrahedron motifs have been suggested to contribute to local chromatin compaction. β-rhombus conformations promote an open ladder-like chromatin structure that forms linear elongated aggregates. **A,B)** Representative PDB structures from cryo-EM and x-ray crystallography of tri- and tetranucleosomes [122, 27]. **B,C)** Representative structures from 1CPN simulations of tetranucleosome fibers.

structures not present under biological conditions [69]. The adhearance of free molecules, facilitated by the increase in hydrophobic interactions, form long and continuous artificial structures which has become an ever more increasing issue in newer imaging experiments such as STORM [69]. Furthermore, long imaging times may mask important short-time nucleosome dynamics and mischaracterize small nucleosome domains [69]. Resolving the emergent structure and stability of tetranucleosome motifs as a function of NRL is of critical biological interest in understanding chromatin packing, DNA accessibility, and epigenetics.

In this work we relied on large-scale simulations of various chromatin fiber lengths using the recently developed 1CPN meso-scale model of chromatin (1-Cylinder-per-Nucleosome) [61]. This model is structured around a rigorous multiscale approach where free energies are

derived from an established and extensively validated model of the nucleosome and mapped onto a single anisotropic topology [35, 59]. The 1CPN model has accurately reproduced a range of chromatin properties, including nucleosome-nucleosome free energy interactions, nucleosome unwrapping free energies, and sedimentation coefficients of short chromatin fibers [61], and its computational efficiency enables access to length and time scales not currently possible with atomistic models.

To investigate the organizational properties of chromatin, we used nonlinear manifold learning to identify the metastable states and collective motions that mediate transitions between such states (see *Diffusion Maps* in Materials and Methods). We identified the formation of two previously characterized tetranucleosomal conformations (α-tetrahedron and β-rhombus motifs) that play an important role in the accessibility of DNA. Our results show that the relative thermodynamic stability of these two motifs are intrinsic properties of the chromatin chain independent of external cellular factors. Our findings and trends with NRL are consistent with experimental observations and provide new mechanistic insight into mesoscale structure formation and the hierarchical structure and conformational transitions of chromatin.

2.4 Results and Discussion

We present the results of 1CPN simulations of three representative chromatin fibers with NRL values of 157, 187, and 197. These repeat lengths were selected due to their structures being available from X-ray and cyro-EM, thereby allowing direct comparisons to experiment [27, 118, 107]. For each NRL system we considered chromatic fibers containing four, eight, and 16 nucleosomes, and identified the metastable states and oligonucleosome organization using diffusion map dimensionality reduction. Simulations were conducted from an initial elongated conformation; convergence was considered achieved when the root mean square deviation (RMSD) with respect to the initial frame approached a steady state (see *Simulating*

the Chromatin Fiber in Material and Methods). Where possible, we make contact with experimental measurements.

2.4.1 NRL 157

Diffusion maps are a non-linear dimensionality reduction technique that extracts collective variables (CVs) characterizing the large-scale collective motions of molecular systems [32]. These CVs are computed as the eigenvectors of a particular transition matrix characterizing the similarity of various system configurations. The appropriate number of eigenvectors to retain in the low-dimensional diffusion map embeddings can be determined by identifying a gap in the eigenvalue spectrum of the transition matrix that demarcate the leading modes characterizing the important large scale collective motions from those associated with smaller amplitude conformational changes. Often the spectrum may exhibit multiple gaps corresponding to a hierarchical splitting into groups of collective modes ordered by decreasing amplitude. In all systems studied in this work we observe the leading spectral gap to occur after the first non-trivial eigenvalue (Figure 2.5). This suggests that the conformational dynamics of the system are dominated by a single CV, and we invariably correlate the leading eigenvector with global contraction or expansion of the fiber. In many cases we also find the second eigenvector to also be informative and we can often correlate it with some additional aspect of the fiber conformational motions as discussed below. Many of the eigenvalue spectra exhibit additional spectral gaps indicative of a hierarchical partitioning of CVs, but our exploration of the higher order eigenvectors did not show any obvious correlation with putative physical order parameters or expose any additional interpretability beyond what we were able to extract from the 2D embeddings. For this reason, in addition to visual clarity and accessibility, all embeddings in this work are 2D in nature.

In Figure 2.2A-C we present 2D embeddings of the 95,000 simulation snapshots of the chromatin fibers with NRL 157 into the two leading diffusion map CVs (ψ_1,ψ_2). To provide

some physical interpretability of the diffusion map embeddings, we computed the RMSD of each contiguous stretch of four nucleosomes relative to a prototypical α/β-structure motif by passing a 4-nucleosome sliding window along the fiber. Dividing by the number of windows defines an order parameter by which to measure the propensity of the complete fiber for either motif. In the four and eight nucleosome fibers, a Pearson correlation analysis reveals the α/β-structure preference to be strongly correlated with the leading diffusion map CV ($\rho^{(4)}_{\alpha,\psi_1} = -0.75, \rho^{(4)}_{\beta,\psi_1} = 0.87; \rho^{(8)}_{\alpha,\psi_1} = 0.68, \rho^{(8)}_{\beta,\psi_1} = -0.72$). For the 16-nucleosome fiber, the leading CV is strongly correlated with the radius of gyration R_g ($\rho^{(16)}_{R_g,\psi_1} = -0.74$). The strong correlation with radius of gyration (R_g) indicates that the leading CV extracted by diffusion maps is strongly associated with the compaction and expansion of the 16-nucleosome fiber. Inspection of the molecular configurations of the elongated and compact fibers in Figure 2.2C, reveals that the elongated fiber contains a high degree of β motif character whereas the compact fiber is composed of primarily α motifs. For the systems studied in this work, it is a general trend that the leading CVs identified by diffusion maps tend to correspond to local α/β motif character for shorter NRL values and fewer nucleosomes versus more global descriptors such as the fiber radius of gyration or end-to-end distance for longer NRL values and more nucleosomes. This transition can be understood as a result of the increased length and flexibility of the fiber leading to the emergence of more global collective variables that subsume and contain the local packing effects.

To resolve the metastable states of the system, in Figure 2.2D-F we present the corresponding free energy surfaces (FES) in the Gibbs free energy $G(\Psi)$ constructed over the 2D intrinsic manifolds and decorated with representative molecular renderings. In the 4-nucleosome system (Figure 2.2D), a global free energy minimum is observed at ($\psi_1 \approx -1.8, \psi_2 \approx -0.7$), and a metastable minimum is found at ($\psi_1 \approx 1.3, \psi_2 \approx -1.5$) lying $\sim$2.21 kcal/mol higher in free energy. Visualization of the simulation snapshots contained within the local minima of the free energy landscape confirms the results of the Pearson correlation analysis that

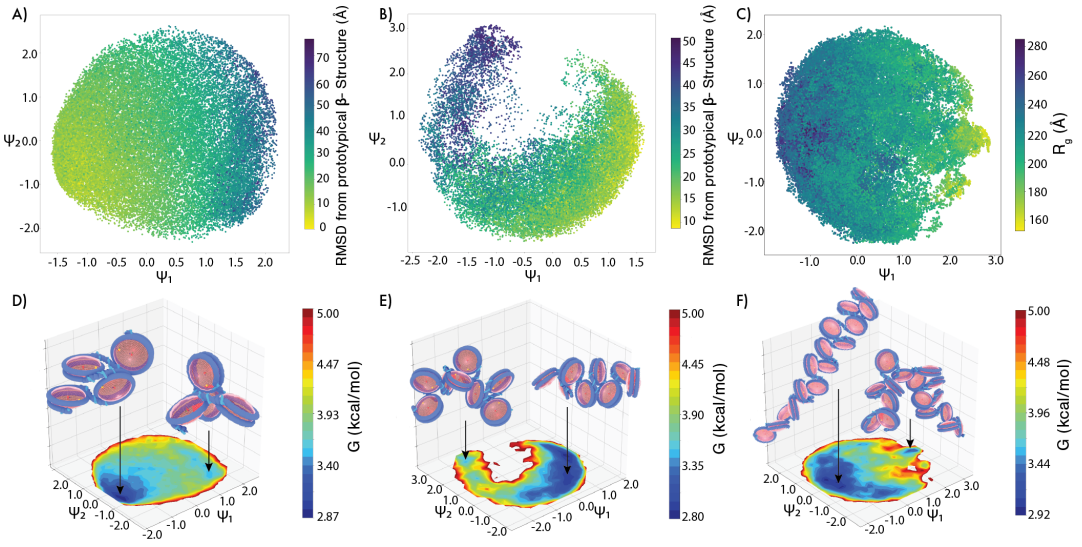

Figure 2.2: Two tetranucleosome motifs, α-tetrahedron and β-rhombus, induce local chromatin compaction or form elongated aggregates at short nucleosome repeat lengths of 157. **A,D)** In 4-nucleosome fibers we observe a global free energy minimum at $(\psi_1 \approx -1.8, \psi_2 \approx -0.7)$ and a metastable minimum at $(\psi_1 \approx 1.3, \psi_2 \approx -1.5)$ that lies $\sim$2.21 kcal/mol higher in free energy. The global minimum contains β-rhombus-like structures, while the local minimum contains compact α-tetrahedron-like packings. **B,E)** The 8-nucleosome fiber possesses two metastable states with a large basin containing the global free energy minimum residing at $(\psi_1 \approx 0.2, \psi_2 \approx -0.8)$ and a weak local minimum at $(\psi_1 \approx -1.5, \psi_2 \approx 2.8)$. The more compact structures reside in the local minimum and contain predominantly α-tetrahedron-like packings, whereas the more extended in the global free energy minimum are β-rhombus-like. **C,F)** The 16-nucleosome fiber possesses a global minimum at $(\psi_1 \approx -0.99, \psi_2 \approx -0.23)$ containing the β-rhombus-like fiber that is $\sim$0.83 kcal/mol more stable than the local minimum at $(\psi_1 \approx 2.44, \psi_2 \approx -0.01)$ containing the α-tetrahedron-like fiber.

ψ_1 should separate configurations based on α/β character, with the global minimum at low-ψ_1 containing β-rhombus-like structures of the four nucleosomes, and the local minimum at high-ψ_1 containing primarily of compact α-tetrahedron-like packings. These two motifs were previously reported in several chromosome conformation capture experiments, which have provided evidence for the existence of tri- and tetranucleosome folding motifs as the primary level of organization in yeast genome [49, 89]. As mentioned earlier, oligonucleosome motifs may serve an important role as functional elements for processes like transcription, replication, and DNA repair [83]. In addition, the rearrangement of tetranucleosomes from open/closed states may serve as a potential target for cellular regulation. For example, previous studies have shown many chromatin regulators have a preference to bind to multinucleosome structures in vitro [74]. We emphasize that our simulations were conducted from an initial elongated conformation, and we have verified that our sampling of the conformational ensemble is converged with the fiber making multiple transitions over the intrinsic manifold and between the two basins. As such, these structures represent inherent metastable states of the 4-nucleosome NRL 157 system.

In the larger 8-nucleosome system we again observe two metastable states with a large basin for the global free energy minimum at ($\psi_1 \approx 0.2, \psi_2 \approx -0.8$), and a weak local minimum at ($\psi_1 \approx -1.5, \psi_2 \approx 2.8$) lying $\sim$2.51 kcal/mol higher in free energy (Figure 2.2E). As for the 4-nucleosome fiber, the global minimum consists of more extended β-rhombus-like arrangements and the local minimum and contain more compact α-tetrahedron-like packings. This analysis reveals that the prototypical α-tetrahedron and β-rhombus tetranucleosome motifs are preserved within the longer 8-nucleosome fiber.

The 16-nucleosome fiber exhibits a similar trend, with a global minimum at ($\psi_1 \approx -0.99, \psi_2 \approx -0.23$) consisting of a β-rhombus-like fiber that is $\sim$0.83 kcal/mol more stable than the local minimum at ($\psi_1 \approx 2.44, \psi_2 \approx -0.01$) that includes the α-tetrahedron-like fiber (Figure 2.2F). As was the case for the 4 and 8-nucleosome fibers, the 16-nucleosome

β-rhombus-like fiber is less compact, more elongated, and more stable than the α-tetrahedron-like fiber. This is due to the condensation and ordered linear packing of the β-rhombus subunits compared to the relatively poorly packed and disordered chain of α-tetrahedron motifs.

Our results confirm the existence of two tetranucleosome motifs, α-tetrahedron and β-rhombus, that induce local chromatin compaction or form elongated aggregates at short nucleosome repeat lengths of 157, respectively. These computational findings are consistent with experimental Hi-C studies that have shown α-tetrahedron motifs to contribute to local chromatin compaction [89]. Tetranucleosomes that take on an α-tetrahedron conformation introduce kinks and bring disorder into the fiber. β-rhombus conformations, on the other hand, form an open ladder-like chromatin structure that packs to form linear elongated aggregates. Our simulations of longer chromatin fibers at an NRL of 157 in the absence of H1, a linker histone protein considered important in maintaining higher-order chromatin structure, adopted a zig-zag conformation in agreement with experiments from reconstituted arrays of nucleosomes without H1 [25, 109]. The crystallization of 157 NRL fibers may be due to the stabilizing interactions between neighboring nucleosomes which intrinsically form these tetranucleosome motifs [109].

2.4.2 NRL 187

In Figure 2.3A-C we present the 2D embeddings into the two leading diffusion map CVs for chromatin fiber with NRL 187. For a 4-nucleosome fiber, the leading diffusion CV is strongly correlated with α-structural preference ($\rho^{(4)}_{\alpha,\psi_1} = 0.95$) and the second leading CV correlated with β-preference ($\rho^{(4)}_{\beta,\psi_2} = 0.97$). For an 8-nucleosome fiber, β-preference highly correlated with the leading diffusion CV ($\rho^{(8)}_{\beta,\psi_1} = 0.91$). For the fiber comprised of 16 nucleosomes, the second CV is moderately correlated with the end-to-end distance of the fiber R_{end} ($\rho_{R_{end}}, \psi^{(16)}_2 = 0.56$). In Figure 2.3D-F we present the corresponding FES and

their representative molecular renderings derived from simulation.

In contrast to the 4-nucleosome system with 157 NRL (Figure 2.2D), the 187 NRL system exhibits two distinct β-rhombus wells in the free energy landscape for the 187 NRL (Figure 2.3D). This is a consequence of the increased degree of freedom permitted by the longer NRL chain that opens up the availability of additional metastable structures. Specifically, the two wells correspond to two distinct alternative packings of the four nucleosomes into the β-rhombus motif. The global minimum at ($\psi_1 \approx 1.0, \psi_2 \approx 0.5$) corresponds to a β-rhombus where the N and (N+1) nucleosomes lie diagonal from each other. The local minimum located at ($\psi_1 \approx -0.82, \psi_2 \approx -1.62$) contains a second β-rhombus conformation where the N and (N+1) nucleosomes lie directly across from one other and exists $\sim$3.55 kcal/mol higher in free energy. The local minimum at ($\psi_1 \approx 1.3, \psi_2 \approx -1.5$) contains the compact α-tetrahedron motif that lies $\sim$1.75 kcal/mol higher in free energy. The gap between clusters suggests limited sampling in the interstitial region between the metastable α/β states. Comparing our results to simulations of NRLs of 157 (Figure 2.2D-F), the transition region is destabilized relative to the metastable states for longer NRL fibers reflecting their larger degree of freedom to stably pack the nucleosomes into the α/β motifs.

The 8 and 16-nucleosome fibers exhibit a relatively broad free energy landscape comprising metastable α/β motifs (Figures 2.3E,F, 2.6, and 2.7). As in the case for chromatin fibers of NRL 157, α-tetrahedron motifs contribute to local chromatin compaction while β-rhombus structures resemble ladder-like chromatin structure. The presence of several β-rhombus structures is consistent with the idea of a tri- or tetranucleosome motif in chromatin fiber folding proposed from Micro-C experiments where the decay of nucleosome-nucleosome interactions as a function of distance revealed no evidence for long-range periodicity in inter-nucleosomal interactions but of short-range structures [49]. In addition, the results of our unbiased simulations where the only driver of the dynamics is thermal fluctuations, suggest that local chromatin motion is spontaneous, as observed from single-particle tracking

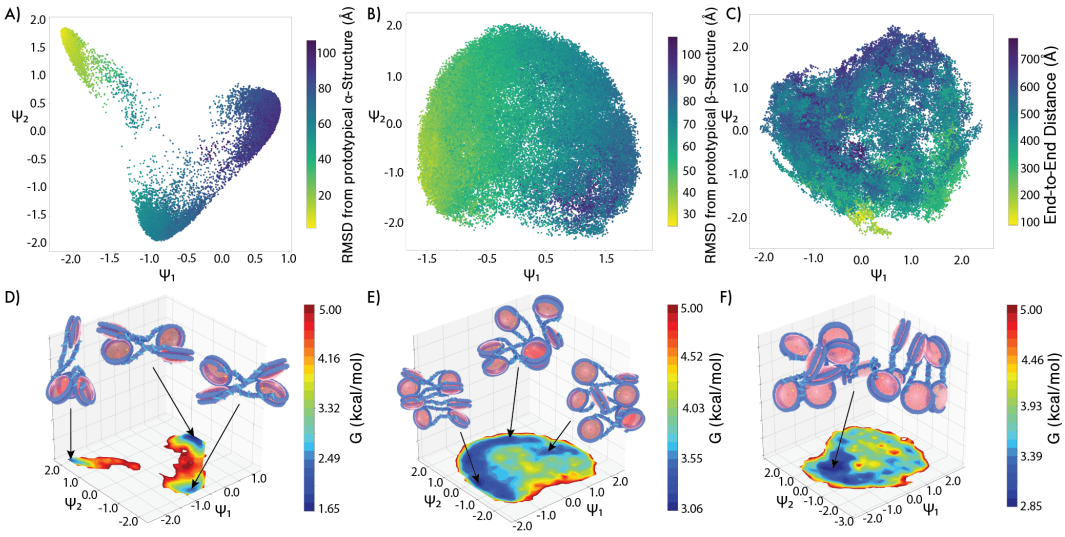

Figure 2.3: Chromatin fibers of 187 NRL show a high propensity for small α-tetrahedron and β-rhombus clusters. Nucleosomes engage in short-lived stacking interactions that form distinct tetranucleosome motifs. **A,D)** In 4-nucleosome fibers, we observe two β-rhombus clusters corresponding to the two distinct ways in which nucleosomes can arrange themselves to form the motif. The leading diffusion CV correlated with α-structure preference ($\rho^{(4)}_{\alpha,\psi_1} = 0.95$) and the second leading CV correlated with β-preference ($\rho^{(4)}_{\beta,\psi_2} = 0.97$). **B,E)** In the 8-nucleosome fibers, α-tetrahedron motifs contribute to local chromatin compaction, while β-rhombus structures resemble the more ladder-like chromatin structure. β-preference is highly correlated with the leading diffusion CV ($\rho^{(8)}_{\beta,\psi_1} = 0.91$). **C,F)** A 16-nucleosome fiber exhibits structural heterogeneity comprised of several α-tetrahedron and β-rhombus motifs. The second CV is moderately correlated with the end-to-end distance of the fiber ($\rho_{R_{end}}, \psi^{(16)}_2 = 0.56$).

experiments [73]. The observed local motion of chromatin at longer NRLs are characteristic of fluid-like behavior, suggesting a more irregular and dynamic view of chromatin. Interestingly, variability in motion seems to be a product of the constraints imposed by physical or geometrical factors. Experimental single-nucleosome tracking data in living human cells have suggested a nucleosome's motion dependence on nucleosome-nucleosome interactions [5]. The fiber makes multiple transitions over the intrinsic manifold and between the two metastable states. As such, these nucleosomes engage in short-lived stacking interactions, relative to total simulation time, that form distinct tetranucleosome motifs which we propose may be responsible for slowing chromatin dynamics through transient trapping in metastable states.

2.4.3 NRL 197

We present in Figure 2.4A-C the 2D embeddings into the two leading diffusion map CVs for chromatin fibers of NRL 197. For the 4-nucleosome fiber, the end-to-end distance R_{end} is strongly correlated with the leading diffusion CV ($\rho^{(4)}_{R_{end},\psi_1} = 0.88$). In the case of 8 and 16 nucleosomes, each leading CVs lacked simple physical interpretation but the second CV correlated moderately with end-to-end distance ($\rho^{(8)}_{R_{end},\psi_2} = 0.42, \rho^{(16)}_{R_{end},\psi_2} = 0.69$). We observed that for longer NRL, fibers are highly irregular and flexible, resembling a "sea of nucleosomes" model [28, 87]. In this view, nucleosomes are able to interact with local and distant partners, leading to a more dynamic, liquid-like behavior that supports structural plasticity and frequent reorganization.

The FES for a 4-nucleosome 197 NRL fiber in Figure 2.4D presents the various ways in which either motif can assemble. As in the case of a 187 NRL fiber, the increase in nucleosome repeat length allows for more complex nucleosome arrangements of α/β-structures. As previous studies have suggested, shorter linker lengths form less flexible fibers, which have the advantage of exposing their DNA to transcription and replication machinery by the displacement of only a few nucleosomes [95]. The lack of a cohesive structure for longer linker lengths suggests that they are less stable and more prone to opening as observed experimentally [99]. At 8 and 16 nucleosomes (Figure 2.4E,F), the liquid-like behavior of chromatin becomes more apparent. Representative structures for both fibers exhibit structural heterogeneity comprised of several α-tetrahedron and β-rhombus motifs. Our results are in line with ChromEMT experiments which have confirmed the presence of small clusters of ∼2-10 nucleosomes, and a lack of larger organized structures [93]. Our results suggest that these motifs arise due to an orchestrated set of interactions that are mainly driven by nucleosome-nucleosome interactions.

The liquid-like behavior of chromatin is critical to gene expression since these dynamic changes directly affect DNA accessibility. Hi-C experiments have shown a correlation between

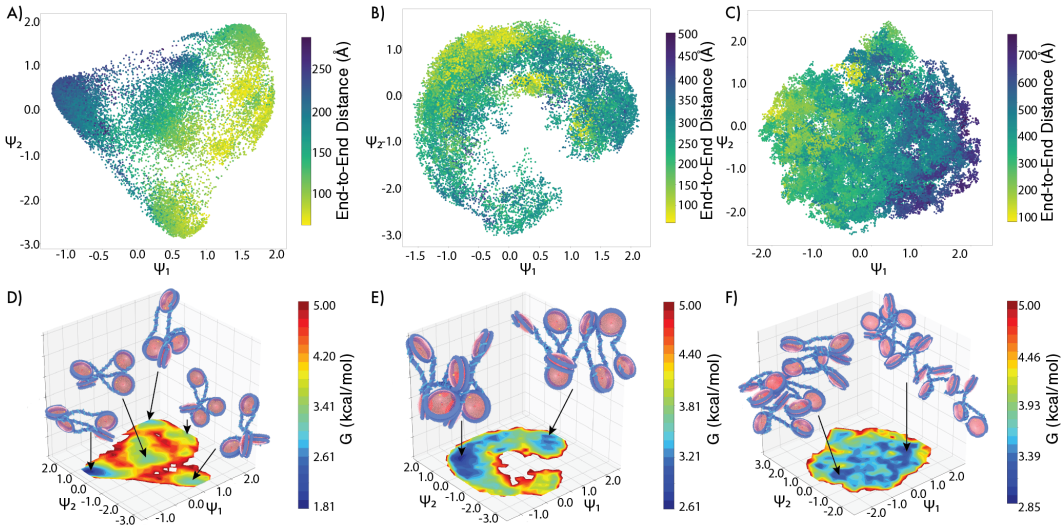

Figure 2.4: Chromatin fibers at NRL 197 are highly irregular and flexible and resemble a "sea of nucleosomes" model. An increase in fiber length is accompanied with an increase in structural irregularity and flexibility. **A,D)** In 4-nucleosome fibers, the increase in nucleosome repeat length allows for more complex nucleosome arrangements of α/β-structures. The leading diffusion map CV is strongly correlated with end-to-end distance of the fiber R_{end} ($\rho^{(4)}_{R_{end},\psi_1} = 0.88$). **B,E)** In 8-nucleosome fibers, the α-tetrahedron motifs contributes to local chromatin compaction while β-rhombus structures resemble the more ladder-like chromatin structure. The second CV is moderately correlated with end-to-end distance ($\rho^{(8)}_{R_{end},\psi_2} = 0.42$). **C,F)** In 16-nucleosome fibers, local chromatin motion is isotropic and largely driven by thermal fluctuations. As for the 8-nucleosome case, the second CV is moderately correlated with end-to-end distance ($\rho^{(16)}_{R_{end},\psi_2} = 0.69$).

the structural changes induced by the liquid-like behavior of chromatin and gene expression levels [8]. In addition, the changes in replication timing during cell differentiation has been suggested to involve the fluctuations of TAD structures [100]. Modeling studies using a simple polymer model of chromatin have shown that these TAD domains are susceptible to the liquid-like movement of chromatin [7]. The regulatory role of these intrinsically stable oligonucleosome motifs remains to be determined.

2.5 Conclusion

By analyzing long simulation trajectories generated by a coarse-grained multi-scale chromatin model using nonlinear manifold learning, we resolved the spontaneous and intrinsic formation within the chromatin fiber of α-tetrahedron and β-rhombus motifs – two previously characterized tetranucleosomal conformations that play an important role in the accessibility of DNA. Our analysis showed that both motifs represent metastable states that correspond to local free energy minima and whose formation are not dependent on external factors such as post-translational modifications or protein remodeling complexes. We also characterized the effects that varying linker DNA length and nucleosome orientation have on the formation of these motifs and, by extension, local chromatin compaction. We observed that local chromatin compaction is induced by α-tetrahedron motifs which allow sterically unfavorable conformations to form such as kinks along the fiber. β-rhombus conformations, on the other hand, were observed to form an open ladder-like chromatin structure which can facilitate DNA accessibility by external machinery such as transcription factors. Longer linker lengths are accompanied by an increase in structural irregularity and flexibility and, ultimately, a dynamic liquid-like "sea of nucleosomes" [28, 87] behavior that allows for constant structural reorganization. The lack of a cohesive structure for longer linker lengths suggests that they are less stable and more prone to opening and would require linker histones or some other chromatin architectural protein to fold into either conformation. This finding is in line with experimental observations which have shown that linker histone levels correlate positively with linker DNA length [135]. Here, we considered homogenous linker lengths and were able to identify the formation these two unique folding motifs. Hi-C methods, however, have revealed that these conserved motifs exist within heterogenous chromatin fibers as well [89]. The effects by which variable linker lengths and tetranucleosome motifs affect the free energy landscape of chromatin folding requires further investigation.

Taken together, our results suggest a two-state model for the local chromatin structure,

wherein extended β-rhombus motifs are more stable than compact α-tetrahedron motifs. It is possible that chromatin remodeling proteins and/or epigenetic status may be responsible for lowering the free energy barrier between the observed states. Future studies will require examining the mechanism by which linker histones and remodeling protein complexes, such as Polycomb, affect the emergence of metastable states. In addition, we are investigating the kinetic rates between each state using Markov state models (MSM) to elucidate dynamical links between chromatin structure and gene expression.

2.6 Materials and Methods

2.6.1 Simulating the Chromatin Fiber

Coarse-grained molecular dynamics simulations of chromatin fibers were conducted using the 1CPN model (Figure 1) [61]. The model was parameterized by explicit experimental measurements and finer-scaled models of DNA and proteins [61] and incorporates and preserves molecular-level nucleosome physics while enabling kilobase-scale simulations of genomic DNA. Each nucleosome is represented by a single anisotropic site and DNA as a spherical particle at a three basepair-per-bead resolution. We refer interested readers to the detailed descriptions found in prior publications [46, 61, 60].

Simulations of four, eight, and 16 nucleosomes at nucleosome repeat lengths (NRL) 157, 187, and 197 were conducted in a periodic box at a temperature of 300 K and a simulation time step of 60 fs. Solvent was represented implicitly with a Langevin thermostat and salt concentration of 150 mM. Nucleosome interactions were represented by the Zewdie potential, which has shown to be well-suited for representing inter-nucleosomal interactions [119]. To account for the stabilizing effects of the histone H3 tail of DNA entering and exiting the nucleosome, an additional pairwise interaction was introduced between the dyad and DNA sites. Electrostatic interactions were incorporated using Debye-Hückel theory

[86]. Simulations were carried out using the LAMMPS molecular dynamics package [98]. Chromatin fibers were initialized in an extended state with the angle between entering and exiting DNA and the nucleosome at 180°. The root mean square deviation (RMSD) with respect to the first given frame achieved a steady state after a ∼100 μs equilibration period. The final configuration from the relaxation procedure was used as the initial starting structure for our production run and analysis. Five 30 μs replicas were conducted totaling 150 μs of simulation time for each permutation and amounting to a total 1.35 ms of overall simulation time.

2.6.2 Diffusion Maps

Diffusion maps are a nonlinear manifold learning technique that have found extensive applications in generating low-dimensional embeddings of high-dimensional molecular trajectories [10, 19, 43]. Assuming that the distance metric used to compare pairs of configurational microstates is a good proxy for short-time kinetic distance and that the conformational dynamics over the state space may be approximated as a diffusion process, the leading collective variables of the diffusion map correspond to the large-scale, high-variance collective motions of the system, and kinetically close configurational microstates are embedded close together [32]. We employ the density-adaptive variant of diffusion maps, which we find to be particularly useful for handling the large inhomogeneities in sampling densities observed in our chromatin simulations [130]. We provide a brief summary of the approach below, but direct the reader to prior publications for mathematical and algorithmic details [10, 19, 32, 43].

Pairwise distances d_{ij} are calculated between data points in our set, x_i and x_j, which correspond to the RMSD between translationally and rotationally aligned nucleosomal coordinates in frames i and j of the simulation. A Gaussian kernel is applied to d_{ij} to

construct a threshold pairwise distance matrix $\mathbf{A}$,

$$A_{ij} = \exp\left(\frac{-d_{ij}^{2\alpha}}{2\epsilon}\right) \tag{2.1}$$

where ϵ is the kernel bandwidth and defines the local neighborhood of each point and α is a parameter that globally rescales pairwise distances to smooth out large density fluctuations between densely and sparsely sampled regions of configurational state space [130]. The matrix $\mathbf{A}$ is then row-normalized to form the transition matrix,

$$\boldsymbol{M} = \boldsymbol{D}^{-1}\,\boldsymbol{A} \tag{2.2}$$

where $\boldsymbol{D}$ is a diagonal matrix with elements,

$$D_{ii} = \sum_j A_{ij}. \tag{2.3}$$

The transition matrix, $\mathbf{M}$, is then diagonalized to calculate its eigenvectors ψ_{i} and eigenvalues λ_{i}. By the Markov property, the top eigenvalue-eigenvector pair ($\psi_0 = \vec{1}$, $\lambda_0 = 1$) is trivial, corresponding to the steady-state distribution of a random walk. A gap in the eigenvalue spectrum after the k^{th} non-trivial eigenvalue identifies the k-leading eigenvectors corresponding to the leading high-variance nonlinear collective modes of the system. Snapshot i of the molecular simulation trajectory is embedded into these collective variables spanning the so-called intrinsic manifold of the system under the mapping,

$$x_i \mapsto \left[\psi_1(i), \psi_2(i), \ldots, \psi_k(i)\right]. \tag{2.4}$$

The ψ_{k} are the leading nonlinear collective variables (CVs) identified by the diffusion map that correspond to the high-variance dynamical modes of the system and are responsible for large-scale conformational rearrangements.

Free energy surfaces over the intrinsic manifold $G(\Psi)$ are computed by collecting histogram approximations $\hat{P}$ to the observed distribution of configurational microstates projected into the leading k-eigenvectors $\Psi = \{\psi_i\}_{i=1}^{k}$ and then inverting this distribution using the relation,

$$\beta G\left(\Psi\right) = -\ln \hat{P}\left(\Psi\right) + C, \tag{2.5}$$

where $\beta = 1/(k_B T)$ is the inverse temperature and C is an arbitrary additive constant that sets an absolute free energy scale [113]. By virtue of the interpretability of the eigenvectors as the leading collective modes of the system, the free energy surface constructed over the intrinsic manifold can resolve both the metastable macrostates of the chromatin structure and the interconversion pathways between them [32]. Diffusion maps have already been used successfully to examine the dynamics of DNA around histone proteins, thereby providing precedent for our approach [43], but we note that we could have employed employing tICA, VAMPnets, or SRVs in conjunction with Markov state models to identify kinetic microstates and macrostates [17, 72, 88, 94, 110]. These approaches have the benefit of furnishing kinetic networks without requiring that the assumption of diffusive dynamics be made. In the present work, it is the structure and thermodynamics of the metastable states are of primary interest, as opposed to the kinetic transition rates, and for this reason we favor the smooth, continuous, and more structurally interpretable free energy surfaces furnished by diffusion maps.

2.7 Acknowledgements

The authors thank Dr. Tobin Sosnick and Dr. Vadim Backman for their generous feedback of this work. We also thank Aria Coraor, Soren Kyhl, Mike Jones, Kirill Shmilovich, Wei Chen, and Yutao Ma for their helpful discussions, and Xinran Lian for her assistance in preparing artwork. This study was supported by the NSF grant EFRI EEC 1830969. This work was completed in part with resources provided by the University of Chicago Research Computing

Center. The authors gratefully acknowledge computing time on the University of Chicago high-performance GPU-based cyberinfrastructure (NSF Grant No. DMR-1828629).

2.8 Supporting Figures

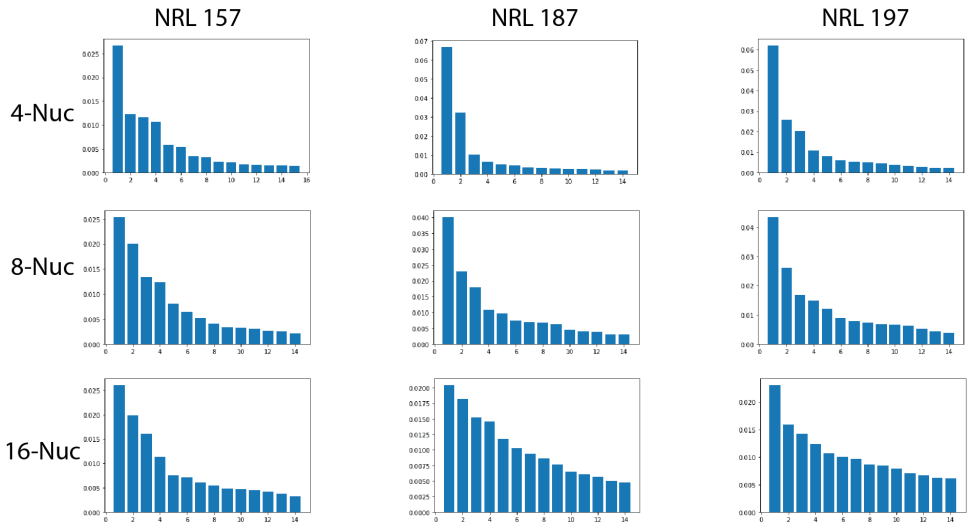

Figure 2.5: Analysis of the 4, 8, and 16-nucleosome systems at varying NRLs using diffusion maps reveal a leading gap in the eigenvalue spectra after the 1st non-trivial mode.

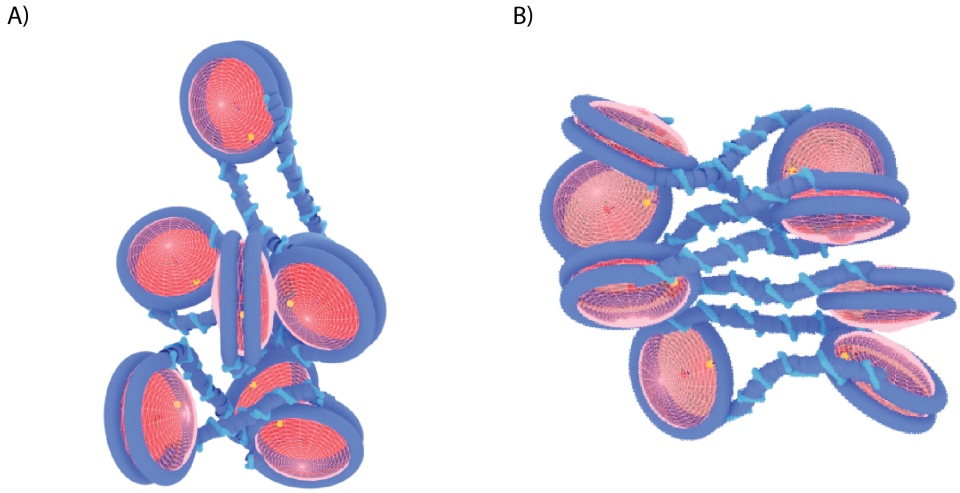

Figure 2.6: Two representative structures of the 8-nucleosome system at NRL 187 shows a high propensity for small α-tetrahedron (A) and β-rhombus clusters (B).

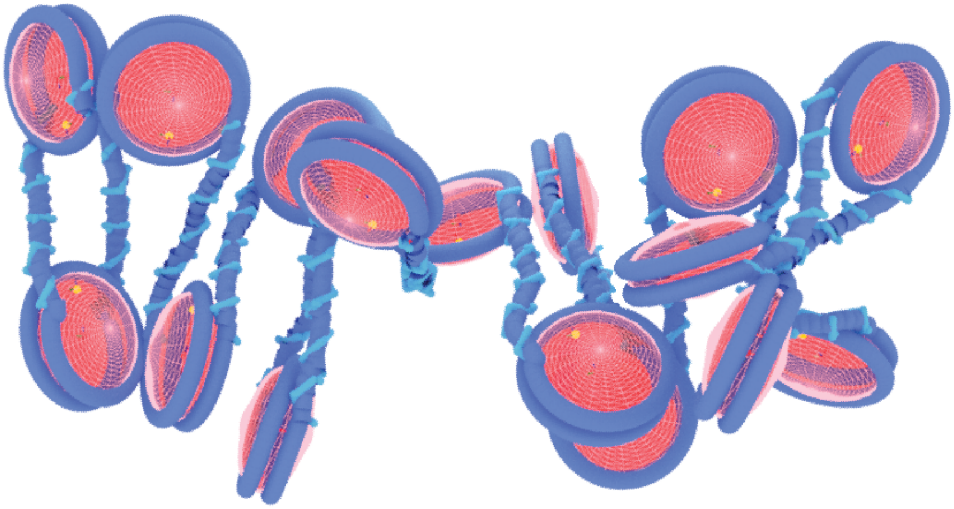

Figure 2.7: A fiber of 16-nucleosomes at an NRL of 197 transitioning between an open (left) to closed state (right). The α-tetrahedron and β-rhombus motifs can be considered as the folding units of the chromatin fiber.

CHAPTER 3

DENOISING AUTOENCODER TRAINED ON SIMULATION-DERIVED STRUCTURES FOR NOISE REDUCTION IN CHROMATIN SCANNING TRANSMISSION ELECTRON MICROSCOPY

Reprinted with permission from **Alvarado, W.**; Agrawal, V.; Li, W. S.; Dravid, V. P.; Backman, V.; de Pablo, J. J.; Ferguson, A. L. ACS Cent. Sci. 2023, 9, 6, 1200–1212. DOI: 10.1021/acscentsci.3c00178.

3.1 Author contributions

W.A., J.D.P., and A.F. conceptualized the study and interpreted findings. V.A. performed the domain analysis. V.A. and W.S.L. co-created the domain analysis platform and wrote custom code for mass scaling and radial density analysis. W.A., V.A, W.S.L., J.D.P., and, A.F. wrote the paper. V.P.D. supervised collecting sample images. J.D.P., V.B., and A.F. supervised the project. All authors participated in reviewing and commenting on the study drafts.

3.2 Abstract

Scanning transmission electron microscopy tomography with ChromEM staining (ChromSTEM), has allowed for the three-dimensional study of genome organization. By leveraging convolutional neural networks and molecular dynamics simulations, we have developed a denoising autoencoder (DAE) capable of post-processing experimental ChromSTEM images to provide nucleosome-level resolution. Our DAE is trained on synthetic images generated from simulations of the chromatin fiber using the 1-cylinder per nucleosome (1CPN) model

of chromatin. We find that our DAE is capable of removing noise commonly found in high-angle annular dark field (HAADF) STEM experiments and is able to learn structural features driven by the physics of chromatin folding. The DAE outperforms other well-known denoising algorithms without degradation of structural features and permits the resolution of α-tetrahedron tetranucleosome motifs that induce local chromatin compaction and mediate DNA accessibility. Notably, we find no evidence for the 30-nm fiber, which has been suggested to serve as the higher-order structure of the chromatin fiber. This approach provides high-resolution STEM images that allows for the resolution of single nucleosomes and organized domains within chromatin dense regions comprising of folding motifs that modulate the accessibility of DNA to external biological machinery.

3.3 Introduction

Chromatin is the highly organized complex of DNA, RNA, and proteins that packages DNA within the cell nucleus, prevents DNA damage, and controls replication and gene expression [127]. The main organizational unit of chromatin is the nucleosome core particle constituting a complex of DNA wrapped around a histone octomer. [76] Structurally, the nucleosome is approximately 146 base pairs (bp) of DNA wrapped in 1.67 left-handed superhelical turns around two copies of the H2A, H2B, H3, and H4 proteins. Chromosomes can contain hundreds of thousands of nucleosomes linked by short strands of DNA which give it the appearance of beads on a string. The structure of these 11-nm wide nucleosomal disks is nearly conserved across all eukaryotic cells and serves as the repeating building block of chromatin [133]. Beyond this basic structural unit, chromatin is believed to have several hierarchical levels of DNA packaging, beginning with a 10-nm fiber that further compacts into a 30-nm fiber, the latter of which has been considered to be a key intermediate level of chromatin organization and compaction within the eukaryotic nucleus [104]. The structure of the 30-nm fiber is characterized as a nucleosomal chain folding into a solenoid or a "one-start"

helical structure. Each nucleosome in this configuration interacts with its fifth and sixth surrounding nucleosomes as the nucleosomes coil around a central cavity at a rate of about six nucleosomes per turn [34]. Though first observed under an electron microscope *in vitro*, the relevance of the 30-nm fiber *in vivo* remains an open question [75, 125, 104]. More recently, studies have suggested nucleosomes can arrange themselves into stable secondary structural arrays comprised of four nucleosomes that play an important regulatory function by controlling the accessibility of DNA to external biological machinery [109, 118, 89, 1]. While these tetranucleosomes have been observed in reconstituted chromatin fibers *in vitro* and suggested by modeling studies *in silico*, current imaging techniques remain insufficient to resolve their existence *in situ* [3].

Recently, chromatin staining coupled with electron and scanning transmission electron microscopy (ChromEM and ChromSTEM, respectively) have resolved the 3D organization of chromatin and observed distinct, anisotropic packing domains [93, 63]. The size and variability of these domains across different cell types has been suggested to regulate gene activity by controlling the size of macromolecular complexes that can access DNA within these clusters thereby affecting processes such as DNA transcription, replication, and repair. In addition, variability in statistical and morphological properties of packing domains may potentially play an important role in the construction of higher-order chromatin structures such as euchromatin and heterochromatin [64]. While these experimental imaging techniques have provided key insights into chromatin structure, nucleosome-level packing remains obscured by statistical noise inherent to STEM imaging [3, 112]. In particular, the spatial organization of nucleosomes within dense chromatin regions suffers from low signal-to-noise ratios at these smaller length scales. Denoising STEM images provides a means to identify folding motifs and advance understanding of the details of chromatin structure, nucleosome packing, and the structure-function relation.

By combining the advances made in STEM imaging for chromatin, molecular dynamics

simulations, and machine learning, we designed a deep convolutional denoising autoencoder (DAE) for STEM image denoising. Since noiseless experimental images upon which to train our denoising models are not available, we instead generate noise-free training data using by molecular dynamics (MD) simulations. This strategy is similar to the approach employed by Ziatdinov *et al.* in studying the surface of molecular structures [141]. We conduct simulations of the chromatin fiber using the 1-cylinder per nucleosome (1CPN) model that has been shown to accurately reflect the possible conformations of oligonucleosomal structures [1, 61, 82]. Snapshots from these MD trajectories are then converted to synthetic ChromSTEM image datasets which are used to train the DAE to remove noise artificially added to the training images and produce images with greatly enhanced structural resolution that enable the identification and analysis of folding motifs within dense DNA regions. The DAE outperforms other well-known denoising algorithms and, as we demonstrate in applications of the trained model to experimental ChromSTEM images, resolves specific tetranucleosome motifs that induce local chromatin compaction and are known to mediate DNA accessibility. Notably, we find no evidence for the 30-nm fiber, which has been suggested to serve as the higher-order structure of the chromatin fiber [28, 68]. Our machine-learning enabled DAE presents a means to bridge experimental ChromSTEM imaging and physics-based molecular dynamics simulations to realize high-resolution, denoised images capable of resolving previously unidentifiable tetranucleosome motifs to advance understanding of the small-scale organization of chromatin and the relationship of structure to function.

3.4 Methods

3.4.1 Coarse-Grained Molecular Dynamics Simulations and Generation of Synthetic STEM Data

We train our DAE on tomographic images generated from MD simulations of the chromatin fiber (Figure 4.1). To generate a synthetic dataset, coarse-grained molecular dynamics simulations were carried out using the 1-cylinder per nucleosome (1CPN) model of chromatin [61]. The 1CPN model is parametrized by explicit experimental measurements and atomistic models of DNA that preserve molecular-level nucleosome physics enabling kilobase-scale simulations of genomic DNA. The 1CPN model is an appropriate choice since it has been extensively validated in the literature as a reliable model for capturing chromatin dynamics [61]. The model was fitted against experimental data and has demonstrated its ability to reproduce a wide range of chromatin processes that include nucleosome unwrapping, sedimentation coefficients, and interactions between nucleosomes, which is a primary mechanism that drives chromatin folding [82, 1].

We conducted the 1CPN simulations under conditions representative of those under which the ChromSTEM images were acquired. As anticipated, the 30-nm fiber was not observed within in our simulations as the conditions that typically involve its formation are due to specific *in vitro* environmental conditions such as the inclusion of high-affinity 601 DNA repeat and a cationic environment (e.g., 1–2 mM Mg^{2+}) [52]. Furthermore, cryo-EM images of the 30-nm fiber have not been reported for mitotic chromosomes *in vivo* [68]. We note however, that our pipeline is designed to be easily adaptable to new conditions, and that transfer learning could be used to augment the existing model by repeating the simulations under the conditions under which the new experimental data was gathered and retraining the DAE.

After equilibration, three 30 μs replicas were conducted totaling 150 μs of simulation time

of chromatin fibers varying from 150-200 nucleosome repeat length (NRL) and comprised of 4-16 nucleosomes. The lengths and sizes were chosen to account for the natural variability in biological systems. We highlight that our simulations cover long time scales that have not been reached by previous studies. This extended simulation time allows for a more comprehensive exploration of the phase space and reduces the risk of being trapped in certain emergy minima. The 1CPN model's effectiveness in representing chromatin behavior helps to ensure that our simulation snapshots are indeed representative of the physical system under study. The combination of long time scale simulations and the use of the 1CPN model provides a strong foundation for generating a diverse and representative training dataset for our denoising autoencoder. We performed an internal consistency verification that the 150 μs simulations of each system were sufficiently long to comprehensively probe the relevant configurational phase space by verifying that the phase space ensemble visited by the first 75 μs and second 75 μs produced similar distributions in key structural order parameters such as radius of gyration and root-mean-square deviation in reference to the initial elongated fiber structure.

Approximately 16,000 snapshots from all simulation trajectories were extracted at 28×28 pixel resolution. These synthetic images represent a variety of conformations of the chromatin fiber at a resolution commensurate with that of typical ChromSTEM imaging experiments [63, 64]. From this dataset, 12,702 conformations were selected for training and 3,176 held out as a validation set. An x-ray crystal structure of the nucleosome core particle at 1.9 Å resolution (PDB:1KX5) was superimposed to the location of each nucleosome bead and linker DNA was built with repeating ATAT bases [23]. Each structure was converted to a point cloud representation and then voxelized to resemble a high-angle annular dark-field scanning transmission electron microscope (HAADF-STEM) tomogram. Each synthetic image stack contained 28×28×9 voxels with a voxel dimension of approximately 3×3×3 nm^3 corresponding to the approximate 27 nm^3 volume captured in an experimental STEM voxel.

Mathematically, the voxel intensity, $I_{m,n}$, is given by the total number of atoms that are enclosed within the volume of a voxel unit, $V_{m,n}$,

$$I_{m,n}(x) = \sum_{i=1}^{N} \left[x_i \in V_{m,n} \right], \tag{3.1}$$

where the position of a given atom is given by x_i, m and n denote a voxel constituting the total image, I, and where we have used Iverson's bracket notation to denote the indicator function. Finally, the synthetic image intensity is normalized to match the distribution of voxel intensity in experimental tomograms [111, 91, 103].

HAADF-STEM has emerged as a powerful imaging technique that provides nanoscale--level structural detail [116, 58]. It is, however, sensitive to environmental and instrumental noise during image acquisition that introduces extraneous signals not associated with the scattering of the sample [112, 9, 12]. For example, images are acquired at different projection angles by tilting the sample stage, at high-tilt angles; however, focusing becomes more difficult which leads to image blurring [30]. In addition, limited beam penetration and focal depth coupled with the restricted tilt range results in a lower set of projections which also introduces artifacts (i.e., "missing cone" artifacts) [79, 22]. Beam damage and environmental noise (e.g. airflow, sound, temperature, etc.) also deteriorate image quality and limits the accuracy of HAADF-STEM tomographic reconstruction [67, 112, 9]. Due to the particle nature of electrons and collection method, Poisson noise remains the dominant form of noise in STEM imaging [78, 112]. To account for these effects within our simulated data, we apply several HAADF-STEM related noise conditions such as Gaussian noise, Poisson noise, and tip-blurring effects to each simulated image similar to the approach implemented by Schwenker *et al.* [111, 91, 103]. Parameters such as broadening effects, counts, and additive background noise were adjusted to account for the different levels of noise that may be encountered during image acquisition. Mathematically, each noise-free image, I, generated

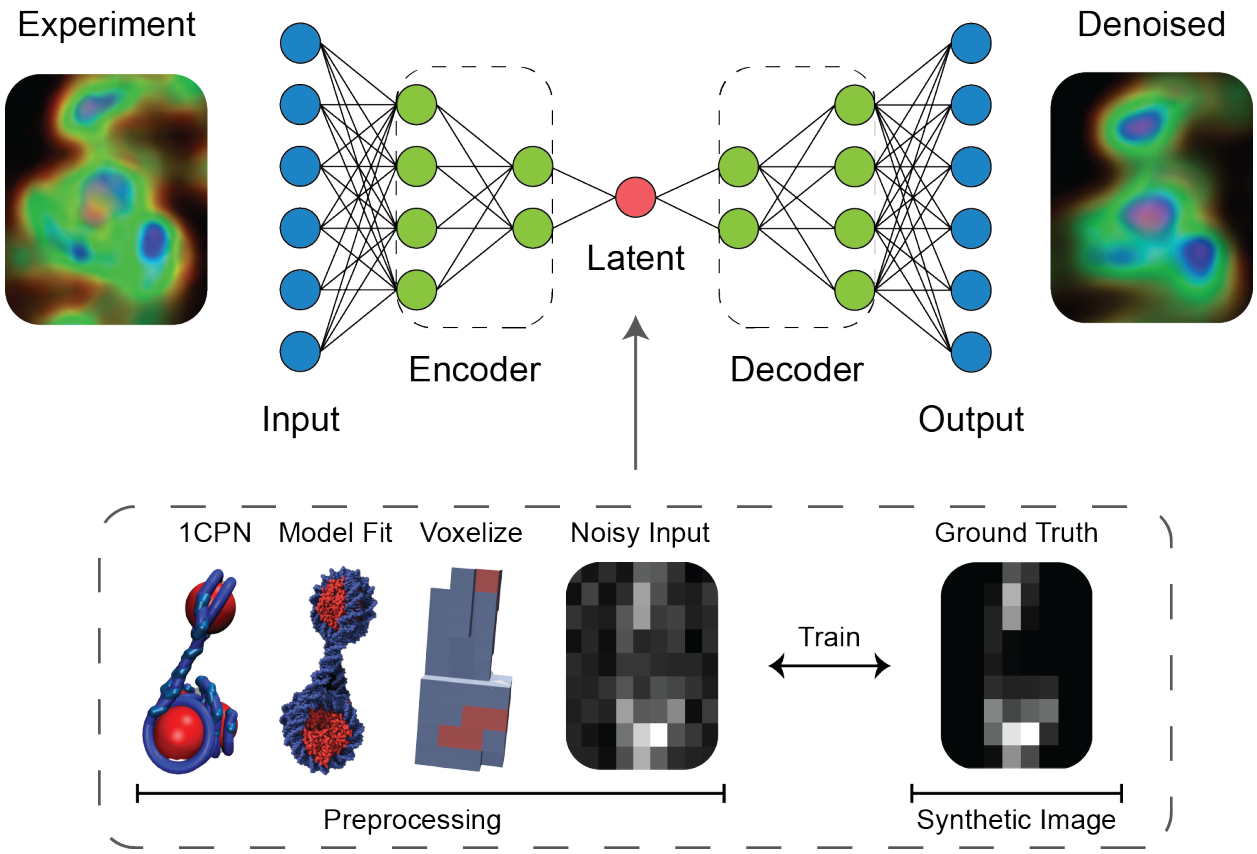

Figure 3.1: A denoising autoencoder (DAE) is constructed and trained on simulations of the chromatin fiber. We simulate nucleosome arrangements using the 1CPN model of chromatin and use the resulting trajectories to generate synthethic STEM images by superimposing crystal structures of the nucleosome (PDB: 1KX5) and DNA snippets. Noise commonly found in angle annular dark field (HAADF) STEM experiments is applied to the images and the DAE trained to remove this noise and preserve the underlying signal.

from the MD simulations is converted into an artificially noisy image, $\tilde{I}$, by corrupting it with articial noise under the noise model:

$$\tilde{I} = I + I_{Poisson} + I_{Gaussian} + I_{Scan}. \tag{3.2}$$

Given that Poisson noise is not additive and correlated with voxel intensity, we instead begin by applying a signal-dependent Poisson noise layer on top of each noise-free image using the

discrete probability distribution:

$$I_{Poisson} \sim Pr(N = k) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{3.3}$$

where N represents the number of photons measured by a given sensor and λ is the expected number of photos per unit time interval. We make the assumption that the number of atoms counted in a given voxel unit ($I_{m,n}$) is similar to photon counting in a classic Poisson process.

STEM images are susceptible to thermal vibrations and electronic noise which can be modeled as a Gaussian process [55]. To account for this, we add a Gaussian noise layer that obeys the distribution:

$$I_{Gaussian} \sim N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(z-\mu)^2}{2\sigma^2}} \tag{3.4}$$

where μ is equal to the mean of the image and σ is the standard deviation which represents the broadening (i.e., "spread") of the signal. Similar to the approach by Schwenker *et al.* to emulate noise and distortion conditions common to the HAADF-STEM imaging mode, we set $\sigma = 0.8$ [111, 91, 103].

Finally, scan line shifts, I_{Scan}, are random, persistent, time-dependent distortions that occur due to positioning errors of the electron beam that result in shifts in the image perpendicular to the scan lines [71]. We generated this type of noise by introducing approximately a 1 subpixel offset randomly along the x-direction and resampling these random shifts via bilinear interpolation:

$$I_{Scan} \sim I_{u_x,u_y} = I_{m,n}(u_x, u_y) \tag{3.5}$$

where u_x and u_y are the desired shift across the range $[-1, 1)$.

3.4.2 Denoising Autoencoder (DAE) Architecture

As the name suggests, denoising autoencoders (DAE) are artificial neural networks designed to remove noise from an input signal, frequently images [129]. A typical autoencoder is comprised of two distinct components: an encoder and decoder. The encoder compresses a high-dimensional image into a low-dimensional representation. These representations are called latent representations or encodings which the decoder uses to reconstruct the original input image. During training, the DAE is provided with training images that have been artificially corrupted with noise generated by a model representative of the noise expected to be encountered in the particular application domain. A loss function is applied that minimizes the difference between the reconstructed image and the original noise-free image. Intuitively, the training process teaches the DAE to learn a latent space representation that filters out the noise while preserving the underlying signal within the training data and permits the decoder to reconstruct denoised images [128]. The trained DAE model may then be applied to noisy images outside of the training data for which the ground truth is unknown to predictively reconstruct denoised images. The success and generalizability of the trained model is contingent on the training images and noise model being sufficiently representative of the new images to which it is applied and it is good practice to perform *post hoc* checks that the model has not introduced artifacts or been applied outside of its domain of applicability.

We employ a fully convolutional DAE architecture that permits variable input image sizes to allow for potential variability in training and experimental image sizes [65]. A training and validation set of 12,702 and 3,176 images (80/20 random split) with 28×28 dimensions at a batch size of 32 was used for training and validation (Figure 4.2). We guard against overfitting by employing early stopping based on the validation error on a 20% randomly sampled hold-out validation partition. These images were harvested from the 1CPN MD simulations and contain a diversity of conformations of chromatin fibers at a resolution

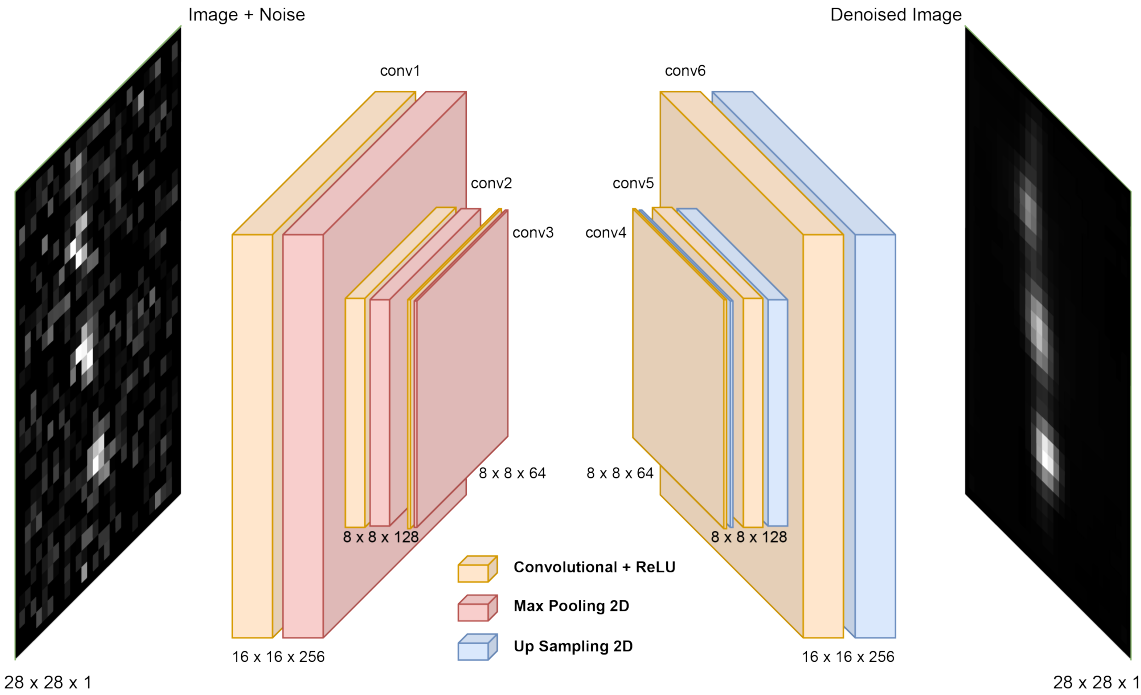

Figure 3.2: A denoising autoencoder (DAE) comprises an encoder that compresses the noisy image into a low-dimensional latent space embedding and a decoder that decompresses this embedding into a denoised image. The latent space presents an information bottleneck that the trained DAE model uses to reject noise and preserve signal, enabling reconstruction of denoised images. The DAE is trained on noise-free images for which the ground truth is known and which are artificially corrupted by noise under a noise model representative of the intended application domain for the trained DAE. The image illustrates a DAE that performs an encoding of a 28×28 pixel greyscale (i.e., single channel) image into a 64-channel 8×8 latent space embedding under three convolution plus max pooling layers, followed by decoding under three convolutional plus upsampling layers to generate a denoised 28×28 pixel image [53] .

commensurate with that of typical ChromSTEM imaging experiment. We use a convolution layer of kernel size (3,3) with 256 output filters, and stride 1 employing ReLU activation functions and followed by a max pooling layer of pool size (2,2). We follow this with a second ReLU convolutional layer of kernel size (3,3), 128 output filters, and stride 1 followed by a max pooling layer of pool size (2,2), and finally a third ReLU convolutional layer of kernel size (3,3), 64 output filters, and stride 1 followed by a max pooling layer of pool size (2,2). The output of the third convolutional layer produces a low-dimensional latent space

embedding of the image that serves as an information bottleneck designed to preserve the image signal and reject noise. The decoder is symmetric to the encoder structure, employing three convolutional upsampling layers used to rebuild images to their original dimension. Our network employs a fully convolutional architecture that does not use any fully connected layers and enables its deployment on images of arbitrary size. Given that images comprise single channel grayscale pixels with intensities normalized between [0,1], the binary cross-entropy (BCE) loss function is used:

$$\mathrm{BCE} = -\frac{1}{\substack{\text{output}\\\text{size}}} \sum_{i=1}^{\substack{\text{output}\\\text{size}}} y_i \cdot \log \hat{y}_i + (1 - y_i) \cdot \log\left(1 - \hat{y}_i\right), \tag{3.6}$$

where $\hat{y}_i$ is the output prediction and y_i is the corresponding target value. It has been shown that when training autoencoders on image data, minimizing the BCE loss function facilitates gradient steps in data space from low to high probability regions under the data-generation distribution [20].

We constructed and trained our DAE in TensorFlow using Keras [18]. Training took ∼3 minutes per epoch on an AMD Ryzen 9 3950X 16-core CPU and Nvidia RTX 3090 GPU card. Training was performed using the Adam algorithm with a learning rate of 1×10^{-3} [56]. We guard against overfitting by employing early stopping based on the validation error on a 20% randomly sampled hold-out validation partition. We explored architectures employing 3-6 convolutional layers, first layer filters ranging from 2×2-5×5, and latent spaces bottlenecks ranging from 2×2×12-16×16×128, but found our result to be relatively insensitive to the precise choice of architecture. Source code for our DAE and training/validation data are available at `https://github.com/Ferg-Lab/ChromSTEM-Denoising-Autoencoder`.

3.4.3 Denoising Performance

Denoising performance was measured using mean-square error (MSE), peak signal-to-noise ratio (PSNR), and structural similarity index (SSIM) [132, 31]. Mean-square error is the total squared error between pixel intensity differences of the original noise-free image, I, and denoised image, $\hat{I}$, defined as:

$$MSE = \frac{\sum_{m=1}^{M} \sum_{n=1}^{N} \left[I_{m,n} - \hat{I}_{m,n} \right]^2}{MN} \tag{3.7}$$

where M and N are the number of rows and columns in the image, and $M{=}N{=}28$ for our training data. The lower the MSE value the lower the error. Similarly, PSNR measures the quality of reconstruction of lossy compression by measuring the peak error and is calculated as:

$$PSNR = 10 \log_{10} \left(\frac{R^2}{MSE} \right) \tag{3.8}$$

where R is the maximum possible pixel value and typically depends on the bit depth of an image (e.g., for 8-bit images $R = 255$ [48]. For PSNR, the higher the value the better the reconstruction.

Whereas MSE and PSNR calculate absolute errors between pixels, the SSIM index considers degradation as the change of perception in structural information by taking into account three key features: luminance, contrast, and structure. An SSIM value can range from (-1) indicating images are structurally different or (+1) indicating they are either the same or very similar, and is defined as:

$$SSIM(x, y) = [l(x, y)]^a \times [c(x, y)]^\beta \times [s(x, y)]^\gamma \tag{3.9}$$

where,

$$l(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}, \tag{3.10}$$

$$c(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}, \tag{3.11}$$

$$s(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}. \tag{3.12}$$

The functions $l(x,y)$, $c(x,y)$, and $s(x,y)$ compare luminance, contrast, and structure between two images x and y, where here we set $x = I$ and $y = \hat{I}$ for our ground-truth and denoised images, respectively [48]. The variables μ_x and μ_y are their respective local means over all pixel values and represent the luminance of each images. Contrast is measured by taking the standard deviation σ_x and σ_y of all pixel values and σ_{xy} is the cross-covariance of the images. The variables α, β, and γ adjust the relative importance of each feature and are typically set to unity. The constants $C_i = (K_i L)^2$ prevent functions from becoming undefined, where L accounts for pixel value range and is set to unity given that our images are normalized in the range of [0,1]. By convention, we adopt $C_3 = C_2/2$ and set $K_1 = 0.01$ and $K_2 = 0.03$ [132].

Denoising performance metrics such as MSE, PSNR, and SSIM are calculated between a ground-truth image (i.e., noise-free image), I, and its denoised counterpart, $\hat{I}$ produced by the DAE from the artificially noisy image $\tilde{I}$. Given that noise-free ChromSTEM images do not exist to serve as a ground-truth comparison, we rely on power spectral density plots (PSD) to compare raw and denoised experimental image sets. PSD represents the total signal power contributed across the frequency domain of a signal. For images, it measures the strength of the features at different resolutions. This allows for comparison of morphological features and noise in the low and high wavelength domains, respectively. We compute the PSD by taking the discrete Fourier transform (DFT) of each image which allows for the

decomposition of resolutions,

$$F(k,l) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} I_{m,n} \exp\left\{-2\pi i \frac{mk}{M}\right\} \exp\left\{-2\pi i \frac{nl}{N}\right\} \tag{3.13}$$

where $I_{m,n}$ is representation of the image in the spatial domain corresponding to the greyscale intensity of the pixel at row (m) and column (n) coordinates, $F(k,l)$ is the representation of the image in the Fourier domain corresponding to the Fourier component at discrete row-wise and column-wise "frequencies" k/M and l/N, and $k = 0, \ldots, (M-1)$ and $l = 0, \ldots, (N-1)$ [117, 45]. Since we only consider square images for which $M = N$, we simplify this expression to equalize the row and column frequency components by setting $k = l$ so that,

$$F(k) = \sum_{m=0}^{M-1} I_{m,n} \exp\left\{-4\pi i \frac{mk}{M}\right\}. \tag{3.14}$$

The PSD follows from the modulus of the DFT as $P(k) = |F(k)|$.

3.4.4 ChromSTEM Sample Preparation, Imaging, and Reconstruction for A549 Cell Nucleus

Adenocarcinoma human lung epithelial cell line, A549 (ATCC Manassas, VA) was cultured in Dulbecco's Modified Eagle Medium (ThermoFisher Scientific, Waltham, MA, #11965092) and maintained at 5% CO2 and 37° C. All culture media were supplemented with 10% fetal bovine serum (Thermo Fisher Scientific, Waltham, MA; #16000044) and penicillin-streptomycin (100 μg/ml; Thermo Fisher Scientific, Waltham, MA; #15140122). The cell line was tested for mycoplasma contamination with Hoechst 33342. Cells were seeded on 35-mm glass-bottom petri dishes (MatTek Corp.) until approximately 40-50% confluent, and were given at least 24 hours to adhere to the dish before fixation.

For ChromSTEM sample preparation, the previously published protocol was adapted [93].

A549 cells cultured on the glass bottom dishes were thoroughly rinsed three times in Hank's balanced salt solution without calcium and magnesium (EMS). A fixation solution (2.5% EM grade glutaraldehyde, 2% paraformaldehyde, 2 mM $CaCl_2$ in 0.1 M sodium cacodylate buffer, pH = 7.4) was prepared. Cells were then fixed at room temperature for five minutes and then replaced with fresh fixative and fixed on ice for an hour. All the succeeding steps, unless mentioned otherwise, were performed on ice. After fixation, the cells were then washed with 0.1 M sodium cacodylate buffer five times on the ice. The samples were incubated in a blocking buffer (10 mM glycine, 10 mM potassium cyanide in 0.1 M sodium cacodylate buffer, pH = 7.4) for 15 minutes. Next, the samples were stained with 10 μM DRAQ5^{TM} (Thermo Fisher) and 0.1% saponin solution in 0.1 M sodium cacodylate buffer, pH = 7.4 for 10 minutes. The cells were washed with a blocking buffer twice, and then incubated in the blocking buffer on ice before photo-bleaching. The blocking buffer was replaced with 2.5 mM of 3-5'-diaminobenzidine (DAB) solution (Sigma Aldrich) in 0.1 M sodium cacodylate buffer, pH = 7.4 during photo-bleaching which was performed on a cold stage developed in-house from a wet chamber and equipped with humidity and temperature control.

A continuous epi-fluorescence illumination (150 W Xenon Lamp) with Cy5 red tilter with a 100× objective was used to bleach a spot – a random field of view with several cells – on the dish for 7 minutes on the cold stage. After photo-bleaching, the cells were washed with 0.1 M sodium cacodylate buffer five times. Reduced osmium solution (EMS) containing 2% osmium tetroxide, 1.5% potassium ferrocyanide, 2 mM $CaCl_2$ in 0.15 M sodium cacodylate buffer, pH = 7.4 was then used to stain the cells for 30 minutes on ice. The cells were then washed with double distilled water for five times on ice. Next, serial ethanol dehydration (30%, 50%, 70%, 85%, 95%, 100% twice) was performed on ice, and the last 100% ethanol wash was performed at room temperature. Durcupan resin (EMS) was used for infiltration and embedding. Resin mixture 1 was prepared by mixing - (i) 10 mL Durcupan TM ACM single component A, M, epoxy resin, (ii) 10 mL Durcupan TM ACM single component B, hardener 964, and (iii) 0.15

mL Durcupan TM ACM single component D. A 1:1 infiltration mixture containing equal proportions of 100% ethanol and Durcupan TM resin mixture 1 was used to infiltrate cells for 30 minutes at room temperature. Next, 2:1 infiltration mixture containing 5 mL 100% ethanol and 10 mL Durcupan TM resin mixture 1 was used to infiltrate the cells for two hours at room temperature. Durcupan TM resin mixture 1 was used to infiltrate the cells at room temperature for one hour. Resin mixture 2 was prepared by adding 0.2 mL Durcupan TM ACM, single component C, accelerator 960 to mixture 1 (10 mL of component A, 10 mL of component B, and 0.15 mL of component D). Durcupan TM resin mixture 2 was used to infiltrate the cells at 50°C in the dry oven for one hour.

The cells were embedded flat with fresh Durcupan TM resin mixture 2 in BEEM capsules and cured at 60°C in the dry oven for 48 hours. An ultramicrotome (UC7, Leica) was used to prepare 100 nm thick sections that were deposited onto a copper slot grid with carbon/Formvar film. Then, 10 nm colloidal gold fiducial markers were deposited on both sides of the sample. A 200 kV cFEG STEM (HD2300, HITACHI) with HAADF mode was used to collect all images. While keeping the field of view constant, the sample was tilted from -60° to 60° with 2° increments on two roughly perpendicular axes, with a pixel dwell time of $\sim$5 μs during image acquisition. Each tilt series was aligned with fiducial markers in IMOD and reconstructed using Tomopy with a penalized maximum likelihood for 40 iterations independently. [57, 44] The final tomogram is a 3D image size of 1230 $\times$ 1230 $\times$ 100 nm with a nominal voxel size of 2.9 nm.

3.5 Results and Discussion

Tetranucleosomes are widely considered the building block of the chromatin fiber and have been crystallized and observed in cryo-EM images of longer chromatin fibers [118]. Recent studies have suggested the existence of two tetranucleosome motifs that regulate gene expression – the α-tetrahedron and β-rhombus (Figure 4.3A) [89, 1]. Experiments and modeling studies

have indicated that these two energetically stable conformations may induce local chromatin compaction (α-tetrahedron) or the formation of elongated aggregates (β-rhombus) and are therefore proposed to play important regulatory and epigenetic roles in the accessibility of DNA to external machinery such as transcription factors [89, 27, 122, 1]. While ChromSTEM has been able to resolve variably packed nucleosomes and linker DNA segments at $\sim$2 nm spatial resolution, the variation of size, density, and shape of chromatin rich regions can obstruct finer-scale resolution of the structural arrangement of nucleosomes (Figure 4.3B). The structural resolution is also degraded by Poisson (i.e., shot) noise associated with electron counting statistics and the relatively poorer performance of segmentation (i.e., differentiation of background and chromatin signal by voxel intensity) within chromatin rich regions relative to regions where nucleosomes are well-separated and have uniform intensity [112]. We develop a machine learning-assisted computational denoising platform by training a denoising autoencoder (DAE) over coarse-grained molecular dynamics simulations, and apply the DAE to *in situ* high-resolution HAADF ChromSTEM microscopy images of chromatin within mammalian cell lines to resolve tetranucleosome motifs.

3.5.1 Testing on Synthetic Data

To validate our trained DAE, we first tested its performance against standard denoising techniques in an application to synthetic ChromSTEM images to which artificial noise was added and the ground truth (i.e., noise-free) images were exactly known. We collected 3000 test images harvested from 1CPN MD simulations of chromatin fibers varying from 150-200 nucleosome repeat length (NRL) and comprised of 4-16 nucleosomes and converted these into noise-free images I and noisy images $\tilde{I}$ using Eqns. 3.1 and 3.2. Importantly, the test set data was never exposed to the DAE at any point during its training. We report in Table 3.1 the denoising performance of our DAE compared to the popular non-local means (NLM) and block-matching and 3D filtering (BM3D) techniques [11, 21]. Performance is assessed

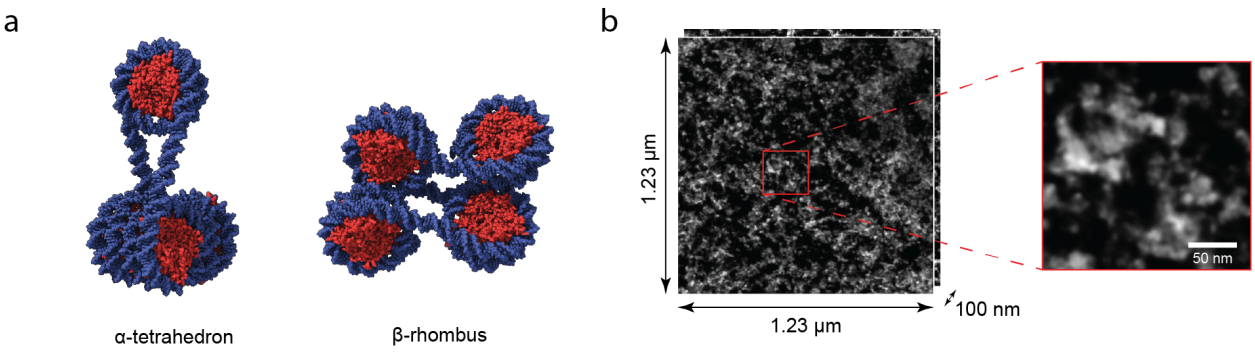

Figure 3.3: Resolution in dense chromatin regions is obstructed by the intrinsic noise of STEM imaging. a) The α-tetrahedron and β-rhombus tetranucleosome motifs have been proposed to play a regulatory and epigenetic role in the accessibility of DNA to external cellular machinery. The α-tetrahedron promotes DNA compaction whereas the β-rhombus results in elongated chomatic structures. Histone proteins are colored in red and DNA is colored in blue. b) In this work we employ high-resolution ChromSTEM tomograms comprised of 33 slices at 1.23 μm $\times$ 1.23 μm $\times$ 100 nm. The structural resolution accessible to experimental ChromSTEM tomograms is limited by the conformational variability of chromatin within chromatin-rich regions, Poisson noise, and the ability of image segmentation approaches to differentiate background and chromatin signal by voxel intensity.

using the mean square error (MSE), structural similarity index (SSIM), and peak signal-to-noise ratio (PSNR) metrics that are commonly used to benchmark denoising methods [31]. Better performance is associated with a reduction in cumulative squared error between the compressed and the original image (lower MSE), an increase in the ratio between the maximum possible power of an image and the power of corrupting noise (higher PSNR), and preservation of structural information between the reference and denoised image (higher SSIM). We present in Figure 4.4 an illustrative example of the application of each of the three denoising approaches to a representative snapshot taken from the 3000 test images.

Our DAE performed the best in all three denoising performance metrics (MSE = 0.003, SSIM = 0.83, PSNR = 26 dB), followed by BM3D (MSE = 0.007, SSIM = 0.55, PSNR = 22 dB) and non-local means (MSE = 0.011, SSIM = 0.15, PSNR = 20 dB). This represents a 57% improvement in MSE relative to BM3D and 72% improvement over non-local means (Table 3.1). From the example in Figure 4.4, we can see that not only is our denoising autoencoder able to remove the applied Gaussian and Poisson noise, but also has the ability to account for

Table 3.1: The mean and standard deviation for 3000 synthetic ChromSTEM test images was calculated to compare the denoising performance of our DAE against non-local means (NLM) and block-matching and 3D filtering (BM3D). Snapshots were harvested from 1CPN MD simulations of chromatin fibers varying from 150-200 nucleosome repeat length (NRL) and comprised of 4-16 nucleosomes and converted into noise-free images I and noisy images $\tilde{I}$ using Eqns. 3.1 and 3.2. Denoising performance is compared using the mean square error (MSE), structural similarity index (SSIM), and peak signal-to-noise ratio (PSNR) metrics. The DAE outperforms non-local means and BM3D along all three performance metrics (low MSE, high PSNR, high SSIM).

Denoiser	MSE	SSIM	PSNR (dB)
NLM	0.011 ± 0.003	0.15 ± 0.04	20 ± 1
BM3D	0.007 ± 0.004	0.55 ± 0.17	22 ± 2
DAE	$\mathbf{0.003 \pm 0.001}$	$\mathbf{0.83 \pm 0.04}$	$\mathbf{26 \pm 2}$

distortions which are typical to STEM experiments by virtue of the fact that it was trained on 1CPN molecular dynamics training data that preserve the physically representative structure of the chromatin strand. Given that denoising autoencoders are inherently lossy compression methods, some fuzzy imaging or loss of information is expected during the encoding process which can lead to broader output signals. The primary goal of our DAE method is to achieve a balance between noise reduction and preservation of structural features in the ChromSTEM images. While it might be possible to reduce these broader signals further, doing so could compromise the performance of the DAE or lead to overfitting.

We do observe that although our test does expose the DAE to novel synthetic ChromSTEM images it has not before encountered, they are generated using the same model as the training data. Conversely, the non-local means and BM3D approaches are standard algorithms that are not trained over images from a particular domain and are more general purpose denoising tools. As expected, the DAE appears to have learned to distinguish the physical arrangement of nucleosomes along the chromatin fiber within the physics-based simulation training data from the applied noise model, and can use these learned patterns to effectively denoise new synthetic ChromSTEM images that it has not previously encountered. A possible cost of this learning is, of course, that the DAE will likely not serve as a good general purpose,

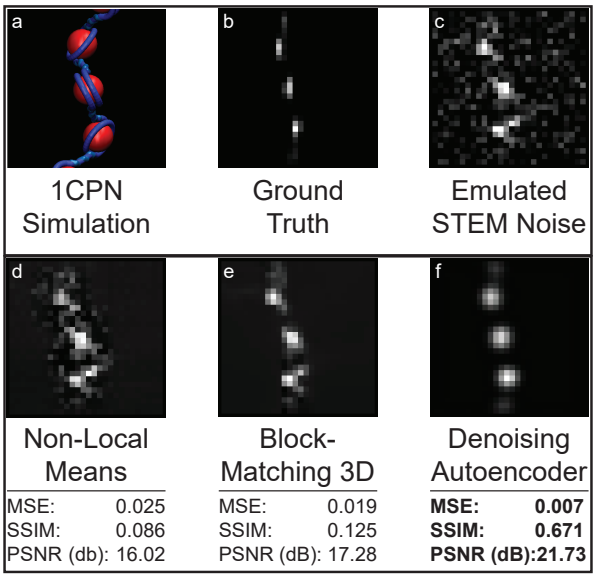

Figure 3.4: Illustrative example of DAE denoising performance to one selected synthetic ChromSTEM test image harvested from the 1CPN MD simulations. a) The selected snapshot was harvested from 1CPN MD simulations of chromatin fibers varying from 150-200 nucleosome repeat length (NRL) and comprised of 4-16 nucleosomes. b) The noise-free synthetic ChromSTEM image I was constructed from the MD snapshot using Eqn. 3.1. This constitutes the ground truth image against which we evaluate denoising performance. c) The noisy image $\tilde{I}$ was generated by adding artificial noise representative of that found in angle annular dark field (HAADF) STEM experiments to the noise-free image using Eqn. 3.2. The denoised image $\hat{I}$ produced from the noisy test image by d) non-local means (NLM), e) block-matching and 3D filtering (BM3D), and f) the DAE. The DAE outperforms NLM and BM3D along all three performance metrics (low MSE, high PSNR, high SSIM) for this particular image and over all 3000 test images (cf. Table 3.1).

application agnostic denoising algorithm in the same manner as non-local means and BM3D.

3.5.2 Application to Experimental Data

After validating that our DAE was capable of removing noise while preserving local structural features from our synthetic dataset, we move to apply it to experimental ChromSTEM images of chromatin. Figure 4.5 shows the difference between a raw and denoised experimental tomogram of an imaged human pulmonary adenocarcinoma epithelial cell (A549 cell). A pseudo-color gradient as opposed to a single grayscale channel is employed to display pixel

intensity for better visibility and to more clearly highlight the features within the image. Visual inspection of the denoised experiment confirms the ability of our DAE to remove noise and its ability to better resolve nucleosomes within chromatin-dense regions. Closer inspection of a randomly selected region of the denoised image (Figure 4.5b,e,f) clearly reveals the existence of clusters of a few nucleosomes that previous studies have suggested may play a role in the formation of topologically associated domains (TADs) in chromatin biology, and which are much less clearly resolved in the original image (Figure 4.5a,c,d) [89]. We also compare the power spectral density (PSD) of the raw and denoised image stacks (Figure 4.5g). We see good agreement of the PSD at lower wavenumbers, which correspond to the large-scale (i.e., low-frequency) structural and morphological features of the image. At higher wavenumbers, the PSD of the denoised image exhibits a linear decrease relative to the raw image, which can be interpreted as the attenuation of small-scale (i.e., high-frequency) noise in the experimental image. Taken together, these results indicate that the important structural signal within the experimental ChromSTEM image is preserved by our denoising approach and produces superior resolution of nucleosome-level features within the chromatin-rich regions of the image.

To determine whether these small nucleosomal clusters are comprised of either of the two recently identified folding motifs (α-tetrahedron or β-rhoumbus), we visually inspect a number of nucleosome clusters extracted from chromatin-rich regions within a 50 $\times$ 50 nm section of the experimental tomogram (Figure 4.6). It is challenging to discern from inspection of the raw image, but after passage through the DAE it is visually apparent that these chromatin-dense regions are primarily composed of tetranucleosome motifs (Figure 3.7). To quantify our assertion, we construct a density map from our denoised STEM image stack and fit a prototypical α-tetrahedral tetranucleosome folding motif reconstructed from a single atomic nucleosome structure (PDB:1KX5) [23]. To find an optimal fit, the cross-correlation coefficient (CCC) score was used to maximize the fit of a simulated map

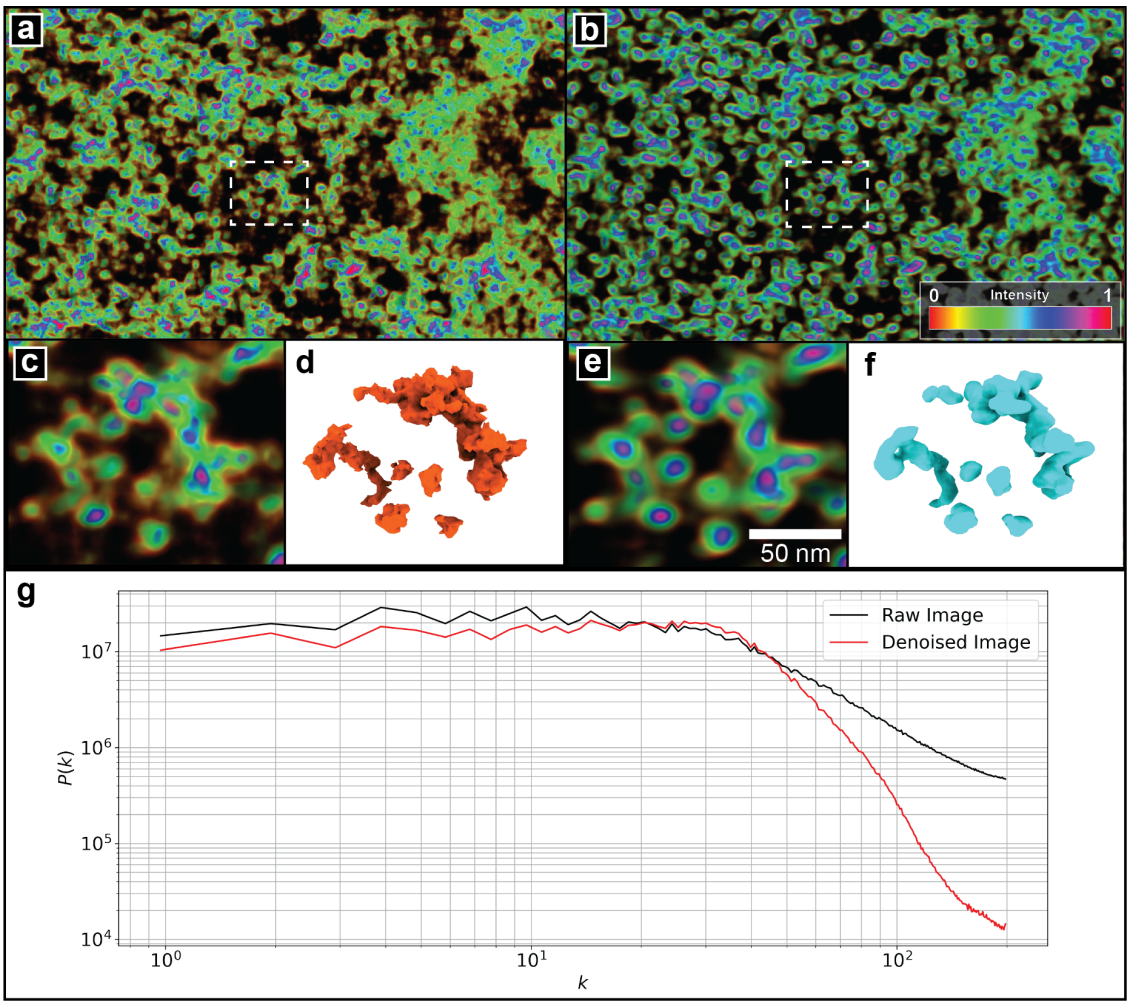

Figure 3.5: Application of the DAE to denoise the experimental tomogram of an imaged A549 cell. The a) original experimental image and b) the image generated after passage through the trained DAE. To improve visual clarity and better highlight features of the images, the pixel intensities are normalized to a [0,1] scale and colored by a pseudo-color gradient indicated by the colorbar as opposed to a single greyscale channel. The denoised image achieves improved resolution of nucleosome-level features within chromatin-rich regions of the experimental image. A subsection comparison between the original c) and denoised experiment e) shows the reduction of noise and results in a smoother 3D reconstruction of the chromatin fiber from the denoised image f) compared to the original d). g) Comparison of the power spectral density (PSD), $P(k)$ between the raw and denoised images shows the denoised image to preserve the large-scale, low-frequency energy density at small wavenumbers k corresponding to the morphological structure of the chromatin fiber, and attenuate the small-scale, high-frequency components at high k that can be primarily attributed to noise.

from the atomic structure and our volume map using the density mapping algorithm from the Chimera software [97]. We find an improved optimal fit with an average high correlation score of 0.87 versus a correlation score of 0.85 for the original tomogram (Figure 4.6). Though comparatively small, incremental quantitative improvements can provide insightful details about chromatin structure. Detecting and quantifying tetranucleosome motifs in raw and denoised images remains an important task and a significant challenge in the field, and expert experimentalists are crucial for interpreting results due to their deep understanding of the biological context and ability to assess image quality and identify relevant features [14, 77]. Our denoising method improved detection of tetranucleosome motifs primarily based on visual cues, resulting in a more accurate representation of chromatin structure in denoised chromSTEM images (Figure 3.7).

These tetranucleosome motifs are known to promote DNA compaction and lead to chromatin condensation, and the preponderance of these structural elements observed within chromatin-dense regions is consistent with prior experiment and simulation [89, 1]. Contrariwise, we do not observe any zig-zag β-rhombus motifs or find any evidence for the formation of the postulated 30-nm fiber [41]. These results support a model in which the *in situ* structural organization of chromatin within chromatin-dense regions in the cell is not a 30-nm fiber, but rather largely composed of smaller tetranucleosome motifs.

3.5.3 Identifying Packing Domains and Their Statistical Properties From Denoised ChromSTEM Stack

The denoised images produced by the DAE enable more robust resolution of chromatin-rich packing domains and improved estimation of statistical distribution of their structural properties such as size, packing scaling exponents, and chromatin volume concentration. We first describe these analyses in the context of the raw ChromSTEM images and then demonstrate how our statistical resolution improves within the denoised images.

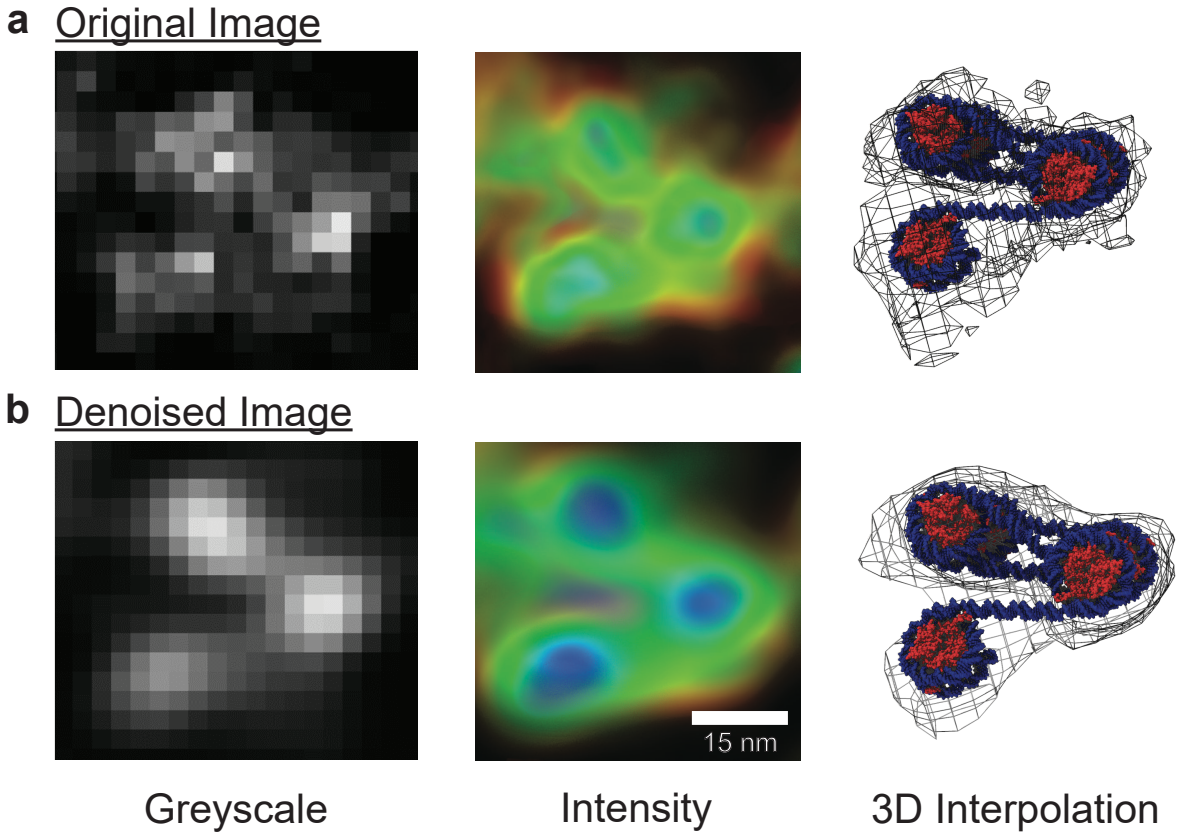

Figure 3.6: Denoised ChromSTEM images reveal tetranucleosomes motifs within a dense chromatin cluster. a) Analysis of nucleosome clusters extracted from chromatin-rich regions of the a) raw experimental tomogram and after passing through the DAE. The denoised image clearly shows the presence of α-tetrahedron motifs that are difficult to discern in the raw image. b) Using Chimera, we construct a prototypical tetranucleosome motif (PDB:1KX5) within the extracted volume of our denoised tomogram and find an optimal fit with an average high correlation score of 0.87 [97]. The construction of the 3D interpolation from the 2D imaging slices is computationally expensive but can, in principle, be extended to large sections of chromatin using high performance computing resources.

Considering first the raw 3D ChromSTEM tomogram presented in Figure 4.5a, we extracted 76 chromatin-rich packing domains and then subjected them to structural analysis to determine the distribution of domain sizes R_f. To do so, we adopted two complementary definitions of domain size. First, we identified the centroid of each domain by creating a local chromatin intensity map by applying Gaussian filtering and local contrast enhancement to the grayscale ChromSTEM z-stacks. We appeal to the fact that ChromSTEM intensity

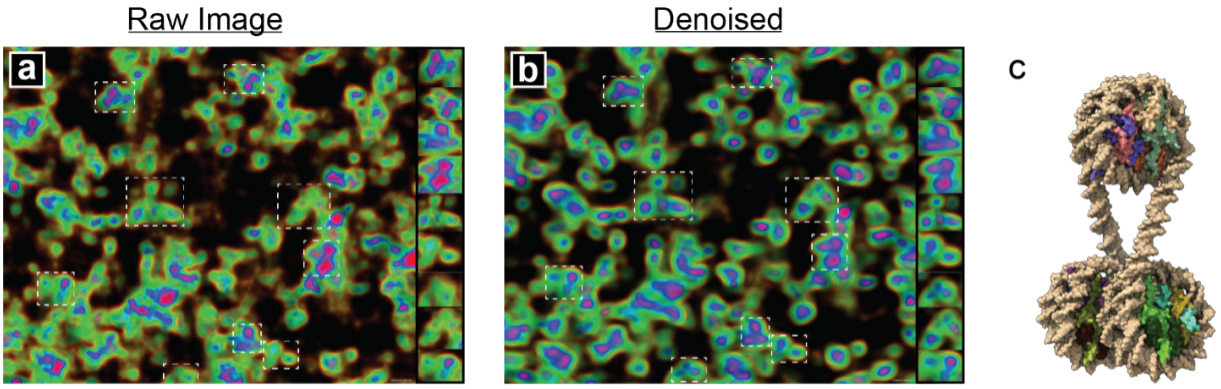

Figure 3.7: Denoised ChromSTEM images reveal tetranucleosomes motifs within dense chromatin clusters. Analysis of nucleosome clusters extracted from chromatin-rich regions within a 200 × 200 nm^2 section of the A) raw experimental tomogram and B) after passing through the DAE. The denoised image clearly shows the presence of C) α-tetrahedron motifs that are difficult to discern in the raw image. We find no evidence for β-rhombus motifs or for the 30-nm fiber.

is approximately linearly proportional to mass to fit a scaling law between mass M and distance r from the centroid of each domain [64]. Following classical power-law polymer scaling relations, mass and distance are expected to be related as $M(r) \propto r^D$, where M is defined as the integrated mass (i.e., intensity) lying at a particular radial distance r from the domain centroid and D is the packing scaling exponent for the polymer that is anticipated to be approximately constant over a particular range of length scales [96]. We computed best-fit values of the packing scaling exponent D by fitting power laws over the range of [0,r] at increasing r and defined the domain size $R_f^{(1)}$ as the distance r at which we observe more than 5% deviation from the best-fit power law. This demarcates the length scale at which a single power-law relationship no longer holds and constitutes our first definition of R_f (Figure 3.9a,b). Second, we calculated the radial density profile of chromatin as a function of distance r from the centroid of the domain. This profile is expected to monotonically decrease until the distance r reaches the boundary of the domain, and then increase again as it begins to encroach upon a neighboring domain (Figure S1c). The minimum in the radial density profile defines our second definition of domain size $R_f^{(2)}$. Finally, we defined the domain

size $R_f = \min\left(R_f^{(1)}, R_f^{(2)}\right)$. We observe that the two complementary definitions of domain size over which we take the minimum are necessary to properly account for the environment in which the domains may be found: in chromatin-poor environments where the domains are isolated, we expect domain-size to be dictated by the mass distribution of the single domain under consideration and $R_f^{(1)} < R_f^{(2)}$; in chromatin-rich environments, we anticipate $R_f^{(2)} < R_f^{(1)}$ and domain size should be more appropriately defined as an multi-body property that defines the boundary between domains.

Having defined R_f and D for each domain, we compute the chromatin volume concentration, CVC, which correlates with binding efficiency of transcriptional reactants and is defined as the fraction of volume occupied by chromatin [93, 64]. The CVC was calculated as the total number of nonzero voxels over the total number of voxels per domain [64]. The distributions of these three quantities for the 76 chromatin-rich domains extracted from the raw A549 3D ChromSTEM tomograms are presented in Figure 3.10 for which we report means and standard deviations of $R_f = (71 \pm 26)$ nm, $D = 2.46 \pm 0.18$, and CVC = $(42 \pm 14)\%$. We previously demonstrated that chromatin forms spatially well-defined higher-order domain structures with radii ranging between an interquartile range of 60 nm to 90 nm in A549 cells, and observe our present measure of mean domain size lies squarely within this range [64].

A concern of applying this structural analysis to the raw ChromSTEM tomograms is the introduction of errors into both the definition of the domains and their structural properties due to the noise inherent in the experimental images. Accordingly, we repeated this analysis for the DAE denoised 3D ChromSTEM tomogram presented in Figure 4.5b. In doing so, our procedure identified 85 chromatin-rich packing domains, nine more than were identified in the raw images. Analysis reveals that application of the domain identification procedure to the denoised image enables identification of more domains and better resolve domains more closely packed in space (Figure 3.11). The improvement in signal to noise ratio in the

denoised tomogram appears to assist in the identification of domain centers that cannot be resolved in the raw tomogram and which are confirmed by manual visual analysis. To assess the possibility of introducing artifacts through the DAE denoising, we present in Figure 3.8 the statistical analysis of R_f, D, and CVC over the 85 denoised ChromSTEM domains. The mean reported values of $R_f = (69 \pm 24)$ nm, $D = 2.65 \pm 0.11$, and CVC $= (60 \pm 13)\%$ are all in good agreement with the analysis of both the raw ChromSTEM images and our prior analyses [64], but are now based on better statistics enabled by the identification of $\sim$12% more domains in the denoised images.

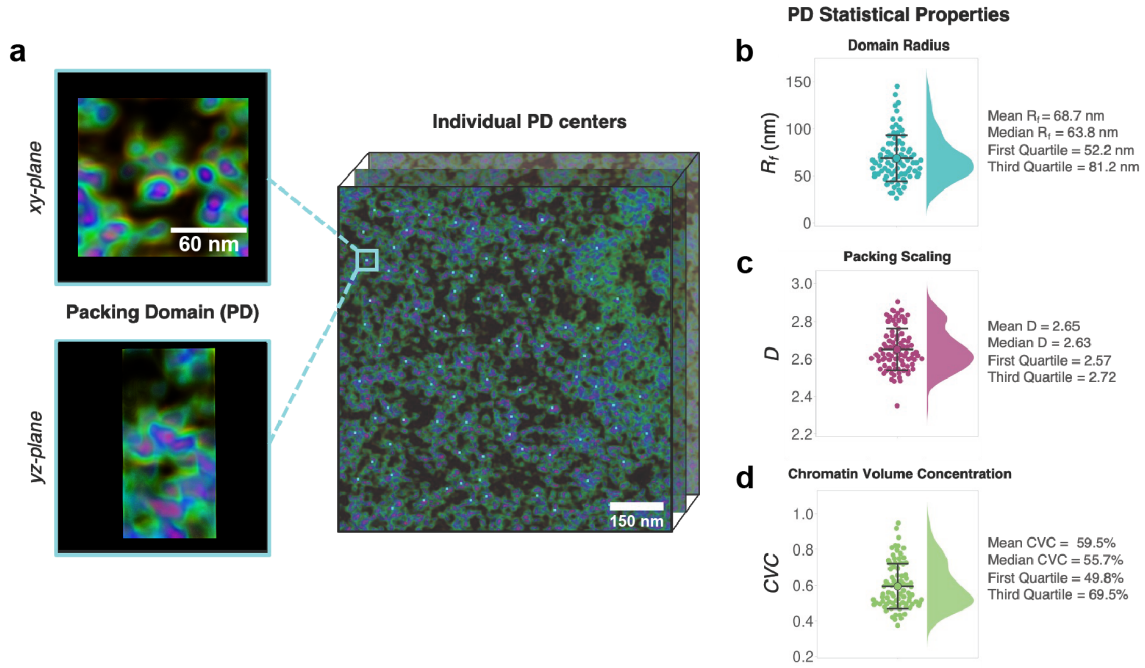

Figure 3.8: Structural analysis of chromatin-rich packing domains from the DAE-denoised A549 3D ChromSTEM tomogram. a) A 3D conformation of a packing domain identified from the denoised ChromSTEM tomogram (Figure 4.5b). Statistical distribution of b) domain size R_f, c) packing scaling exponent D, and d) cluster volume concentration CVC, over the 85 chromatin-rich packing domains identified from the denoised ChromSTEM tomogram. Denoising enables identification of $\sim$12% more domains and domains more closely associated in space relative to analysis of the raw 3D ChromSTEM tomograms.

3.6 Conclusions

By leveraging molecular dynamics and machine learning approaches, we constructed and trained a denoising autoencoder (DAE) capable of removing noise commonly found in scanning transmission electron microscopy tomography with ChromEM staining (ChromSTEM) imaging. The model is trained over physics-based coarse-grained molecular dynamics simulations using the 1CPN model and learns to distinguish the signal from ground truth chromatin structures from artificial noise mimicking the noise profile inherent to experimental STEM imaging. In tests on synthetic ChromSTEM images generated by molecular simulations for which the ground truth is exactly known, the trained outperforms standard denoising approaches, offering a 57% improvement in the mean squared error relative to block-matching and 3D filtering and 72% improvement over non-local means. In applications to *in situ* experimental ChromSTEM images of chromatin within human pulmonary adenocarcinoma epithelial cells (A549 cells), we demonstrate that the DAE eliminates high-frequency noise while preserving the large-scale signal characterizing the chromatin organizational structure. The denoised images enable identification tetranucleosome motifs at a resolution inaccessible within the raw images and expose the α-tetrahedron as the predominant organizational subunit within chromatin-dense regions in the cell and which have been suggested to play a role in chromatin compaction and regulation of gene expression. Notably, we find no evidence for the presence of β-rhombus tetranucleosome motifs or for the 30-nm fiber. The denoised images also permit the identification of $\sim$12% more chromatin-rich packing domains that are obscured by noise within the raw images, enabling improved statistical resolution of the distribution of domain sizes, packing scaling exponents, and chromatin volume concentrations without apparently introducing statistical artifacts. The domain size distributions are consistent with, but have higher statistical resolution and smaller uncertainties than, our prior analyses [64].

The nucleosome motifs exposed by this approach enable new understanding and insight into the small-scale structural organization of chromatin within the cell and how these structures

can influence DNA accessibility and gene regulation. The present work focused primarily on the analysis of tetranucleosome motifs, but in future work we hope to expand our focus to smaller di- and tri-nucleotide motifs. We anticipate that the approaches reported in this study may be applied to ChromSTEM imaging to advance understanding of how stress and epigenetic factors affect chromatin conformation and gene regulation, and may also be applied to other imaging techniques such as cryogenic electron microscopy (cryo-EM). Our study also exemplifies a generic paradigm wherein experimental imaging and theoretical modeling may be bridged via machine learning approaches to enable high-resolution exploration of structural organization within biological systems.

3.7 Acknowledgements

We thank Dr. Yue Li, Eric Roth, and Dr. Reiner Bleher at the Biological-Cryogenic Electron Microscopy (BioCryo) facility at Northwestern University for their assistance in ChromSTEM staining, sectioning, and imaging. We also thank Dr. Tobin Sosnick, Dr. Rebecca Willett, Aria Coraor, Soren Kyhl, Eric Schultz, Yiheng Wu, Mike Jones, Fabian Bylehn, and Kirill Shmilovich for their helpful discussions. This study was supported by NSF Grant EFRI CEE 1830969, NSF grant EFMA-1830961, NIH grants U54CA268084, R01CA228272, R01CA225002, and philanthropic support from Rob and Kristin Goldman. This work was completed in part with resources provided by the University of Chicago Research Computing Center. We gratefully acknowledge computing time on the University of Chicago high-performance GPU-based cyberinfrastructure supported by the National Science Foundation under Grant No. DMR-1828629. This work made use of the BioCryo facility of Northwestern University's NUANCE Center, which has received support from the SHyNE Resource (NSF ECCS-2025633), the IIN, and Northwestern's MRSEC program (NSF DMR-1720139).

3.8 Supporting Figures

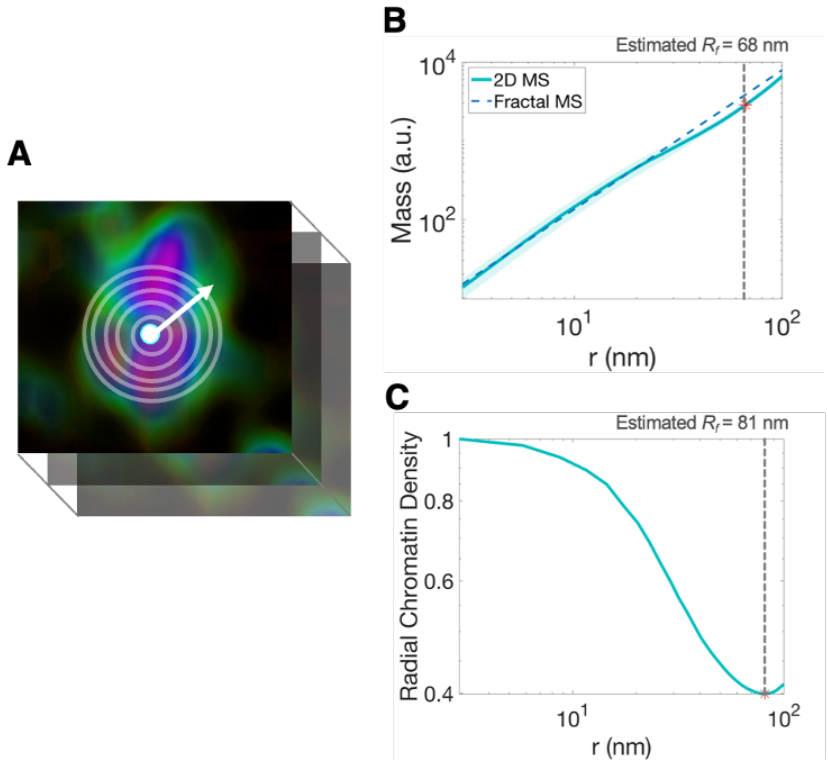

Figure 3.9: Mass scaling and density analysis originating from the domain centers. A) Mass and radial chromatin density are evaluated starting from the center of a domain (white circle with cyan outline) in concentric circles with increasing distance, r. B) Mass scaling of an individual domain in the log-log scale. We performed linear regression on the mass scaling curve and obtained a slope, $D < 3$ for r up to 68 nm (blue dashed line). Beyond the red asterisk, a more significant divergence (>5% error) in the mass scaling behavior is observed. Further, as r increases, there is a sharp transition to the supra-domain regime with D approaching 3. C) Radial chromatin density of an individual domain in the log-log scale. Radial chromatin density of a domain initially is almost constantly high, roughly near the center of the domain. The density then decreases rapidly at moderate distances from the domain center. After a given large distance shown as a red asterisk at 81 nm, radial density increases again. This increase is potentially due to the end of one domain boundary and the interactions with neighboring domains.

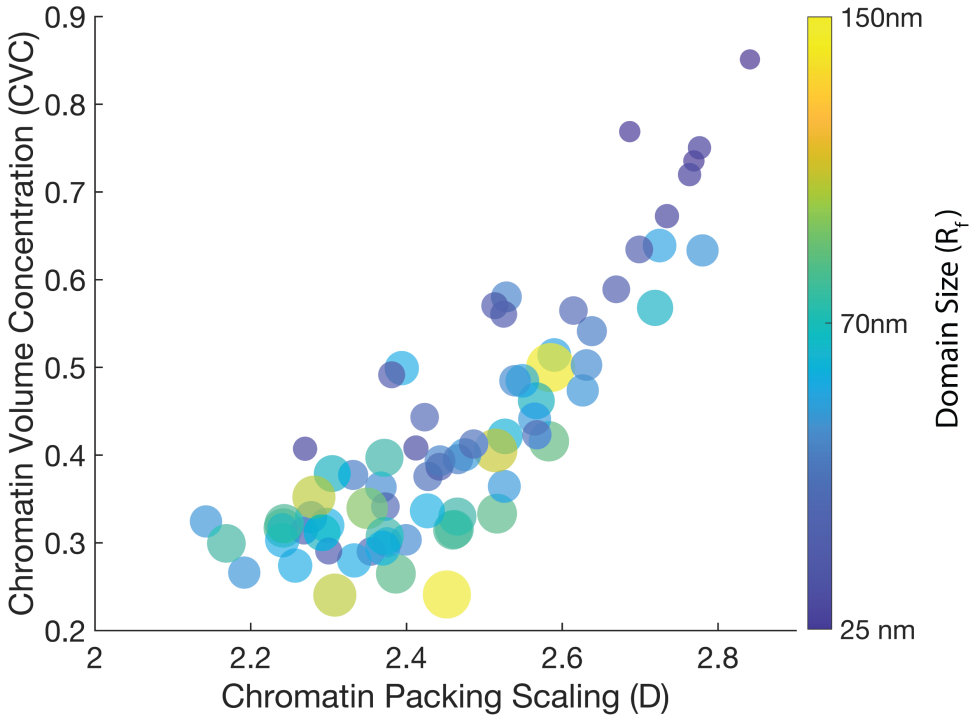

Figure 3.10: Characterization of morphological properties of original higher-noise tomograph of A549 cells. Statistical distribution of chromatin packing scaling D, cluster volume concentration CVC, and domain size R_f.

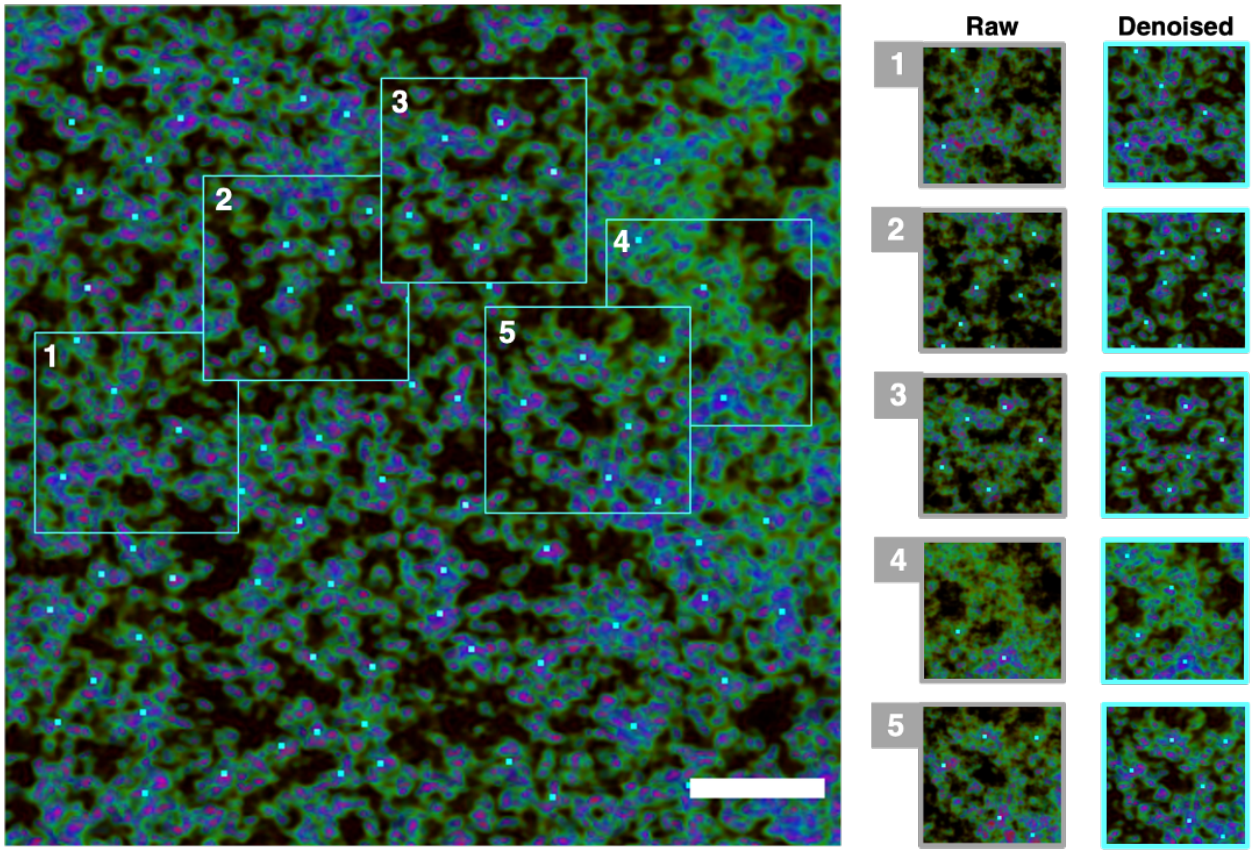

Figure 3.11: Denoising can resolve domains that are closer in space. Left: Domain centers were estimated from denoised tomograms. Right: Five representative regions of the raw and the denoised tomograms show that more domains were identified in the denoised tomogram. Centers are indicated in cyan.

CHAPTER 4

CHARACTERIZING PTM-MODIFIED CHROMATIN FIBERS

4.1 Introduction

Post-translational modifications (PTMs) of histone proteins play an important role in shaping chromatin structure. Despite expectations of uniformity across organizational levels, these modifications sometimes introduce a broad spectrum of effects on chromatin structure, which can appear contradictory [37, 123]. Though extensively researched, the molecular mechanisms underlying the changing dynamics induced by PTMs remains unclear [66]. A mechanistic understanding of how these modifications affect chromatin at the oligonucleosome level is essential in understanding the factors that drive chromatin compaction and decompaction and its effects on regulating gene expression.

PTMs induce structural and dynamic changes to the nucleosome core particle, a complex of a 147 base pair (bp) DNA segment encircling an octamer of histone proteins (H2A, H2B, H3, and H4). Electron micrographs of chromatin segments reveal an array of nucleosomes connected by linker DNA that resembles beads on a string [6]. The process of higher-order chromatin structure folding could potentially be guided by the interactions occurring among nearby nucleosomes [1, 36]. The covalent addition of functional groups can modulate the electrostatic interactions between nucleosomes leading to changes to higher-order structure of chromatin, and consequently, genomic activity.

Of the many types of histone modifications, acetylation has been one of the most extensively studied [114]. Research has shown that acetylation can reduce the compaction of chromatin by decreasing the electrostatic interactions of the K16 residue on the H4 tail with the H2A/H2B acidic patch of neighboring nucleosome [138]. This implies that a reduction in the strength of interactions between nucleosomes should lead to the fiber's elongation as the likelihood of nucleosomes forming compact structures decreases. However, scanning calorimetry studies

have demonstrated an increase in compactness at the local level of individual nucleosomes and groups of nucleosomes from acetylated CHO cells compared to the native state while still inducing relaxation on the overall higher order fiber structure [37]. In addition, of the three endothermic transitions observed in the denaturation profile of chromatin, the peak associated with nucleosome-nucleosome interactions seems to suggest more structurally resilient structures for acetylated chromatin [13, 37]. Conversely, strong interactions between nucleosomes, stimulated by high concentrations of monovalent salts, have been observed to form irregular 3D zig-zag shapes and multiple folded arrangements, signifying an extremely compacted structure [4, 120]. Research also suggests that the elusive 30-nm fiber can be recreated with a high concentration of $MgCl_2$. Interestingly, its structure is composed of tetranucleosome stacking, a configuration resembling a recently identified folding pattern known as the β-rhombus, which is typically associated with open chromatin Figure 4.1a. The dependence of electrostatic interactions on salt concentration highlights the importance of electrostatic interactions in chromatin folding. It also illustrates how chromatin structure disruption at one organizational level can not only be independent but also opposite.

We utilized a combination of molecular dynamics (MD) simulations and machine learning (ML) techniques to investigate how changes to the nucleosome potential affect chromatin folding. It has been observed that long-range nucleosome-nucleosome interactions serve an important role in organizing the chromatin fiber. Experiments have demonstrated that the dynamics of nucleosomes can be well approximated by the Zewdie potential, a variant of the Lennard-Jones potential. By representing chromatin as beads on a string using a 1-nucleosome-per-bead model called 1CPN, we were able to conduct molecular dynamics simulations under previously unattainable lengths and timescales. Considering the vast volume of simulation data generated by our model, we employed diffusion maps, a nonlinear dimensionality reduction technique, to gain insights into the dynamics of the system. This approach assisted us in comprehending the system dynamics and understanding how the

intensity of local interactions between nucleosomes influences the overall structure of the chromatin fiber.

Our key findings align with experimental observations, displaying both localized compaction and somewhat counterintuitive higher-order relaxation. Notably, we observed that when interaction strength decreased, local compaction still occurred, providing a potential explanation for previously reported contrasting results. Through the diffusion mapping of tetranucleosome fibers, we noticed that the motifs maintained equilibrium at the control interaction strength. However, enhanced nucleosome-nucleosome interactions favored a more open beta structure, reflecting the impact of increased salt concentration and the formation of 30-nm fibers. While investigating longer fibers of 16 nucleosomes at varying nucleotide repeat lengths (NRLs), we observed this behavior to decrease as DNA linker lengths increased. Taken together, these results offer a mechanistic perspective on the seemingly contradictory effects observed with these histone modifications, shedding light on their complex interplay.

4.2 Methods

Our research was carried out using the 1CPN model, originally developed by the de Pablo group, to simulate chromatin behavior. In this model, chromatin is depicted as a string of beads, with each bead symbolizing a single nucleosome. Nucleosome interactions were represented by the Zewdie potential, which has been validated for accurately portraying internucleosomal interactions. To account for the stabilizing effects of the H3 histone tail as DNA enters and exits the nucleosome, a pairwise interaction between the dyad and DNA sites was introduced. We incorporated electrostatic interactions using the Debye–Huckel theory. Simulations were conducted using LAMMPS molecular dynamics software. We initiated the chromatin filaments in an extended state by setting the angle between incoming and outgoing DNA and the nucleosome to 180 degrees. Every simulation was executed until it reached a simulation time of 120 microseconds, guaranteeing ample sampling of the system's dynamics.

For our analysis, we utilized the latter half of the simulation trajectories of five replicate runs to increase statistical robustness. This process resulted in a dataset comprising 20,000 snapshots equating to 60 microseconds of simulation time per replica, spanning an aggregate simulation time of 300 microseconds for each interaction strength. Simulations of extended fibers consisting of 16 nucleosomes, with nucleosome repeat lengths of 157, 187, and 197, adhered to the same procedure.

Post-translational modifications (PTMs) can exert effects through several mechanisms, among which changes in electrostatic potential are particularly notable. To probe the implications of PTM-induced changes in the strength of nucleosome-nucleosome interactions, we manipulated the interaction potential's intensity within the 1CPN model by adjusting the well depth, e_0, of the Zewdie potential. We altered the well depth within a range of 50% reduction to 50% increase relative to the control group, thereby simulating an array of conditions possibly arising from PTM-induced variations in interaction strength. Earlier work on nucleosome dynamics parameterized the interaction strength prompted by acetylation of the H4 tail, an effect akin to the removal of H4 tails, was also included resulting in a total of 11 conditions (Figure 4.1b). To investigate how these local alterations influence longer chromatin strands, we executed five replicate simulations for systems with diverse nucleosome repeat lengths (NRLs), including 157, 187, and 197. We then compare the results for various conditions, including H4, control, and a strong interaction (50% increase), to gain insights into how local interactions impact the overarching chromatin structure.

The heat capacity for our tetranucleosome simulations were derived to assess structural stability of our models. Heat capacity is a measure of the ability of a system to absorb heat, and it is commonly used to characterize the stability of protein structures. By expressing the heat capacity in terms of fluctuations of the energy of the system,

$$C_v = k\beta^2 \left[\left\langle E^2 \right\rangle - \langle E \rangle^2 \right] \tag{4.1}$$

we obtained the heat capacity and evaluated the stability for each condition.

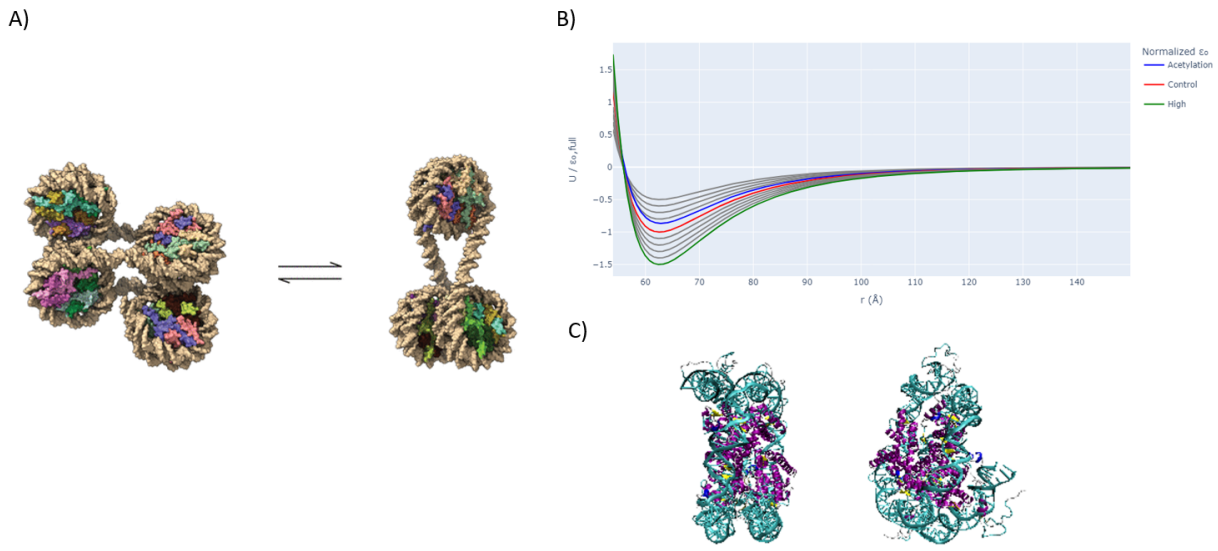

Figure 4.1: PTM induced modifications are modeled as modification to the well depth, e_0, of the Zewdie Potential. A) Two previous motifs, α-tetrahedron and β-rhombus have been suggested to induce chromatin fiber elongation or compaction. B) Adjusts to the well depth, e_0, are assumed to model possible modifications to the nucleosome interaction strength when two nucleosomes are in a stacked configuration C).

We employed Rouse mode analysis, a technique used to study the dynamics of systems composed of beads connected by springs, to study the collective motion of our systems (modes). Each mode is defined as cosine transformation of the position vector r_i at time t,

$$\mathbf{X}_p(t) = \left(\frac{2}{N}\right)^{1/2} \sum_{i=1}^{N} \mathbf{r}_i(t) \cos\left[\frac{p\pi}{N}(i-1/2)\right], \quad p = 0, \ldots, N-1. \tag{4.2}$$

where p > 0 represent the internal relaxation of a chain composed of N/p monomers. Each mode represents a different characteristic motion of the system, with the first mode typically corresponding to the oscillation from end-to-end of the whole fiber, followed by half segments of the fiber (mode 2), and finally neighboring nucleosome interactions (mode 3). By analyzing these modes, we gained further insights into the dynamics and conformational changes of

chromatin.

Diffusion maps, a nonlinear manifold learning technique, have proven invaluable for generating lower-dimensional representations of high-dimensional molecular trajectories, a property we utilized in our study. These maps postulate that a chosen metric for comparing pairs of configuration microstates accurately represents the short-term kinetic distance and that the dynamics over the conformational space can be approximated as a diffusion process. Thus, the leading collective variables of the diffusion map align with the system's large-scale, high-variance collective motions. In addition, configuration microstates that are kinetically proximate are embedded closely together. In our study, we utilized the density-adaptive version of diffusion maps, which we found exceptionally suitable for managing the significant variations in sampling densities observed in our chromatin simulations. This allowed us to identify similar structures based on the shortest diffusion path in the data space, offering valuable insights into chromatin's dynamics and structural changes. We subsequently generated 2D histograms of these maps to pinpoint relative low-energy states. For a detailed description of the methods employed in this study we refer the reader to the "Material and Methods" section of Chapter 2.

4.3 Results and Discussion

Through the utilization of these methods, we aimed to uncover the influence of local nucleosome-nucleosome interactions on the overall structure and dynamics of chromatin fibers. By examining the effects of varying interaction strengths and fiber lengths, we aimed to elucidate the mechanisms underpinning the contradictory behaviors observed at various chromatin organizational levels. Recent research into the local organization of chromatin has identified tetranucleosomes as the building blocks of higher-order structure and suggested that they serve a crucial role in chromatin folding. Recent studies have proposed that tetranucleosomes can take two particular motifs, the α-tetrahedron and β-rhombus, which induce

compaction or decompaction of the chromatin fiber, respectively (Figure 4.1A). The alpha motif has been suggested to be responsible for introducing kinks along the fiber resulting in denser chromatin domains. While the *beta*-rhombus seems to be the structural motif responsible for the elongating of the chromatin fiber, allowing for a more open conformation. Studying how PTMs affect the formation of these motifs may offer clues into their cascading effects on higher-order structure.

4.3.1 Lower interaction strength induces compaction dynamics in tetranucleosome building blocks.

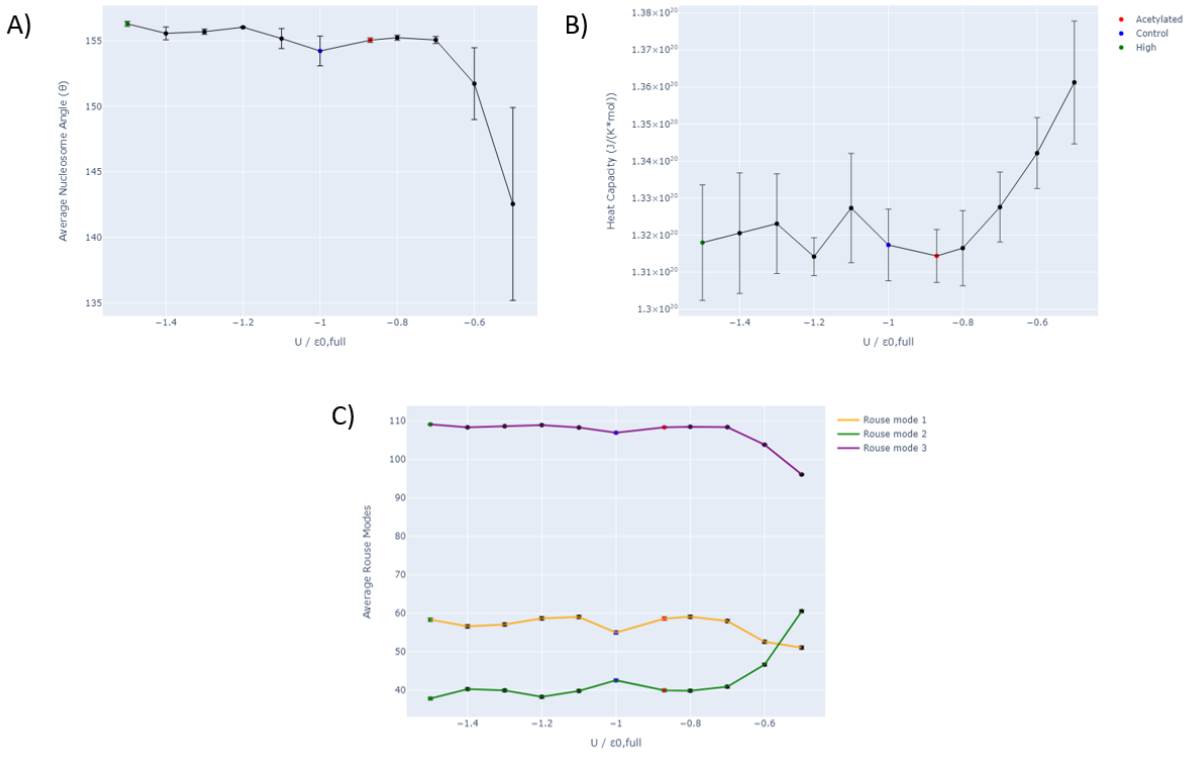

Figure 4.2: Qualitative analysis of tetranucleosomes reveals compaction as intranucleosomal interaction decreases. A) Angle nucleosome angle was calculated to measure folding and compaction. B) Heat capacity of tetranucleosomes models suggest increase structural stability as interaction strength decreases. C) Rouse mode analysis shows interactions with neighboring nucleosomes decreases and interchromatin interaction increases.

Our study focused on the influence of nucleosome-nucleosome interaction strength on the dynamics of tetranucleosome building blocks. We considered 11 cases with nucleosome-nucleosome interaction strengths ranging from 0.5 to 1.5 times that of the control group. We assessed the average nucleosome angle (Figure 4.2A) to determine the degree of folding and observed that as the interaction strength decreased, the average nucleosome angle decreased, indicating increased compaction. Interestingly, we identified a critical interaction strength of around 0.7, where a significant increase in compaction occurred. We found that an increase in the heat capacity at this same inflection point indicating enhanced stability (Figure 4.2B).

Our findings reveal that as the nucleosome-nucleosome interaction strength decreases, the compaction of the chromatin fiber decreases. This behavior aligns with the paradoxical results observed in experiments, where histone acetylation, known to decrease interaction strength, leads to local compaction despite overall chromatin relaxation [37]. Notably, H4K16 has been found to favor the alpha motif, suggesting a preference for a more compact chromatin structure [89].

To further investigate the dynamics of the system, we employed Rouse mode analysis to examine the importance of different motions in the system (Figure 4.2C). The average values of each Rouse mode for the fibers were calculated. Mode 1, representing the whole chain, showed a slight decrease, while Mode 2, associated with interchromatin interactions, displayed an increase. Mode 3, reflecting neighboring nucleosome interactions, decreased as the interaction strength decreased, again at this same inflection point. Our rouse mode analysis indicates a shift in dominance from nucleosome-nucleosome interactions to intrachromatin interactions as the interaction strength decreases. This shift suggests that intrachromatin interactions, driven by promiscuous nucleosomal binding, become increasingly important in mediating chromatin compaction.

The analysis of diffusion maps (Figure 4.3) provides further insights into the structural changes induced by varying interaction strengths. Given that their structural composition

were similar, four nucleosomes of 187 NRL, we applied a single diffusion map for all 11 cases. The diffusion maps revealed two distinct regions corresponding to the alpha and beta motifs (Figure 4.3A). The alpha structure is characterized by a nucleosome having a higher propensity to disassociate. We colored our maps by identifying structures that had at least a single nucleosome as an outlier. We found this to be effective in differentiating alpha and beta motifs. We then color this map by interaction strength and find that the alpha region was comprised of the control, H4, and systems with interaction strengths below 0.7 (Figure 4.3B-D). The observed deviation from the dominant beta structure, which is prevalent when nucleosome-nucleosome interactions are strong, supports the notion that lower interaction strengths disrupt the formation of the beta structure, leading to more open chromatin conformations. The observation that high nucleosome-nucleosome interaction strengths favor the formation of beta structures (Figure 4.3E) also aligns with previous studies showing that increased salt concentration, which strengthens these interactions, promotes the formation of a 30-nm fiber characterized by beta stacking.

Interestingly, our results reveal the existence of alpha structures in the control group, suggesting that there might be a phase transition point where the preference for alpha or beta structures is determined. It seems that for the control model, slight modifications to the potential can shift the equilibrium between these two structural states and thereby affect chromatin compaction. This phase transition point may be finely tuned by the nucleosome-nucleosome interactions and could be influenced by post-translational modifications that perturb these interactions, leading to a shift in preference towards either more alpha or more beta folding motifs.

4.3.2 Longer NRLs seems to mediate expansion

Moving to longer nucleosome repeat lengths (NRLs), our diffusion map and radius of gyration studies allow us to investigate how the disruption of these tetranucleosome building

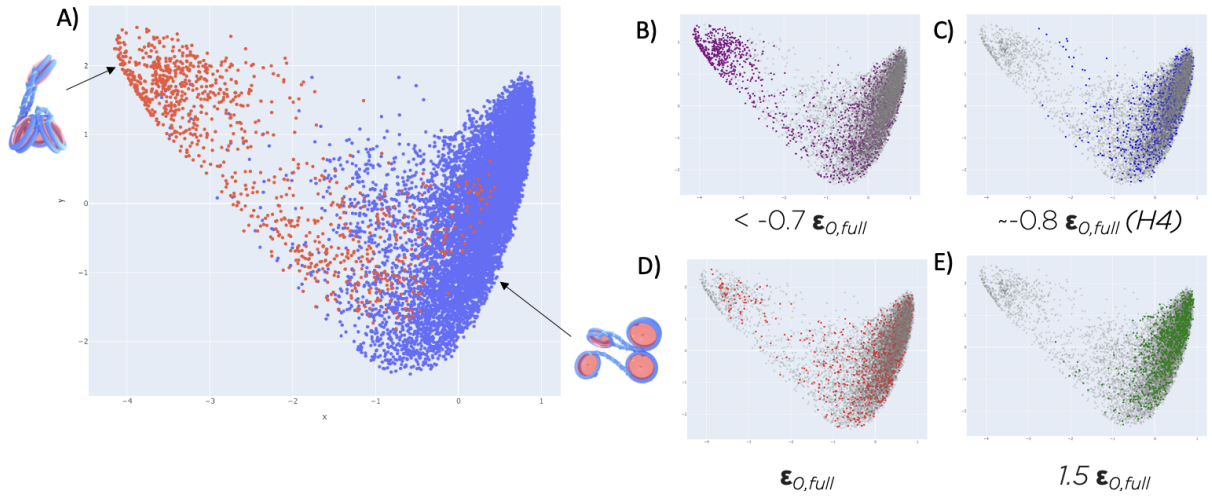

Figure 4.3: Diffusion maps of all conditions reveals two distinct folding motifs accessible by decreasing nucleosome interaction strength. A) A diffusion map colored by identifying motifs with at least a single nucleosome at a distance a deviation away from the mean. Alpha motifs are colored in red and beta motifs are colored in blue. B) Decreasing interaction strength below 0.7 allows access to alpha motif region. C and D) H4 and control nucleosomes can form alpha motifs. E) High interaction strength locks tetranucleosomes in beta state.

blocks affects chromatin at larger length scales. At shorter NRL scales, where neighboring nucleosomes are closer together, the interactions between them still play a significant role. As the NRL increases, we start to observe behavior that is consistent with studies of chromatin fiber opening up upon acetylation, indicating a more open and extended chromatin conformation.

We investigated impact of these modifications to the interactions strength on the longer chromatin fibers with increasing nucleosomes and longer nucleosome repeat lengths. Figure 4.4 compares the average radius of gyration for four, eight, and 16 nucleosomes and indicates significant differences, particularly for the 16-nucleosome fibers. In the case of NRL 157, the high interaction potential system ("High") exhibited higher compaction compared to both H4 and the control group. However, H4, with lower interaction strength, also showed more compaction compared to the control group. Similar trends were observed for the fibers of NRL 187. At NRL 197, the system started to behave as expected, with decreasing radius of

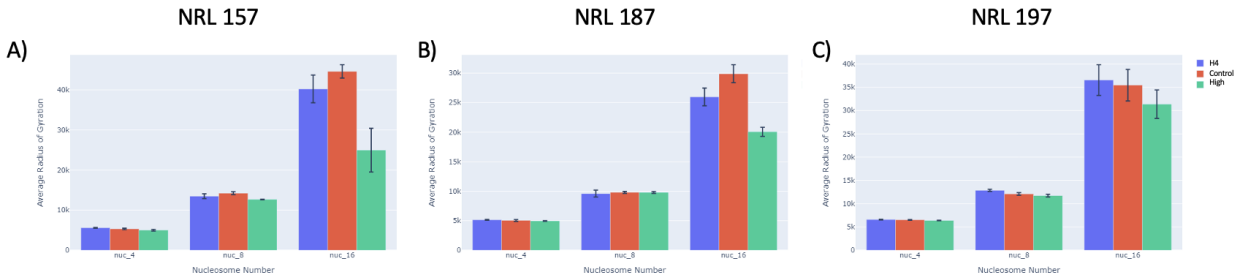

Figure 4.4: An increase in nucleosome repeat length allows chromatin to elongate at low nucleosome interaction strength. The average radius of gyration for all conditions were calculated for fibers of four, eight, and 16 nucleosomes at NRLs of A) 157, B) 187, and C) 197.

gyration associated with increasing interaction strength.

Analyzing the diffusion maps for the 157 case (Figure 4.5) revealed the formation of structures resembling the 30-nm fiber for all three possible conditions and globular systems induced by kinks alpha motifs predominately found in the H4 and "High" models. The fact that all three systems form a elongated fiber, suggests this to be a the preferred state, and the stability of maintaining it would be depended on inter-/intrachromatin interactions that may disrupt it. Higher nucleosome-nucleosome interaction strength allowed for the formation of two lower state compacted conformations (Figure 4.5B). The H4 group exhibited a propensity for beta constructed motif, with an energetic barrier reduction to two of the stable compacted forms occupied by the "High" model (Figure 4.5D). The lower energetic barrier allows for these states two compacted states to be more accessible allowing for the formation of alpha motifs and the observed decrease in radius of gyration compared to the control model. Similar behavior was observed for the 187 case.

Figure 4.6) displays the diffusion embedding of the NRL 197 case, the fiber displayed characteristics of a liquid polymer, comprising motifs associated with both compaction and elongation. A single global minimum is observed in the "High" model corresponding to a semi-compacted state. In the control case, the fiber begins to explore the energy landscape

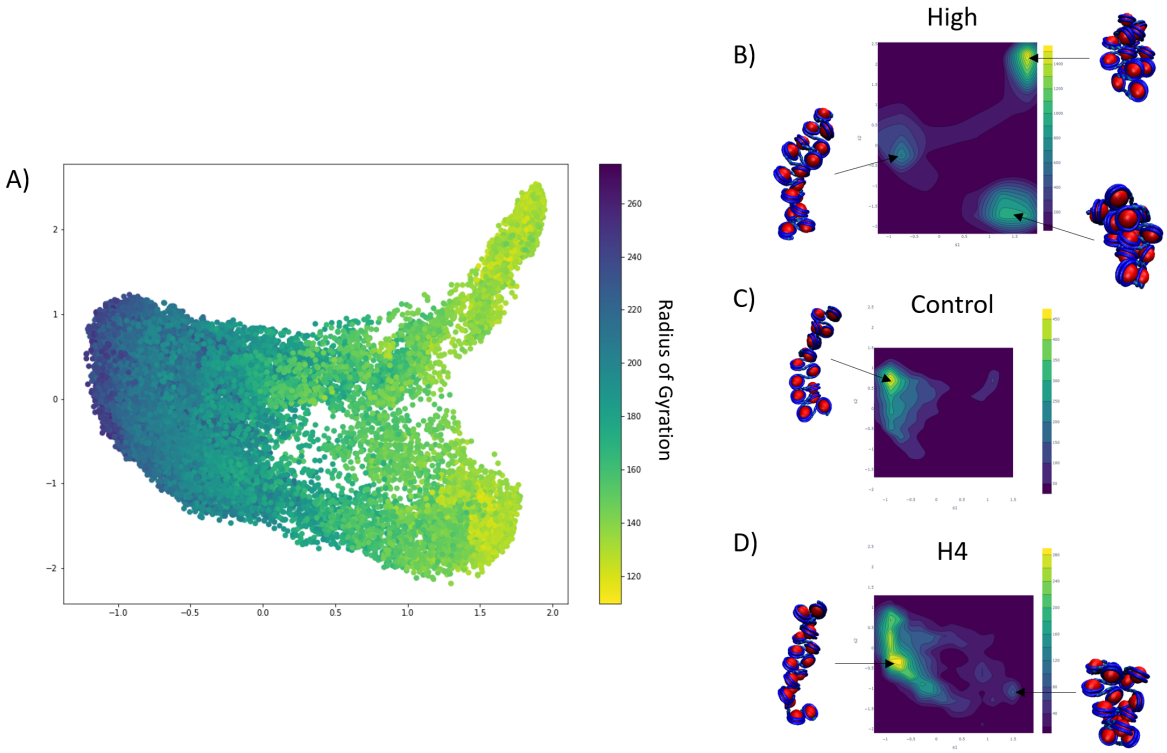

Figure 4.5: Diffusion maps of the 157 NRL system shows formation of systems that resemble a 30 nm fiber and globular system induced by kinks in the fiber from the formation of alpha structure. A) Diffusion map embedding with all conditions considered and colored by radius of gyration display compaction and elongation of the fiber. B) At high interaction strength, compacted structures form an are energetically stable. C) For the control model, beta motifs seem to be the preferred state. D) For acetylated models, both compacted and decompacted conformations are accessible.

allowing access to higher and lower compacted structures. At the H4 interaction strength, four local minima observed corresponding to elongated and compacted states.

These findings begin to unravel the independent nature of chromatin disruptions at diverse scales. Despite the existence of local compaction, the potency of interactions both internally and among nucleosomes weakens, making them more prone to disturbances by intra- and interchromatin interactions. This behavior corresponds to the experimentally observed flexibility and separation of nucleosomes in crowded chromatin spaces.

Additionally, these insights demonstrate that diminished nucleosome-nucleosome interaction strength fosters compaction dynamics in foundational tetranucleosome structures, which

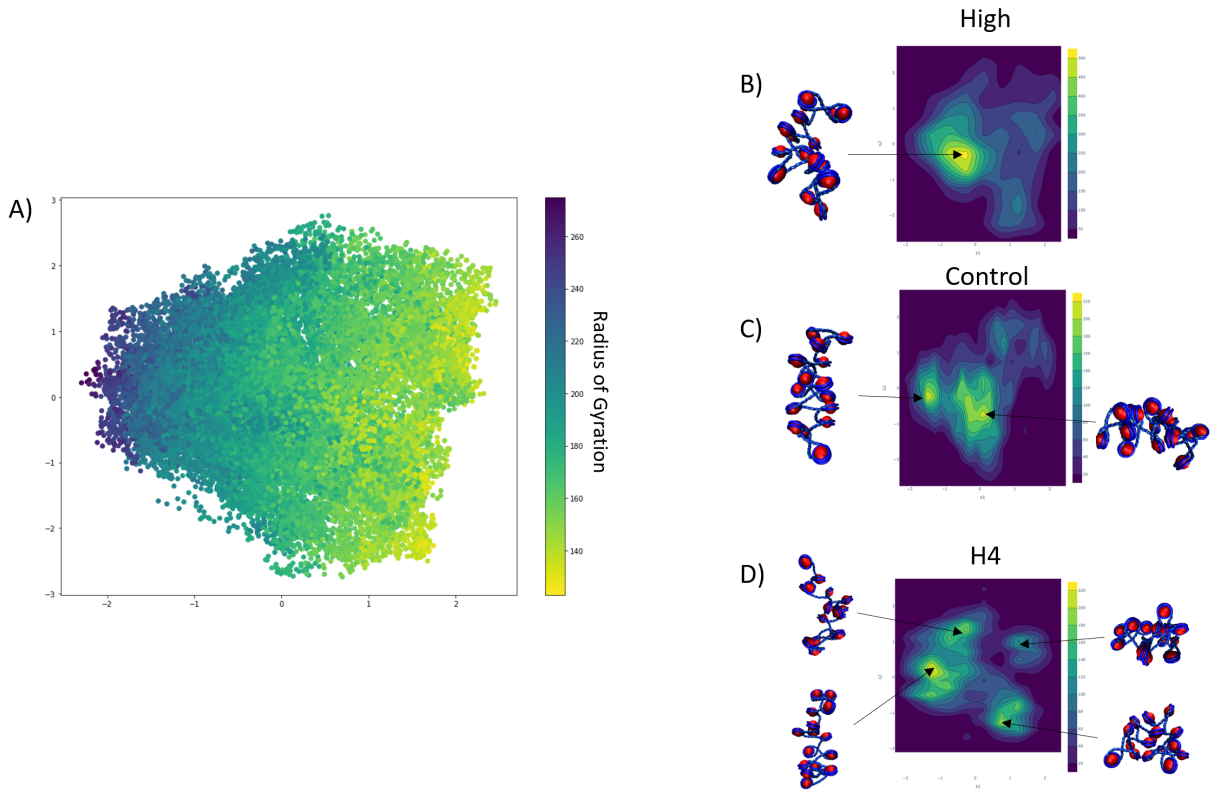

Figure 4.6: Chromatin fiber begins to behave like a liquid polymer. A) First diffusion mode (colored by radius of gyration) of the diffusion map embedding captures compaction and elongation of the fiber. B) At high interaction strength, a single global minimum is observed. As the interaction strength is decreases C,D), the fiber can explore a wider free energy landscape.

ripple up to influence higher-order fiber architectures. We've found that increasing NRL corresponds with a more relaxed fiber, underlining its significance in facilitating higher-order folding.

Collectively, our findings suggest that the magnitude of nucleosome-nucleosome interaction strength is pivotal in shaping chromatin compaction. The equilibrium between alpha and beta structures, manipulated by interaction strengths and PTMs, could function as a control mechanism for organizing chromatin across varying length scales. These insights bolster our understanding of the intricate relationship between local and global chromatin structure, emphasizing the need to consider nucleosome-nucleosome interactions and NRLs in the study of chromatin folding and dynamics.

4.4 Conclusion

In this study, we employed molecular dynamics (MD) simulations and machine learning (ML) techniques to investigate the effects of nucleosome-nucleosome interaction strength on the structure of the chromatin fiber. By utilizing diffusion maps of tetranucleosomes, we observed the previously observed paradoxical behavior of compaction at smaller length scales when the interaction strength decreases. Specifically, we found that tetranucleosomes have a higher propensity to form alpha structures as the interaction strength decreases, while a higher interaction strength promotes the formation of beta structures. The control group appeared to be a transition point between these two structures, aligning with experimental findings.

Furthermore, by studying longer chromatin fibers, we discovered that nucleosome repeat length (NRL) plays a crucial role in the formation of chromatin structures. While local effects on compaction were observed, these effects did not significantly disrupt the overall structure of larger fibers. As NRL increased, intrachromatin interactions emerged as important factors, and the system exhibited more fluid-like behavior. These findings suggest that intrachromatin interactions become increasingly relevant at higher-order length scales, supporting the notion that NRL influences chromatin folding.

The importance of our work lies in elucidating the mechanisms underlying experimental observations of chromatin ordering at different length scales. Through the integration of MD simulations and ML techniques, we provided insights into the complex interplay between nucleosome-nucleosome interactions, NRL, and chromatin structure. This work brings us closer to understanding the PTM code and its relationship to chromatin organization and, consequently, gene activity.

Moving forward, our ongoing research focuses on determining kinetic rates and exploring how these modifications affect chromatin dynamics. By unraveling the intricacies of the PTM code and its impact on the chromatin fiber, we aim to deepen our understanding of gene regulation and the mechanisms that govern chromatin structure and function.

CHAPTER 5

MOLECULAR CHARACTERIZATION OF COVID-19 THERAPEUTICS: LUTEOLIN AS AN ALLOSTERIC MODULATOR OF THE SPIKE PROTEIN OF SARS-COV-2

Reprinted with permission from Alvarado, W.; Perez-Lemus, G. R.; Menéndez, C. A.; Byléhn, F.; de Pablo, J. J. Molecular characterization of COVID-19 therapeutics: luteolin as an allosteric modulator of the spike protein of SARS-CoV-2. Molecular Systems Design & Engineering, 2022, 7, 1, 58-66.

5.1 Author contributions

All authors contributed equally to this work.

5.2 Abstract

The interactions between the receptor binding domain (RBD) of SARS-CoV-2 and the angiotensin-converting enzyme 2 (ACE2) are crucial for viral entry and subsequent replication. Given the large and featureless contact surfaces between both proteins, finding a suitable small-molecule drug that can bind and disrupt regulatory pathways has remained a challenge. A promising therapeutic alternative has been the use of small compounds that bind at the protein-protein interface or at distal "hot spots" and induce conformational changes that inhibit or stabilize protein-protein interactions (PPIs). In this work, we conduct large-scale all-atom explicit solvent simulations of the top scoring compounds from a recent large-scale high-throughput docking screening to investigate their interaction with the RBD domain of the spike (S) protein in complex with ACE2. We identify several promising candidates that exhibit a large negative free energy of binding to the RBD/ACE2 complex based on

ab initio thermodynamic integration calculations. A systematic structural analysis of two glycosylation profiles of the RBD/ACE2 complex reveal the important role glycans play in modulating the binding of small-molecules. Cross correlation, fluctuation and strain analysis identify several of these compounds as effective PPI modulators that inhibit or stabilize the protein-protein interactions of RBD/ACE2 based on glycosylation profile. Among them, Luteolin, a drug currently approved for asthma, exhibits an intense allosteric effect when it binds to a previously unidentified distal binding site of the RBD domain far from the RBD and ACE2 interface which may serve as a potential target for future drug discovery.

5.3 Design, System, Application

The design of small-molecules which target the spike protein of SARS-CoV-2 has been the focus of recent studies in the search for an effective treatment strategy against COVID-19. In this study, molecular dynamics simulations are used to study how several previously identified therapeutics alter the structure and dynamics of the spike protein. We analyze the strain and molecular stiffness induced by each small molecule upon binding and identify a previously uncharacterized distal binding site that induces significant allosteric conformational changes to the protein–protein interaction that takes place between the subunits of the spike protein and the active site of the angiotensin-converting enzyme 2 (ACE2), a key binding partner required for host cell infectivity. The fundamental understanding obtained in this study will enable the development and engineering of novel small-molecules that disrupt viral entry and provide an informed perspective on how allosteric modulators serve as a template for the design of suitable therapeutics.

5.4 Introduction

As of February 2021, the human coronavirus referred to as severe acute respiratory syndrome coronavirus 2 (SARS-CoV2) has infected millions around the world, and is responsible for the COVID-19 disease that has so far led to well over 2.61 million deaths worldwide [24]. As vaccinations begin to roll out, viable treatments for COVID-19 are still necessary to treat those who may contract the disease. One promising treatment strategy is the identification of previously approved drugs that could be repurposed to act on different stages of SARS-CoV-2 infection and host response.

The surface-anchored viral spike (S) glycoprotein mediates coronavirus entry into host cells. The S protein is a type I transmembrane protein comprising two functionally distinct regions, S1 and S2, that mediate receptor binding and membrane fusion, respectively. Similar to SARS-CoV-1, responsible for the 2003 SARS outbreak in Asia, the S1 subunit of the SARS-CoV-2 engages the same angiotensin-converting enzyme 2 (ACE2) receptor for host cell entry and utilizes the serine protease TMPRSS2 for S protein priming [47]. Of the two regions, the S2 subunit is the most conserved among different coronavirus genera and nine residues involved in the interaction between ACE2 and the receptor binding domain (RBD) of the S1 subunit are conserved between SARS-CoV and SARS-CoV2 [140, 50].

Specific RBD-receptor binding determines host cell infection and has been the target of several drug design efforts. Recent flow cytometry studies have shown that SARS-CoV-2 RBD exhibits a higher binding affinity to human ACE2 (hACE2) and bat ACE2 (bACE2) receptors than SARS-CoV-1 RBD.[121] Deletions in the RBD of closely related S protein sequences have also been shown to inhibit binding to hACE2.[131] In addition, several studies indicate that the RBD is an attractive antigen for specific antibody detection.[101, 90] Taken together, these results suggest that SARS-CoV-2 RBD is a promising target for the development of novel vaccines and antiviral drugs.

Recent studies have identified key surface contacts that can be leveraged to disrupt the

RBD/ACE2 protein-protein interaction (PPI) [136, 137]. PPIs are physical contacts of high specificity between proteins that play an important role in cellular function [134]. PPIs are characterized by their large, flat, and featureless binding interfaces, as in the case of the RBD/ACE2 complex. One major challenge in PPI inhibition is designing small molecules that can compete with the high binding affinity of a natural protein partner. In addition, the large and flat nature of the protein interaction surface often lacks clear binding pockets or grooves that can act as binding sites for smaller ligands. A promising alternative is the identification or engineering of allosteric modulators that stabilize protein-protein interactions, thereby interfering with the downstream pathways they mediate [85]. The key strategy in this approach involves indirectly affecting PPI interfaces by targeting distal binding sites that are structurally distinct. Allosteric modulators induce a conformational change that either inhibits or stabilizes association with another protein and have shown promise in the design of suitable therapeutics [38, 92].

In this work, we focus on a set of small-molecule drugs that can be repurposed for the therapeutic treatment of COVID-19. Specifically, we focus on the top ten scoring compounds proposed in a recent high-throughput supercomputer-based docking screen that was performed in vacuum and relied on minimization of the potential energy of the ACE/RBD complex [115]. Here we perform molecular simulations in explicit water and determine the relative and absolute binding free energies of each drug at various binding locations and with different glycosylation profiles. We find the affinity of the ligands to the complex to be different than previously predicted on the basis of energy minimization, with several of the top candidate drugs unbinding from the protein over the course of nanosecond long trajectories. We study the effect glycans have on binding affinity for several ligands by comparing two glycoforms of the RBD/ACE2 complex: a simple and "Abundant" model. Notably, our free energy results show that for an RBD/ACE2 complex glycosylated with small sugar moieties, high-affinity binding ligands stabilize the binding between RBD and ACE2. An RBD/ACE complex with

glycans that were determined as the most abundant from a glycoproteomics study conducted by Zhao *et al.* was analyzed for comparison [139]. The presence of these complex-glycans affect the binding affinity of several drug candidates and their mechanism of action. We then analyze the strain and molecular stiffness induced by each small molecule upon binding and identify a "hotspot", comprising only a few amino acid residues, as a potential target for drug discovery. Surprisingly, binding of Luteolin – one of the drugs considered here, to this distal site, induces a profound allosteric conformational change to residues near the RBD/ACE2 binding interface, which disrupts non-native contacts at the protein-protein interface. We also provide additional quantitative insights into the binding mechanism of Luteolin to the RBD/ACE2 complex using cross-correlation analysis and Principal Component Analysis (PCA). We conclude with a discussion of our findings and several suggestions for future experimental studies.

5.5 Methods

5.5.1 Molecular Docking

Autodock 4.2 was used for the molecular docking between target proteins and ligands using the Lamarckian genetic algorithm (LGA), and pseudo-Solis and Wets local search method [84]. The initial configuration of receptor-binding domain (RBD) of the spike protein of SARS-CoV-2 bound to the ACE2 receptor was taken from crystal structure (PDB: 6M0J) after an NPT relaxation for 100ns and averaged for the last 20ns (see section below for simulation details). The search space was centered around the S-protein and spanning the binding interface of the ACE2 receptor. The initial configurations of ligands were randomized before each docking calculation. A total of 200 docking runs were performed and each run was set to terminate after 25,000,000 energy calculations. The best pose of each ligand was selected for further analysis in MD simulations.

5.5.2 Molecular Dynamics

The recently determined crystal structure of the receptor-binding domain (RBD) of the spike protein of SARS-CoV-2 bound to the ACE2 receptor was used as an initial structure (PDB: 6M0J). Crystal waters were removed force field parameters of the S-protein/ACE2 complex were determined with the Antechamber program using the ff14SB and GLYCAM-06j-1 force fields. An octahedron box with 40,000 TIP3P water molecules and 23 Na+ ions was added. Energy minimization included 3000 steps which involved 1500 steepest descent steps constrained to heavy atoms, followed by a second minimization of 30000 steps involving 15000 steepest descent steps. Equilibration was performed through a gradual temperature increase from 0 K to 300 K over 400 ps using Langevin thermostat with a temperature coupling constant of 1.0 ps in a constant volume ensemble. Density equilibration and production runs were conducted using a constant pressure ensemble (NPT). All simulations were performed using periodic boundary conditions and 2 fs time step. Long-range electrostatic interactions were modeled using the Particle Mesh Ewald method with a non-bonded cut-off of 10 Å and the SHAKE algorithm. The ligands included for MD simulations were described by the General Amber Force Field (GAFF). Partial charges for all small-molecules were generated using the AM1-BCC charge model. Their initial position was selected from the best score in docking calculations.

5.5.3 MM/GBSA Calculations

The relative binding free energy between ACE2 and RBD domains were calculated using the GBSA method implemented in MMPBSA.py with igb2=2 and mbondi2 parameter [80]. This relative free energy is defined as: $\Delta\Delta G_{Binding}=\Delta G^{Lig}-\Delta G^{N}$ where the superscript Lig denotes the RBD/ACE2 binding free energy with the presence of the ligand in simulations and N denotes the RBD/ACE2 system with no ligands. Three different replicas were used for every ligand/no ligand system, each one with 5000 snapshots sampled every 20ps from a

previously 300ns equilibrated system.

5.5.4 Thermodynamic Integration Calculations

The absolute binding free energy for ligands is defined as: $\Delta G_{Absolute}=\Delta G^{L}-\Delta G^{RL}$, where ΔG^{RL} is the free energy change of the ligand annihilation in the RBD/ACE2 complex, and ΔG^{L} is the free energy change of ligand annihilation in water. To calculate these free energy changes, we use Thermodynamic Integration (TI) implemented in PMEMD for Amber20. A one step annihilation protocol with soft core potentials was implemented. Runs were conducted from a starting equilibrated ligand position extracted via K-means clustering from previous 300ns MD simulations. In this way, three independent replicas for each ligand were considered, as well as three replicas for solvation in pure water. Twelve windows were selected using Gaussian quadrature with 10ns of simulation time per window. To keep the ligand from wandering in TI calculations, we used a soft restraint of 10 kcal/mol$Å^2$.

5.5.5 Contact Maps and RMSF Calculations

Contacts maps by residue and Root Mean Squared Fluctuations (RMSF) calculations were generated using the native contacts and RMSF functions in CPPTRAJ [106]. The native contacts were defined relative to crystal PDB (6M0J) with a cutoff distance of 7 Å. The maps were averaged over 1000 snapshots taken every 100ps for each ligand. Root Mean Squared Fluctuations calculations were calculated with respect to a 100ns averaged structure using the mass-weighted average over the CA, N and C atoms and reported by residue. A total of 5000 snapshots were used taken every 20ps.

5.5.6 Strain Analysis

The effects of functional and non-functional fluctuations after ligand binding were performed through the application of the discrete form of the strain formalism from continuum theory

[81, 42]. In this form, derivatives are replaced by differential operators. The local neighborhood for a given central atom, i, is determined by a radius R that contains n number of atoms j. Distances between atoms i and j are related through the deformation matrix $\boldsymbol{F}$,

$$x_j - x_i = \boldsymbol{F}\left(x_{0,j} - x_{0,i}\right) \tag{5.1}$$

where x_i and x_j are the instantaneous position of atom i and j, and $x_{o,i}$ and $x_{o,j}$ correspond to their positions at any other given timestep, respectively. An optimized $\boldsymbol{F^*}$ is then sought, that minimizes the difference between actual distances and projected distances to an affine deformation,

$$\boldsymbol{F}^* = \min \sum_{j=1}^{n} \left[x_j - x_i - F\left(x_{0,j} - x_{0,i}\right)\right]^2. \tag{5.2}$$

The atomic strain tensor is determined by,

$$\varepsilon = \frac{1}{2}\left(F^T F - I\right) \tag{5.3}$$

whose magnitude is defined as the L2-norm of the shear term,

$$\mathrm{Tr}\left(\left(\varepsilon - \frac{1}{3}\,\mathrm{Tr}\,\varepsilon \cdot I\right)^2\right) \tag{5.4}$$

given that proteins are generally incompressible [81, 26].

For this study, a 10Å radius around each Cα atom for our analysis was considered. A simulation of RBD/ACE2 without ligand was used as reference for strain calculations to elucidate the binding of several ligands at different sites.

5.5.7 *Principal Component Analysis*

Principal Component Analysis (PCA) calculations were performed using the Bio3D package for R [40]. PCA is a dimensionality reduction technique that is effective in identifying correlated motions in atomic simulations of proteins from experimental structures or MD trajectories. Essential correlated conformational changes between structures can be represented in this low-dimensional subspace spanned by the first few principal components (PCs).

Mathematically, PCs are evaluated by diagonalizing the correlation matrix C_{ij},

$$\boldsymbol{R}^T \boldsymbol{C} \boldsymbol{R} = \boldsymbol{\Lambda} \tag{5.5}$$

for coordinates i and j,

$$C_{ij} = \left\langle \left(x_i - \langle x_i \rangle\right) \left(x_j - \left\langle x_j \right\rangle\right) \right. \tag{5.6}$$

where $x_1, \ldots, x_{3N}$ are the mass-weighted atomic coordinates of the protein, averaged over all sampled structures from simulation trajectories. The diagonal of the matrix Λ contains eigenvalues that correspond to the eigenvectors of R. Eigenvectors with the largest eigenvalues account for the highest proportion of variance within the dataset (first PC) and decrease sequentially while maintaining orthogonality to the first PC. The first several PCs are often considered "essential dynamics", and the rest are neglected without significant loss of information [2].

5.6 Results and discussion

5.6.1 *Calculating Binding Free Energies*

The configuration of the complex was derived from a recently published crystal structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor (PDB: 6M0J), whose conformation noticeably differed from the homology modeled structure used by Smith

Table 5.1: Ligand Binding Energies for Simple Glycan Model

Molecule	Oak Ridge ΔG (kcal/mol)	Docking ΔG (kcal/mol)	TI ΔG (kcal/mol)	RBD/ACE2 ΔΔG (kcal/mol)
Luteolin	-7.4	-6.19	-11.31 ± 0.62	-13.93 ± 1.63
Protirelin	-7.3	-8.45	-1.22 ± 5.79	-11.22 ± 5.10
Nitrofurantoin	-7.2	-9.22	-28.98 ± 4.83	-13.96 ± 2.84
Sapropterin	-7.1	-5.67	-6.10 ± 1.21	-18.89 ± 3.64
Vidarabine	-7.1	-6.24	-3.39 ± 1.70	-17.77 ± 1.66
Eriodictyol	-7.1	-7.86	-9.36 ± 2.41	-15.99 ± 3.53

et al. [115]. The glycosylation profile of the RBD/ACE2 protein was based on a simple glycan model and a glycomics-informed glycoproteomics structure ("Abundant" model) generated by Zhao *et al.* [139] which included one glycan group on the RBD domain and six on ACE2. Molecular dynamics simulations were carried out using the Amber20 simulation package [15]. To increase sampling, ensure convergence, and extract an equilibrated representative structure, K-means clustering was applied to three independent 100 ns molecular dynamics replicas *sans* ligand. To identify potential binding sites and validate the results of the previous docking study, candidates were docked using the Autodock4 software [84]. Three binding sites were identified for the simple glycan model: the binding interface between the S protein and the ACE2 receptor, a distal RBD region, and a site inside the ACE2 receptor (Figure 5.1). For the "Abundant" model, the interface and distal binding sites were identified. The structure with the highest binding affinity, as identified through docking, was used as the starting structure for production runs. Three 100 ns replicas for each ligand in complex with RBD/ACE2 were conducted to determine whether ligands would remain bound. Six ligands remained bound to the simple glycan model, and only three for the "Abundant" form. A representative structure with the ligand bound was extracted and used as a starting structure for three additional MD replicas totaling 300 ns for each ligand bound to the RBD/ACE2 complex and used for analysis.

Multiple binding free energy calculations were conducted to determine the binding affinity

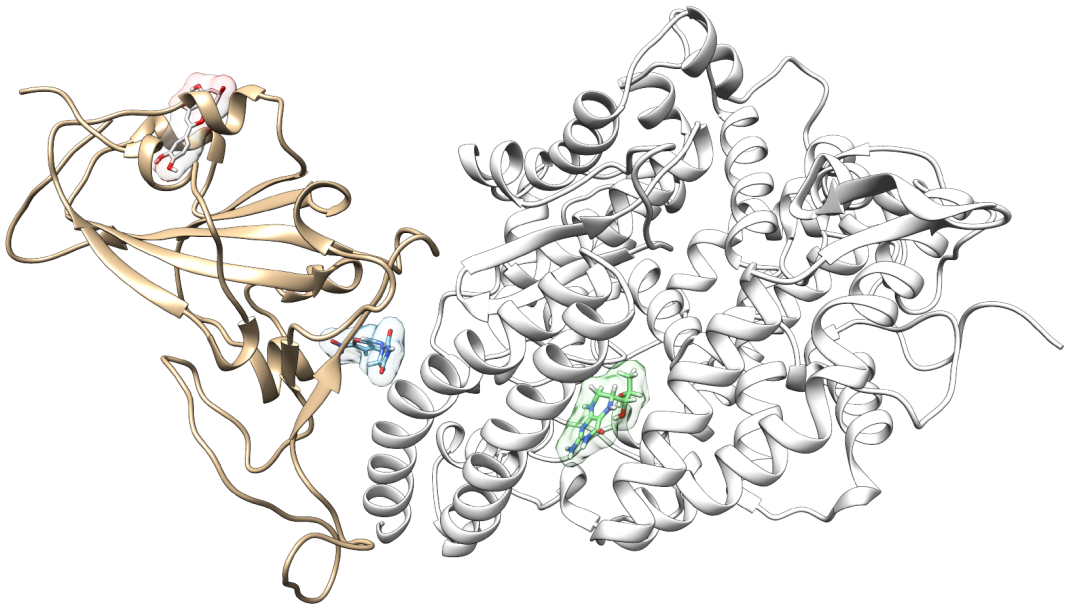

Figure 5.1: The RBD domain (tan) of SARS-CoV-2 recognizes ACE2 (white) as its receptor. We identify and characterize the interactions at three potential small-drug binding sites located at the binding interface between the RBD and the ACE2 receptor, inside the ACE2 protein, and a new previously unidentified distal site to which the drug Luteolin has a high binding affinity (red). Nitrofurantoin is shown in blue and Sapropterin is shown in green.

between the RBD/ACE2 complex and each small ligand (Table 5.1). The results were compared to those of the high-throughput screening [115]. We observed standard deviations ranging from 0.60-0.91 and 0.16-0.53 between our docking binding energies to that of the previous screening study for the simple and "Abundant" models, respectively. This could be due to the differences in the structures used (homology modeled vs. crystal structure) and the fact that important disulfide bonds and glycans groups were accounted for in our structure but were omitted in the Smith *et al.* analysis [115]. The absolute binding free

Table 5.2: Ligand Binding Energies for "Abundant" Glycan Model

Molecule	Oak Ridge ΔG (kcal/mol)	Docking ΔG (kcal/mol)	TI ΔG (kcal/mol)	RBD/ACE2 $\Delta\Delta$G (kcal/mol)
Luteolin	-7.4	-7.71	-15.04 $\pm$ 5.04	0.0027 $\pm$ 8.94
Isoniazid	-7.3	-6.57	-7.29 $\pm$ 4.54	11.87 $\pm$ 7.26
Eriodictyol	-7.1	-8.16	-12.65 $\pm$ 2.36	-8.06 $\pm$ 9.92

energies between ligand and protein reported here were determined using thermodynamic integration (TI) in the presence of explicit water. In our TI protocol, each ligand atom is treated as a softcore atom and is subsequently "removed" in a one-step alchemical cycle both in solution and in complex. This alchemical method, also referred to as free energy perturbation (FEB), allows for the calculation of absolute binding free energies (ABFE). We observed larger binding energies with standard deviations as high as 10.79 for the simple glycan structure and 5.04 for the "Abundant" glycoform in our free energy calculations than the docking binding energies. Of the three ligands studied, Luteolin showed the highest binding affinity for both glycoforms of the RBD/ACE2 complex. Luteolin was also the only ligand to bind at the distal site compared to the other candidates which were found to have a stronger binding affinity for the binding interface of RBD and ACE2 (Figure 5.1).

To determine how each compound affects S protein and ACE2 binding, we calculated the free energy of binding between the RBD domain of the S-protein and the ACE2 receptor in the presence of each ligand using the Molecular Mechanics/Generalized Born Surface Area (MM/GBSA) method. Tables 5.1 and 5.2 show the difference in binding free energy with ligand present versus without ($\Delta\Delta G_{Binding}$) for the simple and "Abundant" glycan models of the complex. In the case of the simple glycan model, our results show that every ligand enhances the binding free energy between the RBD domain and ACE2, acting as a binding stabilizer for the complex. Sapropterin, which binds inside ACE2 (Figure 5.1), appears to be the most potent stabilizer of the analyzed group. For the "Abundant" glycan model, our results show that Isoniazid acts as an allosteric inhibitor as indicated by the positive binding energy between the RBD domain and ACE2 ($\Delta\Delta G_{Binding}$) when the ligand is present. In contrast, Eriodictyol was observed to increase the binding affinity of the RBD domain to ACE2 thereby acting more as an allosteric stabilizer. Luteolin was observed to have negligible effect on the binding energy between RBD and ACE2. We hypothesize that inhibition/stabilization of the complex induced by these small molecules may serve to disrupt

downstream events such as S protein priming by TMPRSS2, an essential step for cellular entry by the virus.

5.6.2 Structural Analysis of Luteolin Binding

Figure 5.2 provides a closer look at the interactions between ligand and protein for each identified binding region. For both glycoprofiles, several of the potential binding sites were observed to localize to the same regions. Given its higher affinity for the RBD/ACE2 complex, we focus on the interactions of Nitrofurantoin (Figure 5.2A) within the binding interface. In this region, hydrogen bonds are the dominant interactions responsible for stabilizing Nitrofurantoin. These hydrogen bonds are formed between the carbonyl oxygens of Nitrofurantoin and ACE2 residue Lys353 and RBD residue Gln493. Figure 5.2B shows a representative configuration for Luteolin and the distal RBD binding site. Luteolin is stabilized by a hydrogen bond between its carbonyl oxygen and Tyr369 and pi-alkyl stacking between its aromatic group and Phe377. For comparison, Sapropterin, which was found to bind to only ACE2, is stabilized by hydrogen bonds formed with residues of the ACE2 protein (Figure 5.2C).

We study the changes of non-native contacts between the S protein and ACE2 receptor upon ligand binding by comparing the non-native contact maps of each compound (Figure 5.3). Similar changes in non-native contacts are observed for compounds Eriodictyol and Nitrofurantoin, which bind at the RBD/ACE2 interface (Figure 5.3A, 5.3B), and Sapropterin, which binds inside ACE2 (Figure 5.3C). Interestingly, binding of Luteolin (Figure 5.3D) to the distal binding region shows a clear difference in the number of contacts between the S protein and ACE2 receptor. These results suggest that an allosteric effect is induced when Luteolin binds to the distal binding site.

Figure 5.4 shows the estimated Root Mean Squared Fluctuations and the shear strain on each residue of the RBD/ACE2 complex for several ligands. Strain is a natural quantity to

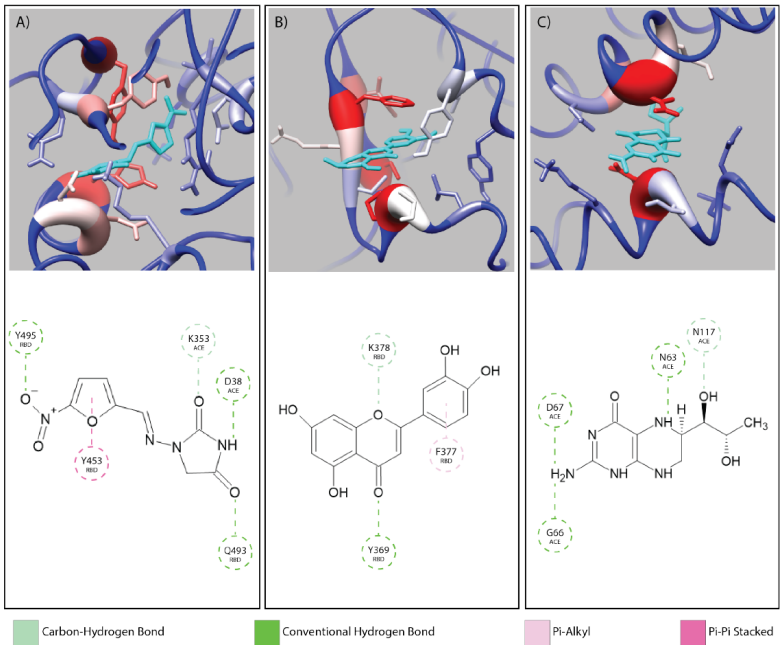

Figure 5.2: The interaction diagrams for equilibrated configurations of ligands at the interface (left, Nitrofurantoin), in the RBD domain (middle, Luteolin) and in ACE2 region (right, Sapropterin) are shown. Hydrogen bonds act as the dominating interactions responsible for stabilizing Nitrofurantoin which are formed between the carbonyl oxygens of Nitrofurantoin and the Lys353 residue of ACE2 and RBD residue Gln493. Luteolin is stabilized by a hydrogen bond between its carbonyl oxygen and Tyr369 and pi-alkyl stacking between its aromatic group and Phe377 of the RBD domain. Sapropterin, which is found buried in the ACE2 cavity is stabilized by hydrogen bonds.

study local protein deformations, which are a good indicator of mechanical allosteric coupling between two binding events on the same protein. This deformation upon ligand binding is calculated as the shear strain measured relative to the average conformation from all frames of the RBD/ACE2 sans ligand. Average Root Mean Square Fluctuations (RMSF), a measure of the displacement of a residue relative to the initial crystal structure (PDB: 6M0J), suggest high levels of deviations for RBD residues Asn481, Gly482 and Val483 near the binding interface of ACE2, particularly for the Luteolin ligand (Figures 5.4A,5.5A). Strain analysis also shows a significant peak at these residues (Figures 5.4B,5.5B). Consistent with the results of Figure 5.3D, this profound allosteric effect is induced when Luteolin is bound to the distal RBD binding site for both the simple and "Abundant" glycan forms of the RBD/ACE2

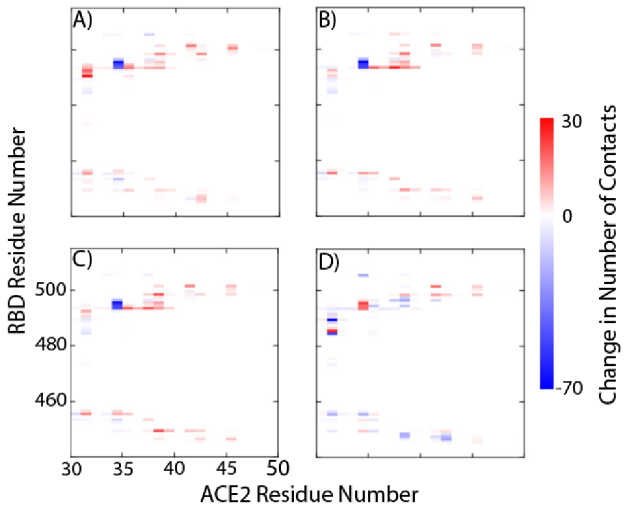

Figure 5.3: Relative protein-protein non-native contact maps in the presence of A) Eriodictyol, B) Nitrofurantoin, C) Sapropterin, and D) Luteolin. The relative non-native contact maps measure the change in contacts relative to the complex with no ligands (red more contacts, blue less contacts). From the graphs we see that the first three panels have identical contact profiles compared to D.

complex.

The formation of cross-correlation networks allows the transmission of information when the binding of a molecule at one site of the protein induces a change in local structure elsewhere in the protein. From our structural analysis, Luteolin was the only ligand to induce a conformational change to the RBD protein after binding to a distal site. To validate the extend by which the atomic fluctuations of the complex are correlated with one another in the presence of Luteolin, a dynamic cross-correlation network analysis was conducted using the Bio3D software [40]. Anti-correlations between -0.40 and -0.60 are represented as blue lines between several distant regions of the RBD, as seen in Figures 5.5C and 5.5D (ACE2 has been omitted for clarity). A comparison of simulations of the RBD/ACE2 with and without Luteolin shows the disruption of the dynamic cross-correlation network by this drug. The disappearance of elements of the network, which include anti-correlated sites between this distal binding site and the RBD/ACE2 interface, may be critical in mediating allosteric transitions and ACE2 binding.

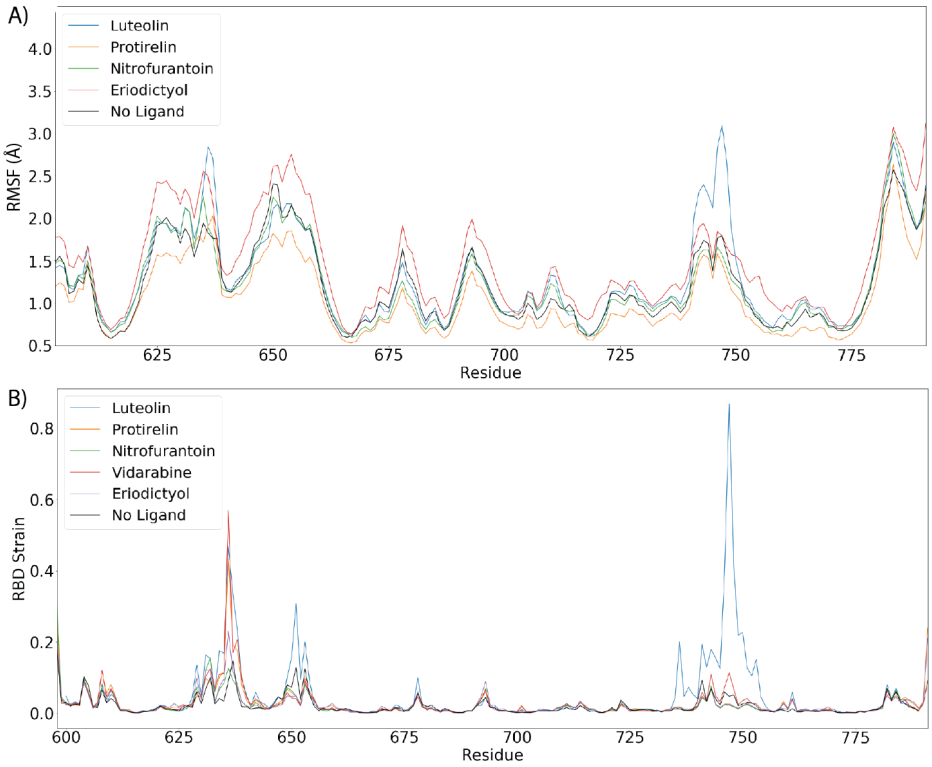

Figure 5.4: Luteolin induces large allosteric strain when RBD domain is bound to ACE2. A) Root Mean Squared Fluctuations (RMSF) between the RBD/ACE2 complex with top scoring ligand. B) Shear strains mapped. For shear train calculations, only Cα atoms are included. Strain analysis suggest a strong allosteric strain to the ACE2 binding region of the RBD domain when RBD is in complex with ACE2 and Luteolin.

Principal component analysis (PCA) was used to gain quantitative insights into the binding of Luteolin to the RBD/ACE2 complex. The orthogonal eigenvectors of the resulting principal components (PCs) describe the maximal variance of the distributions of structures. Details of the data processing are described in the Materials and Methods section. The two dominant principal components (PCs) were sufficient to describe the observed conformational changes at the RBD/ACE2 interface. The first principal component (PC1) corresponds to the rotational motion of the RBD/ACE2 complex. The second principal component (PC2) captures the conformational change observed at the RBD/ACE2 interface induced by Luteolin binding (Figure 5.6). As previously discussed, residues near the binding site were observed to be anti-correlated to residues near the ACE2 binding site. Binding of Luteolin disrupts this

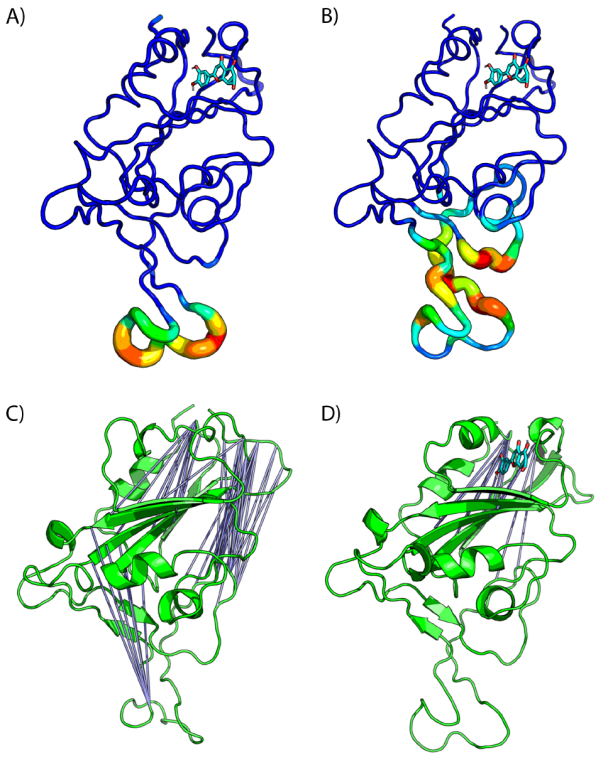

Figure 5.5: Bound Luteolin induces large strain at RBD/ACE binding region. A) Difference in estimated Root Mean Squared Fluctuations (RMSF) between the RBD domain and Luteolin in complex are shown. (B) Shear strains mapped onto RBD/LUT complex. Regions flanking disulfide bonds have the highest atomic fluctuations that contribute the deformation energy of the ACE2 binding region. C,D) Distal site binding disrupts intramolecular RBD interactions inducing conformational changes at ACE2 binding interface. Visualization of residue-residue cross correlations. Blue lines indicate anti-correlation motions with values between -0.4 and -0.6. Higher correlations between distal sites *sans* ligand (C) and in the presence of Luteolin (D).

cross-correlation network, thereby inducing a conformational shift near the ACE2 binding site. Conformational changes to the RBD protein, as captured by PCA, confirm not only correlated motions of the RBD protein but suggest that the "hotspot" identified here may be a useful target in the discovery and design of new therapeutics that modulate the RBD/ACE2 protein-protein interface. Taken together, our results suggest that Luteolin may serve as an important small-molecule allosteric modulator that affects RBD/ACE2 protein dynamics and may provide an alternative therapeutic approach.

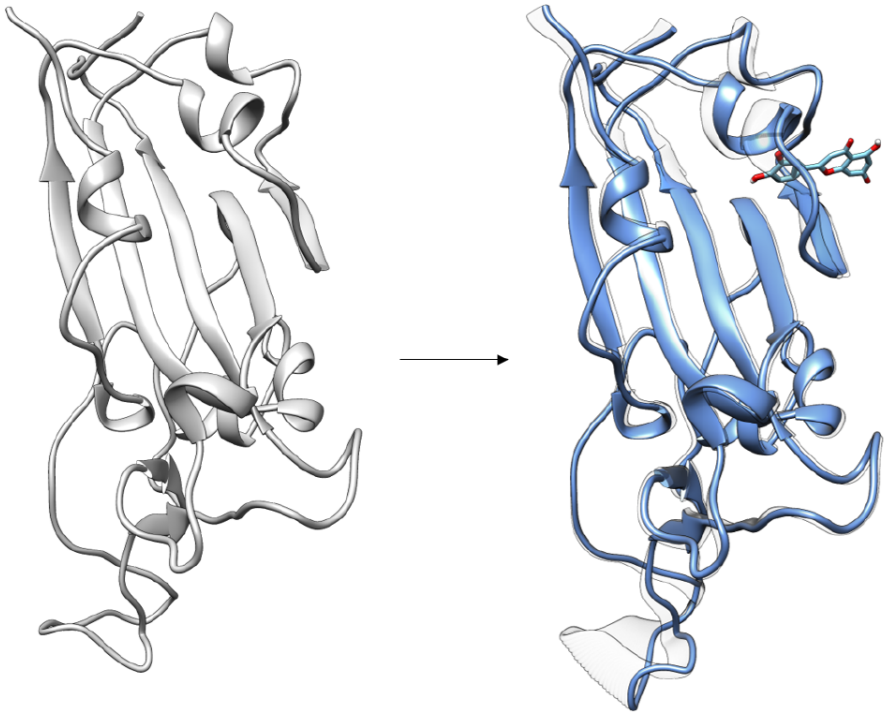

Figure 5.6: Distal binding by Luteolin induces conformational changes at ACE2 binding site. Induced conformational changes at loop binding regions are visualized as captured by the second dominant principal component. Images of important conformational changes are superimposed to emphasize conformations changes introduced after Luteolin binding to distal binding site.

5.7 Conclusion

The repurposing of approved drug candidates provides an alternative strategy to identify lead candidates for viral infections. The time and expense of bringing a drug to market are significantly reduced when the safety and pharmacokinetic profiles of existing drugs are already known. Drug repositioning is guided by a rational approach that requires detailed knowledge of the target structure, the spatial arrangement, interactions and structural conformation of the compound, and the mechanism of action [102]. For protein-protein complexes, the full binding inhibition is hard to achieve using small molecules, and novel approaches based on the concept of interfacial inhibition are needed, where macromolecules are trapped in dead-end complexes that cannot fulfill their biological function [33]. In this sense, allosteric

modulation of PPIs of the RBD/ACE2 complex may serve as an approach in drug discovery. While current high-throughput virtual screenings have identified several potential therapeutic candidates, the mechanism by which they affect protein-protein interactions remains unclear.

In this study, we have focused on determining the binding mechanism and associated conformational changes in the presence of the top compounds from a docking screen recently conducted by Smith *et al.* [115]. After filtering compounds by their residence time in complex through MD simulations, rigorous binding free energy calculations were conducted in explicit water using thermodynamic integration. Our results have revealed a range of free energies and binding positions that were not anticipated in the original docking studies. The majority of such positions lie far from the interface, and the free energy extrema correspond to the shallow interface region of binding (Tables 5.1 and 5.2). The relative RBD/ACE2 binding free energy changes induced by the top ligands considered here indicate a binding enhancement of the RBD/ACE2 complex, even for drugs bound in the interface region. This enhancement is accompanied by a modification of the non-native contact maps of the protein interface, which lead to changes in the manner in which the proteins interact. Taken together, our results suggest that these ligands could not act as ordinary competitive inhibitors for RBD, since there is no binding disruption, but, given the contact modification and the energy differences with respect to the pure RBD/ACE2 complex, we propose that they may inhibit the system by serving as allosteric or direct PPI stabilizers [85, 124].

PPIs with large, flat and featureless surfaces, as in the case in the RBD/ACE2 complex, lack good drug-binding pockets for ligands. Recent approaches have shown that allosteric modulators provide an alternative strategy to target PPIs such as the RBD/ACE2 complex. Our results have uncovered pronounced effects at the RBD/ACE2 interface upon ligand binding at a distal site. In particular, upon Luteolin binding, RMSF and strain analysis unveil significant levels of fluctuations and strain in RBD regions away from the binding site. In addition, cross-correlational analysis reveals the disruption of anti-correlated motions

between the Luteolin binding site and the binding interface of RBD/ACE2. Future research will require investigating the possible binding motifs for Luteolin and its effects on S protein binding to the ACE2 receptor. Beyond the finding that Luteolin might serve to inhibit viral entry, the discovery of this distal binding site offers potential for the design and engineering of future therapeutics for COVID-19.

5.8 Acknowledgments

The simulations reported here were carried out on the GPU cluster supported by the NSF through grant DMR1828629. The authors are grateful to the Research Computing Center of the University of Chicago for additional computational resources.

CHAPTER 6

CONCLUSION

6.1 Summary of contributions

This dissertation investigates the complex structure of chromatin and the various mechanisms that govern its organization. Our initial focus was on the application of machine learning techniques in conjunction with a newly established mesoscale chromatin model. This approach allowed us to explore chromatin's structural characteristics and identify transitional motifs that dictate DNA accessibility.

Building on these findings, we then turned our attention to the development of a denoising autoencoder (DAE). The DAE demonstrated its ability to produce high-resolution STEM images of individual nucleosomes and smaller domains located within chromatin-dense regions. The final part of our research focused on the impacts of nucleosome-nucleosome interaction strength on the overall structure of the chromatin fiber, a phenomenon influenced by post-translational modifications.

Collectively, our studies have significantly broadened our understanding of the intricate dynamics that control the structure and organization of chromatin. We've highlighted the delicate equilibrium between alpha and beta structures, as well as the critical role of interaction strengths and nucleosome repeat lengths (NRLs), offering new insights into a potential regulatory mechanism for chromatin organization across multiple scales.

Our research also serves as an emerging paradigm wherein machine learning methods are leveraged to bridge the gap between experimental imaging and theoretical modeling. Our collaboration with experimentalists resulted in a comprehensive and detailed investigation of chromatin's structural organization from experimental images. Overall, our work underscores how machine learning can facilitate a more thorough exploration of structural organization within biological systems.

6.2 Future directions

Based on the results presented in this thesis, future research could concentrate on deciphering the kinetic rates of tetranucleosomes motifs and determining how post-translational modifications affect their dynamics. Utilizing the framework we established while developing our denoising autoencoder (DAE), future endeavors could pivot towards the examination of smaller di- and trinucleosome motifs. By deciphering the complexities of the post-translational modification code and its effect on the chromatin fiber, we hope to increase our understanding of gene regulation and the mechanisms that govern the structure and function of chromatin. This dissertation lays the groundwork for these future studies and provides a solid basis for further investigation.

REFERENCES

[1] Walter Alvarado, Joshua Moller, Andrew L. Ferguson, and Juan J. de Pablo. Tetranucleosome interactions drive chromatin folding. *ACS Central Science*, 7(6):1019–1027, 2021. doi:10.1021/acscentsci.1c00085. URL https://doi.org/10.1021/acscentsci.1c00085. PMID: 34235262.

[2] Andrea Amadei, Antonius BM Linssen, and Herman JC Berendsen. Essential dynamics of proteins. *Proteins: Structure, Function, and Bioinformatics*, 17(4):412–425, 1993.

[3] Anthony T Annunziato. Dna packaging: nucleosomes and chromatin. *Nature Education*, 1(1):26, 2008. URL https://www.nature.com/scitable/topicpage/dna-packaging-nucleosomes-and-chromatin-310/.

[4] Gaurav Arya and Tamar Schlick. Role of histone tails in chromatin folding revealed by a mesoscopic oligonucleosome model. *Proceedings of the National Academy of Sciences of the United States of America*, 103(44):16236–16241, oct 2006. ISSN 00278424. doi:10.1073/PNAS.0604817103/SUPPL_FILE/04817SUPPAPPENDIXES1-3.PDF. URL https://www.pnas.org/doi/abs/10.1073/pnas.0604817103.

[5] SS Ashwin, Tadasu Nozaki, Kazuhiro Maeshima, and Masaki Sasai. Organization of fast and slow chromatin revealed by single-nucleosome dynamics. *Proceedings of the National Academy of Sciences*, 116(40):19939–19944, 2019.

[6] Sandro Baldi, Philipp Korber, and Peter B. Becker. Beads on a string—nucleosome array arrangements and folding of the chromatin fiber. *Nature Structural & Molecular Biology 2020 27:2*, 27(2):109–118, feb 2020. ISSN 1545-9985. doi:10.1038/s41594-019-0368-x. URL https://www.nature.com/articles/s41594-019-0368-x.

[7] Mariano Barbieri, Mita Chotalia, James Fraser, Liron-Mark Lavitas, Josée Dostie, Ana Pombo, and Mario Nicodemi. Complexity of chromatin folding is captured by the strings and binders switch model. *Proceedings of the National Academy of Sciences*, 109(40):16173–16178, 2012.

[8] Davide Baù, Amartya Sanyal, Bryan R Lajoie, Emidio Capriotti, Meg Byron, Jeanne B Lawrence, Job Dekker, and Marc A Marti-Renom. The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules. *Nature structural & molecular biology*, 18(1):107, 2011.

[9] Peter Binev, Francisco Blanco-Silva, Douglas Blom, Wolfgang Dahmen, Philipp Lamby, Robert Sharpley, and Thomas Vogt. *High-Quality Image Formation by Nonlocal Means Applied to High-Angle Annular Dark-Field Scanning Transmission Electron Microscopy (HAADF–STEM)*, pages 127–145. Springer US, Boston, MA, 2012. ISBN 978-1-4614-2191-7. doi:10.1007/978-1-4614-2191-7_5. URL https://doi.org/10.1007/978-1-4614-2191-7_5.

[10] Lorenzo Boninsegna, Gianpaolo Gobbo, Frank Noé, and Cecilia Clementi. Investigating molecular kinetics by variationally optimized diffusion maps. *Journal of chemical theory and computation*, 11(12):5947–5960, 2015.

[11] Antoni Buades, Bartomeu Coll, and Jean-Michel Morel. Non-local means denoising. *Image Processing On Line*, 1:208–212, 2011.

[12] James P. Buban, Quentin Ramasse, Bryant Gipson, Nigel D. Browning, and Henning Stahlberg. High-resolution low-dose scanning transmission electron microscopy. *Journal of Electron Microscopy*, 59(2):103–112, 11 2009. ISSN 0022-0744. doi:10.1093/jmicro/dfp052. URL `https://doi.org/10.1093/jmicro/dfp052`.

[13] E Cardellini, S Cinelli, G L Gianfranceschi, G Onori, A Santucci, and & L Urbanelli. Differential scanning calorimetry of chromatin at different levels of condensation. *Molecular Biology Reports*, 27:175–180, 2000.

[14] Albert Cardona and Pavel Tomancak. Current challenges in open-source bioimage informatics. *Nature methods*, 9(7):661–665, 2012.

[15] David A Case, Kellon Belfon, Ido Ben-Shalom, Scott R Brozell, David Cerutti, Thomas Cheatham, Vinícius Wilian D Cruzeiro, Tom Darden, Robert E Duke, George Giambasu, et al. Amber 2020. 2020.

[16] Chen Chen, Hong Hwa Lim, Jian Shi, Sachiko Tamura, Kazuhiro Maeshima, Uttam Surana, and Lu Gan. Budding yeast chromatin is dispersed in a crowded nucleoplasm in vivo. *Molecular biology of the cell*, 27(21):3357–3368, 2016.

[17] Wei Chen, Hythem Sidky, and Andrew L Ferguson. Nonlinear discovery of slow molecular modes using state-free reversible vampnets. *The Journal of chemical physics*, 150(21):214114, 2019.

[18] François Chollet. Keras. `https://keras.io`, 2015.

[19] Ronald R Coifman and Stéphane Lafon. Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30, 2006.

[20] Antonia Creswell, Kai Arulkumaran, and Anil A Bharath. On denoising autoencoders trained to minimise binary cross-entropy. *arXiv preprint arXiv:1708.08487*, 2017.

[21] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising with block-matching and 3d filtering. In *Image processing: algorithms and systems, neural networks, and machine learning*, volume 6064, pages 354–365. SPIE, 2006.

[22] Tim Dahmen, Holger Kohr, Andrew R. Lupini, Jean-Pierre Baudoin, Christian Kübel, Patrick Trampert, Philipp Slusallek, and Niels de Jonge. Combined tilt- and focal-series tomography for haadf-stem. *Microscopy Today*, 24(3):26–31, 2016. doi:10.1017/S1551929516000328.

[23] Curt A. Davey, David F. Sargent, Karolin Luger, Armin W. Maeder, and Timothy J. Richmond. Solvent mediated interactions in the structure of the nucleosome core particle at 1.9å resolution††we dedicate this paper to the memory of max perutz who was particularly inspirational and supportive to t.j.r. in the early stages of this study. *Journal of Molecular Biology*, 319(5):1097–1113, 2002. ISSN 0022-2836. doi:https://doi.org/10.1016/S0022-2836(02)00386-8. URL `https://www.sciencedirect.com/science/article/pii/S0022283602003868`.

[24] Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases*, 20(5):533–534, 2020.

[25] Benedetta Dorigo, Thomas Schalch, Alexandra Kulangara, Sylwia Duda, Rasmus R Schroeder, and Timothy J Richmond. Nucleosome arrays reveal the two-start organization of the chromatin fiber. *Science*, 306(5701):1571–1573, 2004.

[26] Jean-Pierre Eckmann, Jacques Rougemont, and Tsvi Tlusty. Colloquium: Proteins: The physics of amorphous evolving matter. *Reviews of Modern Physics*, 91(3):031001, 2019.

[27] Babatunde Ekundayo, Timothy J Richmond, and Thomas Schalch. Capturing structural heterogeneity in chromatin fibers. *Journal of molecular biology*, 429(20):3031–3042, 2017.

[28] Mikhail Eltsov, Kirsty M MacLellan, Kazuhiro Maeshima, Achilleas S Frangakis, and Jacques 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*, 105(50):19732–19737, 2008.

[29] Mikhail Eltsov, Diana Grewe, Nicolas Lemercier, Achilleas Frangakis, Françoise Livolant, and Amélie Leforestier. Nucleosome conformational variability in solution and in interphase nuclei evidenced by cryo-electron microscopy of vitreous sections. *Nucleic acids research*, 46(17):9189–9200, 2018.

[30] Peter Ercius, Osama Alaidi, Matthew J. Rames, and Gang Ren. Electron tomography: A three-dimensional analytic tool for hard and soft materials research. *Advanced Materials*, 27(38):5638–5663, 2015. doi:https://doi.org/10.1002/adma.201501015. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/adma.201501015`.

[31] Linwei Fan, Fan Zhang, Hui Fan, and Caiming Zhang. Brief review of image denoising techniques. *Visual Computing for Industry, Biomedicine, and Art*, 2(1):1–12, 2019.

[32] Andrew L. Ferguson, Athanassios Z. Panagiotopoulos, Ioannis G. Kevrekidis, and Pablo G. Debenedetti. Nonlinear dimensionality reduction in molecular simulation: The diffusion map approach. *Chemical Physics Letters*, 509(1–3):1–11, 2011. ISSN 00092614. doi:10.1016/j.cplett.2011.04.066.

[33] Yann Ferrandez, Wenhua Zhang, François Peurois, Lurlène Akendengué, Anne Blangy, Mahel Zeghouf, and Jacqueline Cherfils. Allosteric inhibition of the guanine nucleotide exchange factor dock5 by a small molecule. *Scientific reports*, 7(1):1–13, 2017.

[34] John T Finch and Aaron Klug. Solenoidal model for superstructure in chromatin. *Proceedings of the National Academy of Sciences*, 73(6):1897–1901, 1976.

[35] Gordon S. Freeman, Joshua P. Lequieu, Daniel M. Hinckley, Jonathan K. Whitmer, and Juan J. de Pablo. Dna shape dominates sequence affinity in nucleosome formation. *Physical Review Letters*, 113:168101, Oct 2014. doi:10.1103/PhysRevLett.113.168101. URL `https://link.aps.org/doi/10.1103/PhysRevLett.113.168101`.

[36] Jonas J Funke, Philip Ketterer, Corinna Lieleg, Sarah Schunter, Philipp Korber, and Hendrik Dietz. Uncovering the forces between nucleosomes using DNA origami. *Science Advances*, 2(11):e1600974, 2016. doi:10.1126/sciadv.1600974. URL `https://www.science.org/doi/abs/10.1126/sciadv.1600974`.

[37] Paola Gavazzo, Laura Vergani, Gian Carlo Mascetti, and Claudio Nicolini. Effects of Histone Acetylation on Chromatin Structure. *J. Cell. Biochem*, 64:466–475, 1997. doi:10.1002/(SICI)1097-4644(19970301)64:3. URL `https://onlinelibrary.wiley.com/doi/10.1002/`.

[38] Michael J Gorczynski, Jolanta Grembecka, Yunpeng Zhou, Yali Kong, Liya Roudaia, Michael G Douvas, Miki Newman, Izabela Bielnicka, Gwen Baber, Takeshi Corpora, et al. Allosteric inhibition of the protein-protein interaction between the leukemia-associated proteins runx1 and cbfβ. *Chemistry & biology*, 14(10):1186–1197, 2007.

[39] JAR Gordon, RA Grandy, JB Lian, JL Stein, Andre J van Wijnen, and GS Stein. *Chromatin*. Elsevier Inc., 2013.

[40] Barry J Grant, Ana PC Rodrigues, Karim M ElSawy, J Andrew McCammon, and Leo SD Caves. Bio3d: an r package for the comparative analysis of protein structures. *Bioinformatics*, 22(21):2695–2696, 2006.

[41] Sergei A. Grigoryev and Christopher L. Woodcock. Chromatin organization - the 30nm fiber. *Experimental Cell Research*, 318(12):1448–1455, 2012. ISSN 0014-4827. doi:https://doi.org/10.1016/j.yexcr.2012.02.014. URL `https://www.sciencedirect.com/science/article/pii/S0014482712000900`.

[42] PM Gullett, MF Horstemeyer, MI Baskes, and H Fang. A deformation gradient tensor and strain tensors for atomistic simulations. *Modelling and Simulation in Materials Science and Engineering*, 16(1):015001, 2007.

[43] Ashley Z Guo, Joshua Lequieu, and Juan J de Pablo. Extracting collective motions underlying nucleosome dynamics via nonlinear manifold learning. *The Journal of Chemical Physics*, 150(5):054902, 2019.

[44] Doga Gürsoy, Francesco De Carlo, Xianghui Xiao, and Chris Jacobsen. Tomopy: a framework for the analysis of synchrotron tomographic data. *Journal of Synchrotron Radiation*, 21(5):1188–1193, 2014.

[45] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi:10.1038/s41586-020-2649-2. URL `https://doi.org/10.1038/s41586-020-2649-2`.

[46] Daniel M Hinckley, Gordon S Freeman, Jonathan K Whitmer, and Juan 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*, 139 (14):10B604_1, 2013.

[47] Markus Hoffmann, Hannah Kleine-Weber, Simon Schroeder, Nadine Krüger, Tanja Herrler, Sandra Erichsen, Tobias S Schiergens, Georg Herrler, Nai-Huei Wu, Andreas Nitsche, et al. Sars-cov-2 cell entry depends on ace2 and tmprss2 and is blocked by a clinically proven protease inhibitor. *cell*, 181(2):271–280, 2020.

[48] Alain Horé and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *20th International Conference on Pattern Recognition*, pages 2366–2369, 2010. doi:10.1109/ICPR.2010.579.

[49] Tsung-Han S Hsieh, Assaf Weiner, Bryan Lajoie, Job Dekker, Nir Friedman, and Oliver J Rando. Mapping nucleosome resolution chromosome folding in yeast by micro-c. *Cell*, 162(1):108–119, 2015.

[50] Yuan Huang, Chan Yang, Xin-feng Xu, Wei Xu, and Shu-wen Liu. Structural and functional properties of sars-cov-2 spike protein: potential antivirus drug development for covid-19. *Acta Pharmacologica Sinica*, 41(9):1141–1149, 2020.

[51] William Humphrey, Andrew Dalke, and Klaus Schulten. Vmd: visual molecular dynamics. *Journal of molecular graphics*, 14(1):33–38, 1996.

[52] Van AT Huynh, Philip JJ Robinson, and Daniela Rhodes. A method for the in vitro reconstitution of a defined "30 nm" chromatin fibre containing stoichiometric amounts of the linker histone. *Journal of molecular biology*, 345(5):957–968, 2005.

[53] JGraph. Diagrams.net, 10 2021. URL `https://github.com/jgraph/drawio`.

[54] Jacob Joffe, Michael Keene, and Harold Weintraub. Histones h2a, h2b, h3, and h4 are present in equimolar amounts in chick erythroblasts. *Biochemistry*, 16(6):1236–1238, mar 1977. ISSN 15204995. doi:10.1021/bi00625a032. URL `https://pubs.acs.org/sharingguidelines`.

[55] Friedrich K Jondral. White gaussian noise–models for engineers. *Frequenz*, 72(5-6): 293–299, 2018.

[56] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. URL `https://arxiv.org/abs/1412.6980`.

[57] James R Kremer, David N Mastronarde, and J Richard McIntosh. Computer visualization of three-dimensional image data using imod. *Journal of Structural Biology*, 116(1): 71–76, 1996.

[58] Christian Kübel, Andreas Voigt, Remco Schoenmakers, Max Otten, David Su, Tan-Chen Lee, Anna Carlsson, and John Bradley. Recent advances in electron tomography: Tem and haadf-stem tomography for materials science and semiconductor applications. *Microscopy and Microanalysis*, 11(5):378–400, 2005.

[59] Joshua Lequieu, Andrés Córdoba, David C Schwartz, and Juan J de Pablo. Tension-dependent free energies of nucleosome unwrapping. *ACS central science*, 2(9):660–666, 2016.

[60] Joshua Lequieu, David C. Schwartz, and Juan J. de Pablo. In silico evidence for sequence-dependent nucleosome sliding. *Proceedings of the National Academy of Sciences of the United States of America*, 114(44):E9197–E9205, 2017. ISSN 0027-8424. doi:10.1073/pnas.1705685114. URL `https://www.pnas.org/content/114/44/E9197`.

[61] Joshua Lequieu, Andrés Córdoba, Joshua Moller, and Juan J De Pablo. 1cpn: A coarse-grained multi-scale model of chromatin. *The Journal of Chemical Physics*, 150 (21):215102, 2019.

[62] Yue Li, Eric Roth, Vasundhara Agrawal, Adam Eshein, Jane Fredrick, Luay Almassalha, Anne Shim, Reiner Bleher, Vinayak P Dravid, and Vadim Backman. Quantifying three-dimensional chromatin organization utilizing scanning transmission electron microscopy: Chromstem. *bioRxiv*, page 636209, 2019.

[63] Yue Li, Adam Eshein, Ranya K.A. Virk, Aya Eid, Wenli Wu, Jane Frederick, David VanDerway, Scott Gladstein, Kai Huang, Anne R. Shim, Nicholas M. Anthony, Greta M. Bauer, Xiang Zhou, Vasundhara Agrawal, Emily M. Pujadas, Surbhi Jain, George Esteve, John E. Chandler, The-Quyen Nguyen, Reiner Bleher, Juan J. de Pablo, Igal Szleifer, Vinayak P. Dravid, Luay M. Almassalha, and Vadim Backman. Nanoscale chromatin imaging and analysis platform bridges 4d chromatin organization with molecular function. *Science Advances*, 7(1):eabe4310, 2021.

[64] Yue Li, Vasundhara Agrawal, Ranya KA Virk, Eric Roth, Wing Shun Li, Adam Eshein, Jane Frederick, Kai Huang, Luay Almassalha, Reiner Bleher, Marcelo A Carignano, Igal Szleifer, Vinayak P Dravid, and Vadim Backman. Analysis of three-dimensional chromatin packing domains by chromatin scanning transmission electron microscopy (chromstem). *Scientific Reports*, 12(1):1–15, 2022.

[65] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[66] Karolin Luger, Mekonnen L Dechassa, and David J Tremethick. New insights into nucleosome and chromatin structure: an ordered state or a disordered affair? *Nature Reviews Molecular Cell Biology*, 13(7):436–447, 2012. ISSN 1471-0080. doi:10.1038/nrm3382. URL `https://doi.org/10.1038/nrm3382`.

[67] Pradeep K. Luther. *Sample Shrinkage and Radiation Damage of Plastic Sections*, pages 17–48. Springer New York, New York, NY, 2006. ISBN 978-0-387-69008-7. doi:10.1007/978-0-387-69008-7_2. URL `https://doi.org/10.1007/978-0-387-69008-7_2`.

[68] Kazuhiro Maeshima, Saera Hihara, and Mikhail Eltsov. Chromatin structure: does the 30-nm fibre exist in vivo? *Current Opinion in Cell Biology*, 22(3):291–297, 2010.

[69] Kazuhiro Maeshima, Sachiko Tamura, Jeffrey C Hansen, and Yuji Itoh. Fluid-like chromatin: Toward understanding the real chromatin organization present in the cell. *Current opinion in cell biology*, 64:77–89, 2020.

[70] Angshul Majumdar and Aditay Tripathi. Asymmetric stacked autoencoder. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 911–918, 2017. doi:10.1109/IJCNN.2017.7965949.

[71] Siavash Maraghechi, Johan P.M. Hoefnagels, Ron H.J. Peerlings, and Marc G.D. Geers. Correction of scan line shift artifacts in scanning electron microscopy: An extended digital image correlation framework. *Ultramicroscopy*, 187:144–163, 2018. ISSN 0304-3991. doi:https://doi.org/10.1016/j.ultramic.2018.01.002. URL `https://www.sciencedirect.com/science/article/pii/S0304399117300918`.

[72] Andreas Mardt, Luca Pasquali, Hao Wu, and Frank Noé. Vampnets for deep learning of molecular kinetics. *Nature communications*, 9(1):5, 2018.

[73] WF Marshall, A Straight, JF Marko, J Swedlow, A Dernburg, A Belmont, AW Murray, DA Agard, and JW Sedat. Interphase chromosomes undergo constrained diffusional motion in living cells. *Current Biology*, 7(12):930–939, 1997.

[74] Fabrizio Martino, Stephanie Kueng, Philip Robinson, Monika Tsai-Pflugfelder, Fred van Leeuwen, Mathias Ziegler, Fabien Cubizolles, Moira M Cockell, Daniela Rhodes, and Susan M Gasser. Reconstitution of yeast silent chromatin: multiple contact sites and o-aadpr binding load sir complexes onto nucleosomes in vitro. *Molecular cell*, 33 (3):323–334, 2009.

[75] James D McGhee, Donald C Rau, Elliot Charney, and Gary Felsenfeld. Orientation of the nucleosome within the higher order structure of chromatin. *Cell*, 22(1):87–96, 1980.

[76] Robert K. McGinty and Song Tan. Nucleosome structure and function. *Chemical Reviews*, 115(6):2255–2273, 2015. doi:10.1021/cr500373h. URL https://doi.org/10.1021/cr500373h. PMID: 25495456.

[77] Erik Meijering, Anne E Carpenter, Hanchuan Peng, Fred A Hamprecht, and Jean-Christophe Olivo-Marin. Imagining the future of bioimage analysis. *Nature biotechnology*, 34(12):1250–1255, 2016.

[78] Niklas Mevenkamp, Peter Binev, Wolfgang Dahmen, Paul M Voyles, Andrew B Yankovich, and Benjamin Berkels. Poisson noise removal from high-resolution stem images based on periodic block matching. *Advanced Structural and Chemical Imaging*, 1(1):1–19, 2015.

[79] P.A. Midgley and M. Weyland. 3d electron microscopy in the physical sciences: the development of z-contrast and eftem tomography. *Ultramicroscopy*, 96(3):413–431, 2003. ISSN 0304-3991. doi:https://doi.org/10.1016/S0304-3991(03)00105-0. URL https://www.sciencedirect.com/science/article/pii/S0304399103001050. Proceedings of the International Workshop on Strategies and Advances in Atomic Level Spectroscopy and Analysis.

[80] Bill R Miller III, T Dwight McGee Jr, Jason M Swails, Nadine Homeyer, Holger Gohlke, and Adrian E Roitberg. Mmpbsa. py: an efficient program for end-state free energy calculations. *Journal of chemical theory and computation*, 8(9):3314–3321, 2012.

[81] Michael R Mitchell, Tsvi Tlusty, and Stanislas Leibler. Strain analysis of protein structures and low dimensionality of mechanical allosteric couplings. *Proceedings of the National Academy of Sciences*, 113(40):E5847–E5855, 2016.

[82] Joshua Moller, Joshua Lequieu, and Juan J de Pablo. The free energy landscape of internucleosome interactions and its relation to chromatin fiber structure. *ACS Central Science*, 5(2):341–348, 2019.

[83] Manuela Moraru and Thomas Schalch. Chromatin fiber structural motifs as regulatory hubs of genome function? *Essays in biochemistry*, 63(1):123–132, 2019.

[84] Garrett M Morris, Ruth Huey, William Lindstrom, Michel F Sanner, Richard K Belew, David S Goodsell, and Arthur J Olson. Autodock4 and autodocktools4: Automated docking with selective receptor flexibility. *Journal of computational chemistry*, 30(16): 2785–2791, 2009.

[85] Duan Ni, Shaoyong Lu, and Jian Zhang. Emerging roles of allosteric modulators in the regulation of protein-protein interactions (ppis): A new paradigm for ppi drug discovery. *Medicinal research reviews*, 39(6):2314–2342, 2019.

[86] C Niedermeier and K Schulten. Molecular dynamics simulations in heterogeneous dielectrica and debye-hückel media-application to the protein bovine pancreatic trypsin inhibitor. *Molecular simulation*, 8(6):361–387, 1992.

[87] Yoshinori Nishino, Mikhail Eltsov, Yasumasa Joti, Kazuki Ito, Hideaki Takata, Yukio Takahashi, Saera Hihara, Achilleas S Frangakis, Naoko Imamoto, Tetsuya Ishikawa, et al. Human mitotic chromosomes consist predominantly of irregularly folded nucleosome fibres without a 30-nm chromatin structure. *The EMBO journal*, 31(7):1644–1653, 2012.

[88] Frank Noé and Cecilia Clementi. Kinetic distance and kinetic maps from molecular dynamics simulation. *Journal of chemical theory and computation*, 11(10):5002–5011, 2015.

[89] Masae Ohno, Tadashi Ando, David G Priest, Vipin Kumar, Yamato Yoshida, and Yuichi Taniguchi. Sub-nucleosomal genome structure reveals distinct nucleosome folding motifs. *Cell*, 176(3):520–534, 2019.

[90] Nisreen MA Okba, Marcel A Müller, Wentao Li, Chunyan Wang, Corine H GeurtsvanKessel, Victor M Corman, Mart M Lamers, Reina S Sikkema, Erwin de Bruin, Felicity D Chandler, et al. Severe acute respiratory syndrome coronavirus 2- specific antibody responses in coronavirus disease patients. *Emerging infectious diseases*, 26(7): 1478–1488, 2020.

[91] Colin Ophus. A fast image simulation algorithm for scanning transmission electron microscopy. *Advanced Structural and Chemical Imaging*, 3(1):1–11, 2017.

[92] Jonathan M Ostrem, Ulf Peters, Martin L Sos, James A Wells, and Kevan M Shokat. K-ras (g12c) inhibitors allosterically control gtp affinity and effector interactions. *Nature*, 503(7477):548–551, 2013.

[93] Horng D. Ou, Sébastien Phan, Thomas J. Deerinck, Andrea Thor, Mark H. Ellisman, and Clodagh C. O'Shea. Chromemt: Visualizing 3d chromatin structure and compaction in interphase and mitotic cells. *Science*, 357(6349), 2017. ISSN 0036-8075. doi:10.1126/science.aag0025. URL `https://science.sciencemag.org/content/357/6349/eaag0025`.

[94] Vijay S Pande, Kyle Beauchamp, and Gregory R Bowman. Everything you wanted to know about markov state models but were afraid to ask. *Methods*, 52(1):99–105, 2010.

[95] Ognjen Perišić, Rosana Collepardo-Guevara, and Tamar Schlick. Modeling studies of chromatin fiber structure as a function of dna linker length. *Journal of molecular biology*, 403(5):777–802, 2010.

[96] RA Pethrick. Polymer physics. edited by michael rubinstein and ralph h colby oxford university press, oxford, 2003. isbn 019852059x. pp 440. *Polymer International*, 53(9): 1394–1395, 2004.

[97] Eric F Pettersen, Thomas D Goddard, Conrad C Huang, Gregory S Couch, Daniel M Greenblatt, Elaine C Meng, and Thomas E Ferrin. Ucsf chimera–a visualization system for exploratory research and analysis. *J Comput Chem*, 25(13):1605–1612, October 2004.

[98] Steve Plimpton. Fast parallel algorithms for short-range molecular dynamics. *Journal of Computational Physics*, 117(1):1–19, 1995. ISSN 0021-9991. doi:https://doi.org/10.1006/jcph.1995.1039. URL http://www.sciencedirect.com/science/article/pii/S002199918571039X.

[99] Michael G Poirier, Eugene Oh, Hannah S Tims, and Jonathan Widom. Dynamics and function of compact nucleosome arrays. *Nature structural & molecular biology*, 16(9): 938–944, 2009.

[100] Benjamin D Pope, Tyrone Ryba, Vishnu Dileep, Feng Yue, Weisheng Wu, Olgert Denas, Daniel L Vera, Yanli Wang, R Scott Hansen, Theresa K Canfield, et al. Topologically associating domains are stable units of replication-timing regulation. *Nature*, 515(7527): 402–405, 2014.

[101] Lakshmanane Premkumar, Bruno Segovia-Chumbez, Ramesh Jadi, David R Martinez, Rajendra Raut, Alena Markmann, Caleb Cornaby, Luther Bartelt, Susan Weiss, Yara Park, et al. The receptor binding domain of the viral spike protein is an immunodominant and highly specific target of antibodies in sars-cov-2 patients. *Science immunology*, 5 (48), 2020.

[102] WILLIAM H Prusoff, TS Lin, E MICHAEL August, THOMAS G Wood, and MARIA ELENA Marongiu. Approaches to antiviral drug development. *The Yale journal of biology and medicine*, 62(2):215, 1989.

[103] Alan Pryor, Colin Ophus, and Jianwei Miao. A streaming multi-gpu implementation of image simulation algorithms for scanning transmission electron microscopy. *Advanced Structural and Chemical Imaging*, 3(1):1–14, 2017.

[104] Sergey V Razin and Alexey A Gavrilov. Chromatin without the 30-nm fiber: constrained disorder instead of hierarchical folding. *epigenetics*, 9(5):653–657, 2014.

[105] Maria Aurelia Ricci, Carlo Manzo, María Filomena García-Parajo, Melike Lakadamyali, and Maria Pia Cosma. Chromatin fibers are formed by heterogeneous groups of nucleosomes in vivo. *Cell*, 160(6):1145–1158, 2015.

[106] Daniel R Roe and Thomas E Cheatham III. Ptraj and cpptraj: software for processing and analysis of molecular dynamics trajectory data. *Journal of chemical theory and computation*, 9(7):3084–3095, 2013.

[107] Andrew Routh, Sara Sandin, and Daniela Rhodes. Nucleosome repeat length and linker histone stoichiometry determine chromatin fiber structure. *Proceedings of the National Academy of Sciences of the United States of America*, 105(26):8872–8877, jul 2008. ISSN 00278424. doi:10.1073/pnas.0802336105. URL /pmc/articles/PMC2440727/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC2440727/.

[108] Adrian L Sanborn, Suhas SP Rao, Su-Chen Huang, Neva C Durand, Miriam H Huntley, Andrew I Jewett, Ivan D Bochkov, Dharmaraj Chinnappan, Ashok Cutkosky, Jian

Li, et al. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proceedings of the National Academy of Sciences*, 112(47):E6456–E6465, 2015.

[109] Thomas Schalch, Sylwia Duda, David F Sargent, and Timothy J Richmond. X-ray structure of a tetranucleosome and its implications for the chromatin fibre. *Nature*, 436 (7047):138–141, 2005.

[110] Martin K Scherer, Benjamin Trendelkamp-Schroer, Fabian Paul, Guillermo Pérez-Hernández, Moritz Hoffmann, Nuria Plattner, Christoph Wehmeyer, Jan-Hendrik Prinz, and Frank Noé. Pyemma 2: A software package for estimation, validation, and analysis of markov models. *Journal of chemical theory and computation*, 11(11):5525–5542, 2015.

[111] Eric Schwenker. Image matching for computer vision in atomic-resolution electron microscopy, 03 2020. URL `https://github.com/MaterialEyes/atomagined`.

[112] Takehito Seki, Yuichi Ikuhara, and Naoya Shibata. Theoretical framework of statistical noise in scanning transmission electron microscopy. *Ultramicroscopy*, 193:118–125, 2018. ISSN 0304-3991. doi:https://doi.org/10.1016/j.ultramic.2018.06.014. URL `https://www.sciencedirect.com/science/article/pii/S0304399118300603`.

[113] M Scott Shell. *Thermodynamics and Statistical Mechanics: An integrated approach*. Cambridge University Press, 2015.

[114] Michael Shogren-Knaak and Craig L. Peterson. Switching on Chromatin: Mechanistic Role of Histone H4-K16 Acetylation. *http://dx.doi.org/10.4161/cc.5.13.2891*, 5(13): 1361–1365, jul 2006. ISSN 15514005. doi:10.4161/CC.5.13.2891. URL `https://www.tandfonline.com/doi/abs/10.4161/cc.5.13.2891`.

[115] Micholas Dean Smith and Jeremy C Smith. Repurposing therapeutics for covid-19: Supercomputer-based docking to the sars-cov-2 viral spike protein and viral spike protein-human ace2 interface.

[116] Karl Sohlberg, Timothy J Pennycook, Wu Zhou, and Stephen J Pennycook. Insights into the physical chemistry of materials from advances in haadf-stem. *Physical Chemistry Chemical Physics*, 17(6):3982–4006, 2015.

[117] Chris Solomon and Toby Breckon. *Fundamentals of Digital Image Processing: A practical approach with examples in Matlab*. John Wiley & Sons, 2011.

[118] Feng Song, Ping Chen, Dapeng Sun, Mingzhu Wang, Liping Dong, Dan Liang, Rui-Ming Xu, Ping Zhu, and Guohong Li. Cryo-em study of the chromatin fiber reveals a double helix twisted by tetranucleosomal units. *Science*, 344(6182):376–380, 2014.

[119] René Stehr, Nick Kepper, Karsten Rippe, and Gero Wedemann. The effect of internucleosomal interaction on folding of the chromatin fiber. *Biophysical journal*, 95(8): 3677–3691, 2008.

[120] Jian Sun, Qing Zhang, and Tamar Schlick. Electrostatic mechanism of nucleosomal array folding revealed by computer simulation. *Proceedings of the National Academy of Sciences of the United States of America*, 102(23):8180–8185, jun 2005. ISSN 00278424. doi:10.1073/PNAS.0408867102/SUPPL_FILE/08867FIG8.PDF. URL https://www.pnas.org/doi/abs/10.1073/pnas.0408867102.

[121] Wanbo Tai, Lei He, Xiujuan Zhang, Jing Pu, Denis Voronin, Shibo Jiang, Yusen Zhou, and Lanying Du. Characterization of the receptor-binding domain (rbd) of 2019 novel coronavirus: implication for development of rbd protein as a viral attachment inhibitor and vaccine. *Cellular & molecular immunology*, 17(6):613–620, 2020.

[122] Yoshimasa Takizawa, Cheng-Han Ho, Hiroaki Tachiwana, Hideyuki Matsunami, Wataru Kobayashi, Midori Suzuki, Yasuhiro Arimura, Tetsuya Hori, Tatsuo Fukagawa, Melanie D. Ohi, Matthias Wolf, and Hitoshi Kurumizaka. Cryo-em structures of centromeric tri-nucleosomes containing a central cenp-a nucleosome. *Structure*, 28(1): 44–53.e4, 2020. ISSN 0969-2126. doi:https://doi.org/10.1016/j.str.2019.10.016. URL https://www.sciencedirect.com/science/article/pii/S0969212619303570.

[123] Iva A Tchasovnikarova and Robert E Kingston. Beyond the Histone Code: A Physical Map of Chromatin States. *Molecular Cell*, 69:5–7, 2018. doi:10.1016/j.molcel.2017.12.018. URL https://doi.org/10.1016/j.molcel.2017.12.018.

[124] Philipp Thiel, Markus Kaiser, and Christian Ottmann. Small-molecule stabilization of protein–protein interactions: An underestimated concept in drug discovery? *Angewandte Chemie International Edition*, 51(9):2012–2018, 2012.

[125] David J Tremethick. Higher-order structures of chromatin: the elusive 30 nm fiber. *Cell*, 128(4):651–654, 2007.

[126] Kensal E van Holde. *Chromatin structure and transcription*. Springer, 1989.

[127] Kensal E van Holde. *Chromatin*. Springer Science & Business Media, 2012.

[128] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.

[129] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11 (110):3371–3408, 2010. URL http://jmlr.org/papers/v11/vincent10a.html.

[130] Jiang Wang, Mohit A Gayatri, and Andrew L Ferguson. Mesoscale simulation and machine learning of asphaltene aggregation phase behavior and molecular assembly landscapes. *The Journal of Physical Chemistry B*, 121(18):4923–4944, 2017.

[131] Qihui Wang, Yanfang Zhang, Lili Wu, Sheng Niu, Chunli Song, Zengyuan Zhang, Guangwen Lu, Chengpeng Qiao, Yu Hu, Kwok-Yung Yuen, et al. Structural and functional basis of sars-cov-2 entry by using human ace2. *Cell*, 181(4):894–904, 2020.

[132] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi:10.1109/TIP.2003.819861.

[133] Tobias Warnecke, Erin A. Becker, Marc T. Facciotti, Corey Nislow, and Ben Lehner. Conserved substitution patterns around nucleosome footprints in eukaryotes and archaea derive from frequent nucleosome repositioning through evolution. *PLOS Computational Biology*, 9(11):1–10, 11 2013. doi:10.1371/journal.pcbi.1003373. URL `https://doi.org/10.1371/journal.pcbi.1003373`.

[134] Jukka Westermarck, Johanna Ivaska, and Garry L Corthals. Identification of protein interactions involved in cellular signaling. *Molecular & Cellular Proteomics*, 12(7): 1752–1763, 2013.

[135] Christopher L Woodcock, Arthur I Skoultchi, and Yuhong Fan. Role of linker histone in chromatin structure and function: H1 stoichiometry and nucleosome repeat length. *Chromosome Research*, 14:17–25, 2006.

[136] Daniel Wrapp, Nianshuang Wang, Kizzmekia S Corbett, Jory A Goldsmith, Ching-Lin Hsieh, Olubukola Abiona, Barney S Graham, and Jason S McLellan. Cryo-em structure of the 2019-ncov spike in the prefusion conformation. *Science*, 367(6483):1260–1263, 2020.

[137] Renhong Yan, Yuanyuan Zhang, Yaning Li, Lu Xia, Yingying Guo, and Qiang Zhou. Structural basis for the recognition of sars-cov-2 by full-length human ace2. *Science*, 367(6485):1444–1448, 2020.

[138] Ruihan Zhang, Jochen Erler, and Jörg Langowski. Histone Acetylation Regulates Chromatin Accessibility: Role of H4K16 in Inter-nucleosome Interaction. *Biophysical Journal*, 112(3):450–459, feb 2017. ISSN 0006-3495. doi:10.1016/J.BPJ.2016.11.015.

[139] Peng Zhao, Jeremy L Praissman, Oliver C Grant, Yongfei Cai, Tianshu Xiao, Katelyn E Rosenbalm, Kazuhiro Aoki, Benjamin P Kellman, Robert Bridger, Dan H Barouch, et al. Virus-receptor interactions of glycosylated sars-cov-2 spike and human ace2 receptor. *Cell host & microbe*, 28(4):586–601, 2020.

[140] Peng Zhou, Xing-Lou Yang, Xian-Guang Wang, Ben Hu, Lei Zhang, Wei Zhang, Hao-Rui Si, Yan Zhu, Bei Li, Chao-Lin Huang, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *nature*, 579(7798):270–273, 2020.

[141] Maxim Ziatdinov, Artem Maksov, and Sergei V Kalinin. Learning surface molecular structures via machine vision. *npj Computational Materials*, 3(1):1–9, 2017.