

Electronic Thesis and Dissertation Repository

12-9-2021 12:15 PM

A computational study on a globular protein and an intrinsically disordered protein

Cecilia Chavez Garcia, *The University of Western Ontario*

Supervisor: Karttunen, Mikko, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Chemistry

© Cecilia Chavez Garcia 2021

Follow this and additional works at: <https://ir.lib.uwo.ca/etd>

 Part of the [Other Chemistry Commons](#)

Recommended Citation

Chavez Garcia, Cecilia, "A computational study on a globular protein and an intrinsically disordered protein" (2021). *Electronic Thesis and Dissertation Repository*. 8343.
<https://ir.lib.uwo.ca/etd/8343>

This Dissertation/Thesis is brought to you for free and open access by Scholarship@Western. It has been accepted for inclusion in Electronic Thesis and Dissertation Repository by an authorized administrator of Scholarship@Western. For more information, please contact wlsadmin@uwo.ca.

Abstract

We present molecular dynamics (MD) simulations of two protein targets for drug design: Triosephosphate isomerase (TIM) and Methyl CpG binding protein 2 (MeCP2). First, we studied three TIM proteins: TcTIM, TbTIM and a chimeric protein (Mut1). The first two are homologous enzymes with high sequence similarity, albeit different biophysical parameters. The chimeric protein has TbTIM's sequence and 13 single point mutations, which are sufficient to obtain TcTIM-like behaviour in reactivation experiments. We analyzed the residue interaction networks observed in the all-atom MD simulations, as well as their electrostatic interactions and the impact of simulation length on them. A conserved salt bridge between catalytic residues Lys 14 and Glu 98 was observed in all three proteins, but key differences were found in other interactions concerning the catalytic amino acids. Although TcTIM forms less hydrogen bonds than TbTIM and Mut1, its hydrogen bond network spans almost the entire protein, connecting the residues in both monomers. Some of these interactions appeared only after the first microsecond of the simulation, and convergence in the number of hydrogen bonds was only reached during the last of the 3 μ s of the simulation. Second, we performed MD simulations of the methyl DNA binding domain (MBD), which is the only domain in MeCP2 with an available structure. After characterizing its structure both in solution and in the presence of a surface in order to compare with high-speed atomic force microscopy experiments (HS-AFM), we built the rest of the protein structure by *ab initio* modelling using Modeller. This model was simulated in both all-atom and coarse-grained force fields. Two main conformations were sampled in the coarse-grained simulations: a globular structure similar to the one observed in the all-atom force field and a two-globule conformation. A similar two-globule conformation has been observed in the HS-AFM experiments. Our results are in good agreement with available experimental data. They predicted 4.1% of α -helical content, the experimental result is 4%. Finally, we compared the model predicted by AlphaFold to our Modeller model. Together, these simulations represent the first attempt to characterize the structure and dynamics of the full-length MeCP2 protein.

Keywords

Intrinsically disordered protein, molecular dynamics simulation, coarse-grained, triosephosphate isomerase, Methyl CpG binding protein 2

Summary for Lay Audience

We present molecular dynamics (MD) simulations of two proteins. The aim of an MD simulation is to provide the time-evolution of a system by solving iteratively its equations of motion. We first studied two Triosephosphate isomerase (TIM) proteins, one from *Trypanosoma cruzi* (TcTIM), the parasite that causes Chagas' disease, and one from *Trypanosoma brucei* (TbTIM), causative agent of the African sleeping sickness, as well as a chimeric protein with some characteristics of both of them. Our simulations allowed us to study the electrostatic interactions between these proteins and explain why they behave differently even though they are extremely similar. Next, we focused our study on the Methyl CpG binding protein 2 (MeCP2). This protein is essential for growth and synaptic activity of neurons. Its malfunction is associated to Rett syndrome, the most common cause of cognitive impairment in females. This protein is an intrinsically disordered protein (IDP), a type of protein which does not have a unique tertiary structure. IDPs are highly flexible and conventional methods to study proteins are often not directly applicable to them. This is why the full-length structure of MeCP2 has not been solved yet. The only available structure solely contains ~17% of its amino acids, which represents the most ordered domain of this protein. We first performed MD simulations on this structure, and then used ab initio modelling to complete the rest of the protein. Since all-atom simulations of this model were not enough to guarantee adequate sampling of its conformational space, coarse-grained modeling was used to complement the atomistic picture. The coarse-grained simulations sampled a conformation that had not been observed in the all-atom simulations but that was in good agreement with a conformation previously observed in experimental data. Furthermore, our simulations predicted an α -helical content of 4.1% (experimental value: 4%). Together, our simulations represent the first effort to characterize the structure and dynamics of the full-length MeCP2 protein.

Co-Authorship Statement

The works presented in chapters 4 and 6 in this thesis represent published, or to be published, first author publications. In these works, my supervisor Dr. Karttunen appears as author. While I have been the primary planner, performer, analyzer, writer and editor of each of them, he has assisted in all of these aspects as well. He has also provided financial resources.

In chapter 5, I planned, performed, analyzed and wrote everything that pertained computational simulations. My supervisor Dr. Karttunen also participated in each of these aspects as well.

In chapter 6, Dr. Jérôme Hénin contributed to planning and manuscript editing.

Besides the work presented in this thesis, the following articles were published during the course of my PhD:

1. Chávez-García, C.; Aguayo-Ortiz, R.; Dominguez, L. Quantifying Correlations between Mutational Sites in the Catalytic Subunit of γ -Secretase. *J. Mol. Graph. Model.* 2019, 88, 221–227. <https://doi.org/10.1016/j.jmgm.2019.02.002>.
2. Rohoullah, F.; Sowlati-Hashjin, S.; Chávez-García, C.; Mitra, A., Hossein Karimi-Jafarif, M.; Karttunen, M. Identification of Catechins Binding Pockets in Monomeric A β 42 Through Ensemble Docking and MD Simulations. *To be submitted*

Acknowledgments

I couldn't have asked for better guidance or support from my amazing supervisor, Dr. Mikko Karttunen. He encouraged me, gave me the opportunity to exchange ideas with people in other countries, and helped me keep on track during the difficult months of the pandemic. Special thanks to Dr. Jérôme Hénin, for hosting me during a summer and providing valuable input in the MeCP2 project.

I would also like to acknowledge the invaluable financial support from the province of Ontario Trillium Scholarship Program, as well as from the department of Chemistry at Western University. SHARCNET and Compute Canada provided most of the computational resources.

On a personal note, I would like to thank all our group members, especially Ali and David, with whom I shared this journey. In addition, thank you to Botsing, Maricarmen, Los Juanos and my Lehua group just for being there. Last, but certainly not least, I would like to thank César and my parents for their unconditional support.

Table of Contents

Abstract.....	ii
Summary for Lay Audience.....	iv
Co-Authorship Statement.....	v
Acknowledgments.....	vi
Table of Contents.....	vii
List of Tables.....	x
List of Figures.....	xi
1 Introduction.....	1
1.1 Proteins.....	1
1.2 Protein structure and function.....	8
1.3 Intrinsically disordered proteins.....	9
1.4 Triosephosphate isomerase - TIM.....	12
1.5 Methyl CpG binding protein 2 - MeCP2.....	14
1.6 References.....	15
2 Molecular dynamics simulations.....	26
2.1 Introduction.....	26
2.2 Integration algorithms.....	30
2.3 Thermostats and Barostats.....	32
2.4 Force fields.....	34
2.5 Time scales.....	36
2.6 Coarse-graining methods.....	38
2.7 References.....	41
3 About this thesis.....	51
3.1 Significance and aims.....	51

3.2	Thesis outline.....	52
3.3	References	52
4	Highly similar sequence and structure yet different biophysical behaviour: A computational study of two triosephosphate isomerases	57
4.1	Abstract	58
4.2	Introduction	58
4.3	Materials and methods	61
4.4	Results	62
4.5	Discussion.....	70
4.6	Conclusions	75
4.7	Supplemental information	76
4.8	Acknowledgments.....	91
4.9	References	91
5	Structure and dynamics of the Rett syndrome protein, MeCP2	101
5.1	Abstract	102
5.2	Introduction	102
5.3	Materials and methods	104
5.4	Results	111
5.5	Discussion.....	120
5.6	Supplemental information	123
5.7	Acknowledgments.....	135
5.8	References	135
6	A multiscale computational study of the conformation of the full-length intrinsically disordered protein MeCP2.....	142
6.1	Abstract	143
6.2	Introduction	143
6.3	Materials and methods	145

6.4 Results	148
6.5 Conclusions	166
6.6 Supplemental information	168
6.7 Acknowledgments.....	183
6.8 References	183
7 Conclusions and future directions.....	192
7.1 Conclusions	192
7.1.1 Conclusions.....	192
7.1.2 All-atom MD simulations can be enough to provide insight into globular proteins	192
7.1.3 MD simulations can be a complimentary technique to experimental procedures.....	192
7.1.4 MD simulations can provide new insights when it's difficult to obtain experimental data	193
7.2 Future directions	193
7.3 References	194
Curriculum Vitae	195

List of Tables

Table 4.1 Summary of interactions defined in RIP-MD.....	65
Table S5.1 Decay rates of auto-correlation functions	134
Table S6.1 Details of all-atom simulations.....	169
Table S6.2 Details of coarse-grained simulations	170
Table S6.3 Secondary structure content in the last 400 ns of the all-atom MeCP2_1 simulation.....	171
Table S6.4 Relative solvent accessible surface area (rSA) of residue W104	175
Table S6.5 Relative solvent accessible surface area (rSA) of residue R111	175
Table S6.6 Relative solvent accessible surface area (rSA) of residue R133	176
Table S6.7 Salt bridges in the last 400 ns of the full-length all-atom MeCP2_1 simulation	176
Table S6.8 Clusters sampled in the coarse-grained simulations	179
Table S6.9 Templates used to generate models MeCP2_2 and MeCP2_3.....	180
Table S6.10 Conformations sampled by the ID+TRD domains simulations.....	181

List of Figures

Figure 1.1 Protein denaturation.....	2
Figure 1.2 Structural hierarchy in proteins	3
Figure 1.3 Proteins are built by amino acids.....	4
Figure 1.4 The genetic code specifies 21 amino acids	6
Figure 1.5 The most common secondary structures in proteins.....	8
Figure 1.6 IDPs lack a well-defined secondary or tertiary structure.....	10
Figure 1.7 The TIM barrel is an eightfold repeat of ($\beta\alpha$) units.....	13
Figure 1.8 Three-dimensional structure of the MBD domain.....	14
Figure 2.1 Schematic view of force field interactions.....	27
Figure 2.2 Range of timescale for atomistic MD simulations	37
Figure 2.3 All-atom vs coarse-grained energy landscape.....	38
Figure 2.4 Mapping between the chemical structure and the coarse grained model.....	40
Figure 4.1 Alignment of TcTIM and TbTIM sequences	60
Figure 4.2 Root-mean-square deviation with respect to the crystal structure.....	62
Figure 4.3 Root mean square fluctuations for the last microsecond	63
Figure 4.4 Number of contacts between monomers.....	64
Figure 4.5 Minimum distance between loops 6 and 7 in TcTIM.....	65
Figure 4.6 Cation- π interaction between Tyr 103 in monomer A and Arg 99 in monomer B in TcTIM.....	67
Figure 4.7 Salt bridges in TcTIM.....	68

Figure 4.8 Total number of intramolecular hydrogen bonds in each simulation	69
Figure 4.9 Changes over time in the RMSF of the residues of TcTIM.....	72
Figure 4.10 Number of hydrogen bonds in TcTIM.....	73
Figure 4.11 Changes in the hydrogen bond network of TcTIM	74
Figure 4.12 π - π interactions in TcTIM	75
Figure S4.1 Minimum distance between loops 6 and 7 in TbTIM and Mut1.....	76
Figure S4.2 Cation- π interactions for TcTIM.....	77
Figure S4.3 π - π interactions for TcTIM, TbTIM and Mut1	77
Figure S4.4 Salt bridges for TcTIM, TbTIM and Mut1	78
Figure S4.5 Main hydrogen bond network in TcTIM	79
Figure S4.6 Hydrogen bonds in TcTIM	80
Figure S4.7 Hydrogen bonds in TcTIM	81
Figure S4.8 Hydrogen bonds in monomer B throughout the last 2 μ s of the TcTIM simulation	82
Figure S4.9 Hydrogen bonds involving amino acids at the interface between monomers in Mut1 and TbTIM.....	82
Figure S4.10 Hydrogen bonds in Mut1	83
Figure S4.11 Hydrogen bonds in monomer A throughout the last 2 μ s of the Mut1 simulation	84
Figure S4.12 Hydrogen bonds in monomer B throughout the last 2 μ s of the Mut1 simulation	84
Figure S4.13 Hydrogen bonds in TbTIM throughout the last 2 μ s of the simulation	85

Figure S4.14 Hydrogen bonds in monomer A throughout the last 2 μ s of the TbTIM simulation.....	86
Figure S4.15 Hydrogen bonds in monomer B throughout the last 2 μ s of the TbTIM simulation.....	87
Figure S4.16 Main hydrogen bond networks throughout the last 2 μ s for TcTIM and TbTIM	87
Figure S4.17 Hydrogen bonds in Cys 14/15	88
Figure S4.18 Changes over time in the RMSF of the TbTIM simulation	89
Figure S4.19 Changes over time in the RMSF of the Mut1 simulation	89
Figure S4.20 Main hydrogen bond networks for TcTIM in the most populated cluster of the first 500 ns of the simulation.....	90
Figure S4.21 Hydrogen bond networks for TcTIM in the most populated cluster of the last 500 ns of the simulation.....	91
Figure 5.1 Structural features of WT MeCP2	113
Figure 5.2 Domain identification and dynamic conformational changes of MBD in WT MeCP2	114
Figure 5.3 MBD–CTD interactions in cis and trans.....	116
Figure 5.4 Structural features of MBD and CTD in MeCP2 bearing RTT point mutations in MBD	119
Figure S5.1 Domain diagrams of wild type MeCP2 and its mutants	129
Figure S5.2 Structural features of MeCP2 fused to GFT at the C-terminus of the former ...	130
Figure S5.3 Structural features of TRD–CTD	131
Figure S5.4 Structural features of d311–328 and d370–415 mutants.....	132

Figure S5.5 Analysis of ultracentrifugation data.....	133
Figure 6.1 MeCP2 is composed of six domains.....	144
Figure 6.2 Radius of gyration of the full-length MeCP2.....	149
Figure 6.3 All-atom MD simulation of the full-length MeCP2	150
Figure 6.4 Percentage of frames in the last 400 ns of the MBD domain with every type of secondary structure.....	151
Figure 6.5 Coarse-grained simulations of MeCP2 using the PLUM model.....	153
Figure 6.6 RMSD from the initial structure of the protein in the CG1 PLUM simulation. ..	154
Figure 6.7 Minimum distance between the two globules in the CG1 PLUM simulation	155
Figure 6.8 RMSD from the initial structure of the CG2 and CG3 PLUM simulations.....	156
Figure 6.9 RMSD from the initial structure and minimum distance between the two globules of simulation CG4	157
Figure 6.10 Principal Component Analysis of the coarse-grained trajectories.....	158
Figure 6.11 RMSD to the initial structure of models MeCP2_2 and MeCP2_3	159
Figure 6.12 α -helical content of the protein structure vs radius of gyration in the all-atom and two-globule all-atom simulations	161
Figure 6.13 Acylindricity and asphericity vs radius of gyration in the all-atom, coarse-grained and two-globule all-atom simulations	162
Figure 6.14 Radius of gyration vs end-to-end distance of five all-atoms simulations of the ID and TRD domains.....	163
Figure 6.15 Radius of gyration vs end-to-end distance in all-atom simulations of the two connecting loops found in the coarse-grained simulations	164
Figure 6.16 Comparison of the models generated by Modeller and AlphaFold.....	165

Figure S6.1 Alignment of ten of the rejected models of the full-length MeCP2 protein.	168
Figure S6.2 RMSD of the TRD domain in the all-atom full-length protein simulation.....	171
Figure S6.3 Percentage of frames in the last 400 ns of the NTD domain with every type of secondary structure.....	172
Figure S6.4 Percentage of frames in the last 400 ns of the ID domain with every type of secondary structure.....	173
Figure S6.5 Percentage of frames in the last 400 ns of the TRD domain with every type of secondary structure.....	173
Figure S6.6 Percentage of frames in the last 400 ns of the CTD α domain with every type of secondary structure.....	173
Figure S6.7 Percentage of frames in the last 400 ns of the CTD β domain with every type of secondary structure.....	174
Figure S6.8 Principal Component Analysis. Projection of the MeCP2_1 simulation.....	174
Figure S6.9 Salt bridge interaction between residues Lys 119 and Asp 121 and between residues Arg 133 and Glu 137.....	178
Figure S6.10 RMSD of the three AlphaFold simulations.....	181
Figure S6.11 Prediction of protein disorder for MeCP2.....	182
Figure S6.12 Autocorrelation of the radius of gyration.....	182
Figure S6.13 Distribution of structures sampled in the MeCP2_1 simulation	183

1 Introduction

This thesis focuses on the study of a globular protein (triosephosphate isomerase) and an intrinsically disordered protein (Methyl CpG binding protein 2). Molecular dynamics simulations were used to characterize their structure and dynamics. In this chapter, a brief introduction to proteins is made. Sections 1.4 and 1.5 briefly discuss the relevance of these two proteins.

1.1 Proteins

Cells are rich in highly complex molecules termed macromolecules, and proteins are the most abundant of them¹. In fact, biochemical methods for protein detection suggest that each human cell may express up to 15,000 distinct proteins². They are functionally diverse and involved in virtually all life processes in biological organisms, including the catalysis of metabolic processes, energy transfer, gene expression, transport of solutes across membranes, cellular communication, molecular recognition, defense mechanisms and forming intracellular and extracellular structures¹. Given the numerous roles of proteins, the overall health of an organism depends on their normal function, and any significant loss of it may lead to the development of a pathological process. Consequently, proteins constitute ~80% of current pharmaceutical targets¹.

Proteins have evolved to perform their function in a specific cellular environment. They have therefore adapted to its biophysical characteristics, including temperature, pH, salinity and pressure. Proteins denature at both high (typically ~60 °C) and low (typically ~-20°C) temperatures³, and their pH-optimum of activity corresponds to their pH-optimum of stability⁴. The term “denaturation” refers to the phenomenon of loss of the three-dimensional structure a protein has under physiological conditions, by either heat or cooling (Figure 1.1)⁵. There is an ongoing debate between two opposing views that explain cold denaturation: hydrophobic hydration and hydrophilic hydration. Each theory claims that the dominant energetic contribution to cold denaturation comes from one of these two types of residues, and the problem remains an open question in the scientific community^{3,6}.

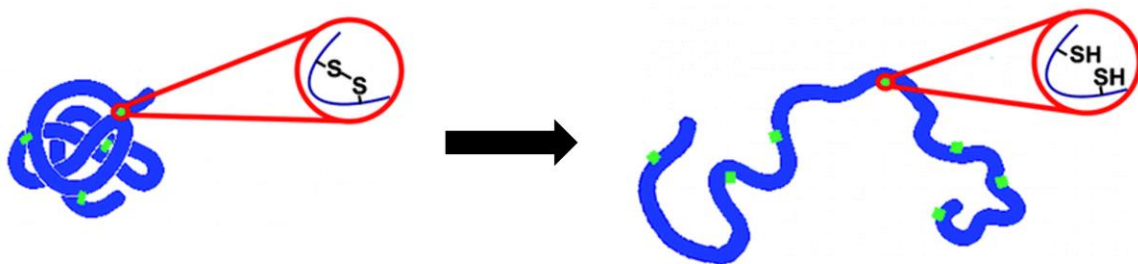


Figure 1.1 Protein denaturation. The protein loses the three-dimensional structure it has under physiological conditions. Green: stabilizing interactions, such as disulfide bridges, hydrogen bonding and ionic bonds. Adapted with permission from Killian *et al*¹. Copyright 2021 American Chemical Society.

The structural organization of proteins is commonly described in terms of four different aspects of covalent structure and folding patterns¹. The levels of this hierarchy are known as primary, secondary, tertiary and quaternary structure (Figure 1.2). The primary structure is the ordered sequence of amino acids composing the protein chain. The secondary structure refers to the initial folding of the sequence into helices and sheets. The overall chain folds further into a three-dimensional compact tertiary structure, which constitutes the third level of the hierarchy. All proteins have these three levels of structural hierarchy, but there are some proteins that include more than one chain. In such cases, the spatial arrangement of the different subunits constitutes its quaternary structure¹.

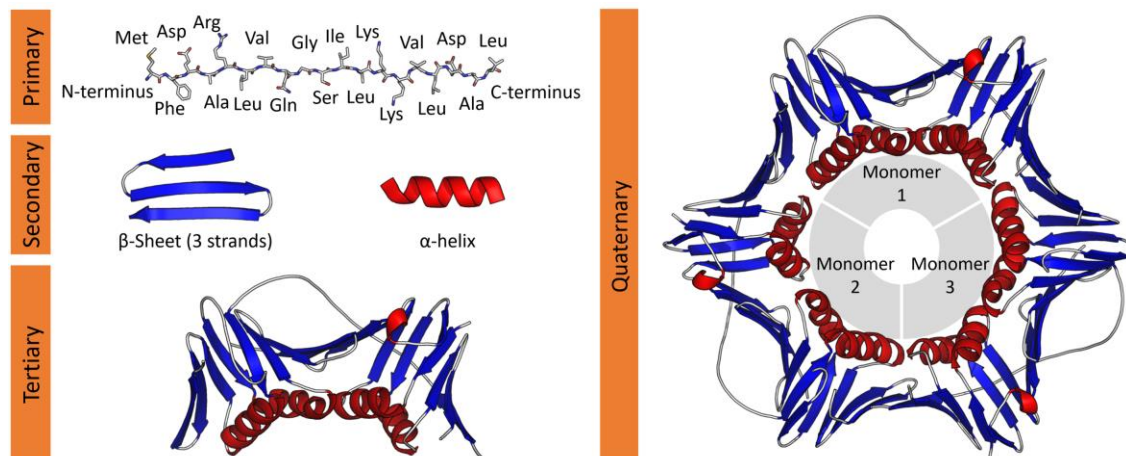


Figure 1.2 Structural hierarchy in proteins. The primary structure is the amino acid sequence of the protein chain, the secondary structure is the initial folding of the chain into helices and sheets and the tertiary structure is the three-dimensional structure of the entire chain. If the protein includes more than one chain, the quaternary structure constitutes the spatial arrangement of the different subunits. From Wikipedia, under a CC BY 4.0 license⁸.

The amino acid chain is the primary and central component of the protein. It is formed by linking amino acids via peptide bonds. A peptide bond forms when the carboxyl group of one amino acid condenses with the amino group of another amino acid to eliminate water. The succession of peptide bonds generates a backbone, from which the side chains are projected. All amino acids have in common a central carbon atom ($C\alpha$) to which a hydrogen atom, an amino group (NH_2) and a carboxyl group ($COOH$) are attached. They are distinguished by the side chain attached to the $C\alpha$ through its fourth valence (Figure 1.3). There are 21 amino acids specified by the genetic code but a few others occur in rare cases by post-translational modifications. Of those, nine are termed “essential” amino acids since humans and other vertebrates cannot synthesise them from metabolic intermediates⁹.

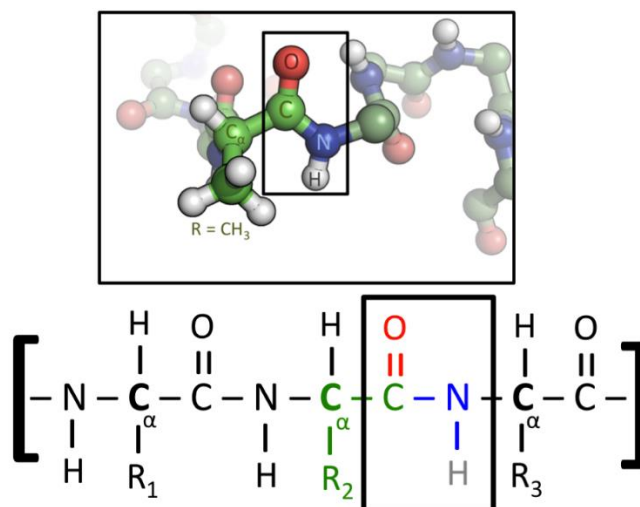


Figure 1.3 Proteins are built by amino acids linked via a peptide bond, generating a backbone. All amino acids have in common a central carbon atom ($C\alpha$) to which a hydrogen atom, an amino group (NH_2) and a carboxyl group (COOH) are attached. They are distinguished by the side chain (R_i) attached to the $C\alpha$ through its fourth valence. From Wikipedia, under a CC BY-SA 3.0 license¹⁰.

Amino acids are divided into three different classes (plus some special cases) according to the chemical nature of their side chain: hydrophobic, polar or charged (Figure 1.4). There are two amino acids with acidic side chains: aspartic and glutamic acid. At neutral pH they are fully ionized, containing a negatively charged carboxylate group ($-\text{COO}^-$). Three amino acids have basic side chains. Two of them are fully ionized and positively charged at neutral pH: lysine and arginine. Histidine is only weakly basic and can be either positively charged or neutral, depending on the ionic environment provided by the nearby residues in the protein. Polar amino acids have zero net charge at neutral pH and contain at least one atom with electron pairs available for hydrogen bonding to water. Hydrophobic amino acids have a side chain that does not bind or give off protons. They do not participate in hydrogen or ionic bonds and instead promote hydrophobic interactions¹¹. Four amino acids are considered to be special cases: 1) Cysteines can form a disulfide bond ($-\text{S}-\text{S}-$) when the sulfhydryl group ($-\text{SH}$) in two of them becomes oxidized to form a covalent cross-link. Disulfide bridges stabilize the folding of proteins, making them less susceptible to degradation⁹. 2) Selenocysteine requires an elaborate synthetic and translational apparatus

that does not resemble the canonical enzymatic system employed for the rest of the amino acids¹². Finally, 3) glycine and proline are both special but in opposite ways. Glycine contains a single hydrogen as its side chain (Figure 1.4) and it therefore has a huge conformational flexibility. In contrast, proline is geometrically limited due to the fusion of its backbone and sidechain. This fusion prevents the N atom from participating in hydrogen-bonding and also provides some steric hindrance to the α -helical conformation¹.

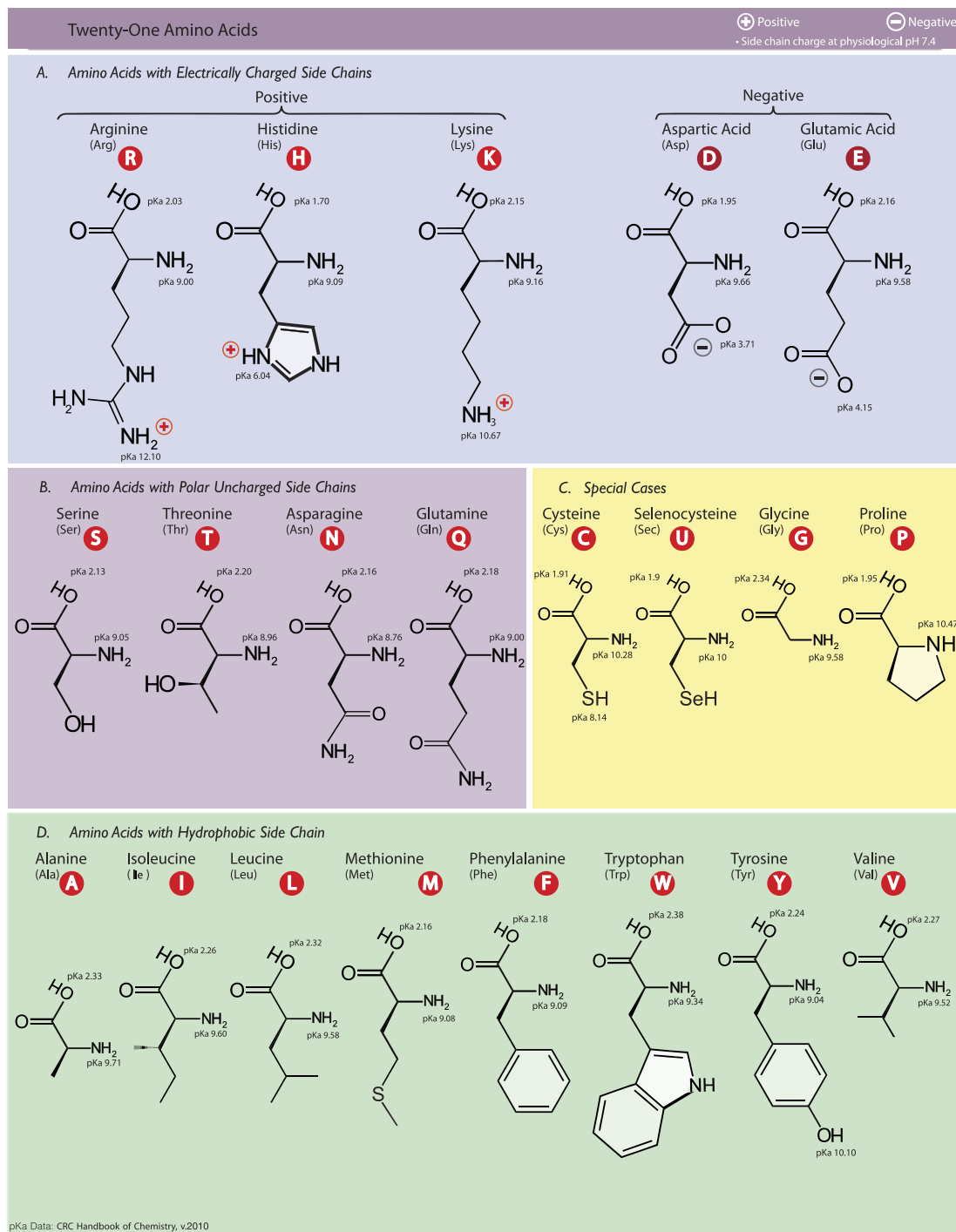


Figure 1.4 The genetic code specifies 21 amino acids which are divided into four groups according to the chemical nature of their side chain. From Wikipedia, under a CC BY-SA 3.0 license¹³.

The main driving force for the folding of globular proteins is the hydrophobic effect^{14,15}. It plays an important role in the stability of biomolecules and is associated to cold denaturation³. It is comprised by two energetic components: one enthalpic and one entropic. The enthalpic hydrophobic effect is associated with the expulsion of disordered water from hydrophobic regions. The entropic component is the result of an increase in water disorder when hydrophobic surfaces aggregate and lessen the surface area around which water molecules are more aligned¹⁶. The hydrophobic effect results in the burial of the hydrophobic side chains in the core of the protein, creating a hydrophilic surface. The hydrophobic core is densely packed and in the few cases where space remains, one or more water molecules will hydrogen-bond to internal polar groups⁹. These are firmly bound and can be regarded as integral parts of the protein structure. In order to bring the hydrophobic side chains into the core, the main chain must also fold into the interior of the protein. It is, however, highly polar, with one hydrogen bond donor (NH) and one acceptor (C'=O) for each peptide unit. These polar groups are neutralized by the formation of hydrogen bonds, giving rise to the secondary structure of proteins: α -helices and β -sheets (Figure 1.5)⁹.

The secondary structure provides a solid framework to the protein. It is relatively rigid and is therefore the best-defined part of a protein structure when it is determined by both X-ray and nuclear magnetic resonance (NMR) techniques.

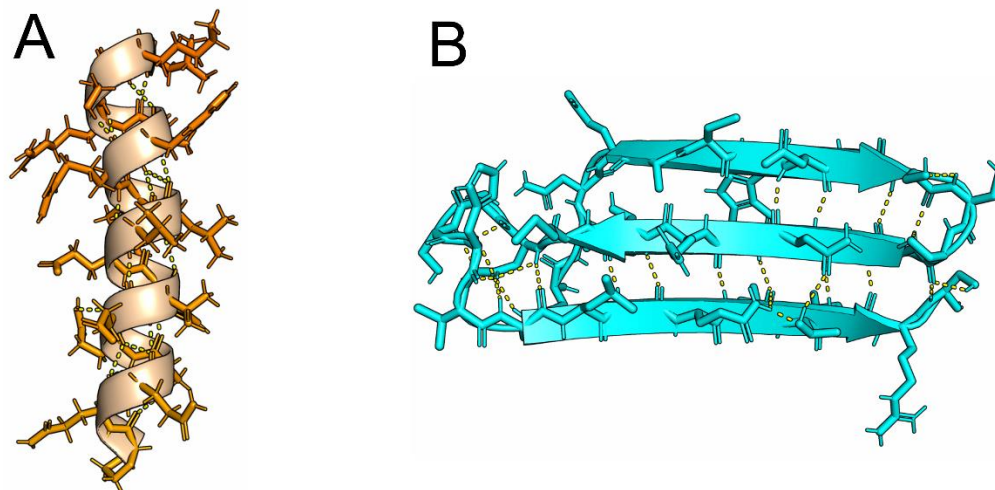


Figure 1.5 The most common secondary structures in proteins are the (A) α -helix and (B) β -sheet. They are stabilized by the formation of hydrogen bonds between the side chains of the amino acids that form them.

Motifs are combinations of secondary structure elements with a specific geometric arrangement that are frequently found in protein structures. Some of them can be associated with a specific function such as DNA binding, but others have no specific biological function on their own. Several motifs combine into domains, which is the fundamental unit of tertiary structure⁹. A domain can fold independently into a stable tertiary structure and is structurally independent of the other domains. Polypeptide chains that are more than 200 amino acids long generally consist of two or more domains¹¹. The process by which a polypeptide chain acquires its correct three-dimensional structure and reaches the biologically active native state is called protein folding⁹.

1.2 Protein structure and function

Initial efforts to crystallize proteins focused on hemoglobin, and although the first photographs of hemoglobin crystals date from 1909, it took another 50 years before the three-dimensional structure of this protein was solved¹⁷. At least 42 scientists have received Nobel Prizes in Physics, Chemistry or Medicine for contributions that included the use of X-rays or neutrons and crystallography. Recording X-ray diffraction images of

macromolecular crystals turned out to be very challenging because of how easy they deteriorate, they are sensitive to over-drying when exposed to air, and are temperature sensitive¹⁷.

The thermodynamic hypothesis of protein folding, also known as “Anfinsen’s dogma”, is a theoretic milestone. It states that the native structure of a protein represents a free energy minimum determined by the totality of interatomic interactions and hence by the amino acid sequence⁵. However, how the correct folding of a protein is selected from the astronomically large number (10^{47}) of possible conformations to give the native state in a timescale of seconds or less, remained a paradox, known as Levinthal’s paradox, for a long time¹⁸. It is now clear that folding pathways drive the protein efficiently towards a topology close to that of the native state¹⁹. Moreover, proteins actually assume a large number of nearly isoenergetic conformations and its motion can be discussed in terms of energy landscapes, which describe the potential energy of the protein as a function of conformational coordinates²⁰.

For a long time, it was assumed that all proteins have a well-defined and stable three-dimensional structure that is fully determined by the amino acid sequence. Early experiments showed that proteins lost their function upon losing their structure and thus it was believed that the native and functional state of a protein was necessarily a stable structure. Results that ran counter this assumption were considered to be mistakes due to either the experimental setup or the experimenter²¹. As experimental evidence of disordered proteins accumulated, it became impossible to ignore them, but they were often considered as being functionally irrelevant. Many terms have been used to describe these proteins, and it was not until 2005 that the field started to use consistent terminology and the term “intrinsically disordered” became predominant²¹.

1.3 Intrinsically disordered proteins

Intrinsically disordered proteins (IDPs) are a class of proteins that contain extensive disorder, either local or global, that is important for function^{22,23}. In contrast to globular

proteins, IDPs do not have a well-defined secondary or tertiary structure and can adopt a wide range of configurations (Figure 1.6). The structure-function paradigm, supported by numerous reports of structures determined using X-ray crystallography and NMR spectroscopy, slowed the acceptance of the biological role of highly dynamic and disordered protein states. Additionally, the lack of common terminology precluded the appearance of the idea that this class of proteins constitutes a separate and important category of proteins²¹.

IDPs play a key role in signalling and regulatory functions, including the regulation of transcription, translation and the cell cycle²³. Their inherent flexibility enables them to interact promiscuously with different targets on different occasions, they offer accessible sites for post-translational modification and their extremely fast association rates allow signals to rapidly turn on²³. Studies have shown that about 10-35% of prokaryotic and 15-45% of eukaryotic proteins contain disordered regions of at least 30 amino acids in length^{24,25}.

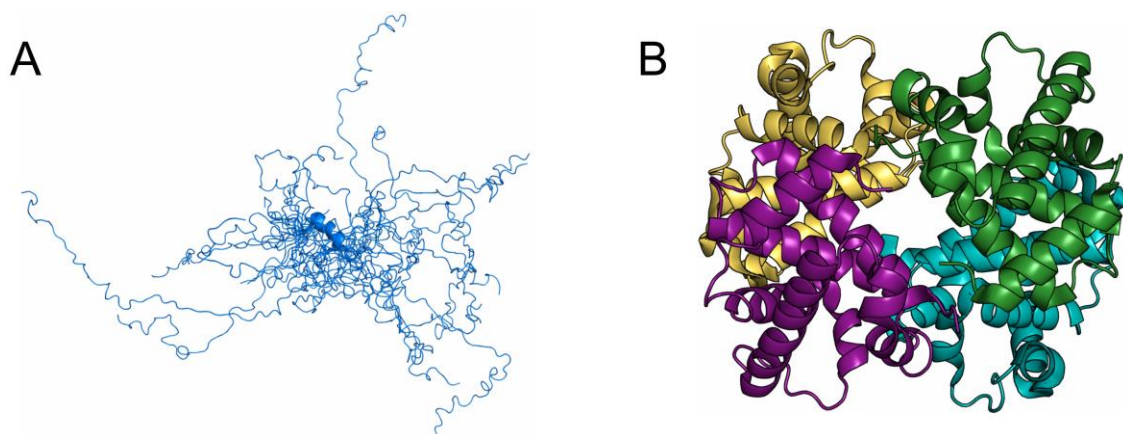


Figure 1.6 In contrast with globular proteins, IDPs lack a well-defined secondary or tertiary structure. (A) A spinach thylakoid soluble phosphoprotein (PDBid: 2FFT), an IDP and (B) human hemoglobin (PDBid: 1SI4), a globular protein.

The structural characterization of IDPs faces many challenges, they are extremely difficult to prepare intact and often degrade during purification²³, and since they do not have a single well-defined structure, crystal-structure analysis can only indicate their presence through the absence of electron density. Solution state NMR, residual dipolar couplings (RDCs),

small angle X-ray scattering (SAXS) and paramagnetic relaxation enhancement (PRE) are some of the tools that can provide detailed information on residual secondary structure, transient long-range contacts and dynamics of disordered proteins, nevertheless, describing their ensemble of conformations remains a challenge^{26,27}.

One way to complement experimental studies is to use computer simulations and statistical thermodynamics as tools for atomic-level characterizations and thermodynamic descriptions. Since 1994 the Critical Assessment of Structure Prediction (CASP), a large-scale community experiment, has been held every two years²⁸. CASP provides an avenue for objective testing and assessment of protein structure modeling methods. In CASP14, the neural network-based model AlphaFold²⁹, demonstrated accuracy competitive with experimental structures in a majority of cases. However, the regions with very low confidence in the predictions overlap with intrinsically disordered regions³⁰.

Achieving an accurate characterization of IDPs via simulations is also challenging, because they rely on the accuracy of the force field. Protein force fields were developed to target globular proteins and their applicability to IDPs is not straightforward³¹. In fact, studies have shown that prediction of native-state structures and folding rates appear to be more robust than the detailed kinetics and the properties of unfolded states, which share some characteristics with disordered proteins³²⁻³⁴. A community-based blind test based on CASP, the Critical Assessment of protein Intrinsic Disorder prediction (CAID) experiment, was established in 2020. With the objective to determine the state-of-the-art in prediction of intrinsically disordered regions, the experiment evaluated 43 methods on a dataset of 646 proteins from DisProt³⁵. Interestingly, the best methods used deep learning techniques and notably outperformed physicochemical methods³⁶.

Force fields have two major shortcomings in describing IDP structures. The first one is that they present variations in their structural propensities, as force fields tend to overpopulate either α -helical or β -sheet structures^{34,37}. This problem has been addressed via explicit optimization of backbone torsion parameters against NMR data³⁸. These improvements can be found in the latest published force fields, including CHARMM22*³², CHARMM36³⁹, Amber ff03*⁴⁰ and Amber ff99SB*-ILDNP⁴¹ amongst others. The second problem is the

prediction of structures that are too compact. This issue has been addressed by either strengthening protein-water interactions (Amber ff03ws force field⁴²), or by using a water model with an increased Lennard-Jones well-depth (TIP4P-D water model⁴³). In spite of these advances, studies that compare different force fields for protein folding and to sample the structural ensembles of IDPs have found large differences across force fields^{32,34,44,45}. Although the accuracy of force fields for IDP simulations is not well-characterized, improving the existing force fields is an ongoing effort in the scientific community⁴⁶⁻⁴⁸.

IDPs have relatively flat energy landscapes and consequently, extensive simulations are needed to ensure that the conformational space has been adequately sampled.

1.4 Triosephosphate isomerase - TIM

The protein described in this section was studied in this work.

Triosephosphate isomerase (TIM) is an enzyme that catalyzes the interconversion of dihydroxyacetone phosphate (DHAP) into D-glyceraldehyde 3-phosphate (GAP), an essential step in the glycolytic pathway⁴⁹. TIM is considered a “perfect” catalyst because the rate of the overall reaction is diffusion controlled⁵⁰. Its first crystal structure revealed for the first time the TIM barrel topology, an eightfold repeat of ($\beta\alpha$) units in such a way that β -strands in the inside are surrounded by α -helices on the outside (Figure 1.7)⁵¹. This is now one of the most common structural motifs in proteins, it is present in ~10% of all known proteins and is the most common enzyme fold in the Protein Data Bank (PDB) database⁵¹⁻⁵³.

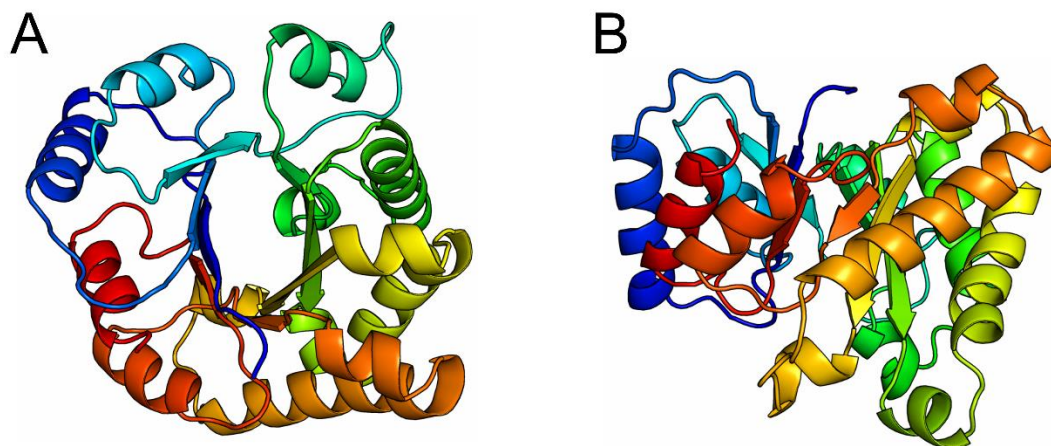


Figure 1.7 The TIM barrel is an eightfold repeat of ($\beta\alpha$) units in such a way that β -strands in the inside are surrounded by α -helices on the outside. Schematic view of the TIM barrel from (A) the top and (B) its side. Colored by region.

The four catalytic residues are strictly conserved throughout the TIM family⁵². They are located in three of the eight $\beta\alpha$ -loops, being loop 1 (N11, K13), loop 3 (H95) and loop 6 (E167)⁵¹. This last loop is highly flexible and it moves from open to close, sampling multiple conformational states⁵⁴. TIM is completely active only in the dimeric form, even though the catalytic residues of each active site are provided by the same subunit. Regions 1, 4 and 8 become more rigid when the dimer is formed, which in turn rigidifies the two separate active sites, providing full catalytic power^{51,55}.

TIM activity is of critical importance for the proper functioning of cells and it is essential for maintaining life under anaerobic conditions. Consequently, it has been used as a target for drug design when dealing with human parasites⁵⁶⁻⁵⁸. In particular, the TIM proteins of *Trypanosoma cruzi* (TcTIM), the parasite that causes Chagas' disease, and *Trypanosoma brucei* (TbTIM), causative agent of the African sleeping sickness, have been the object of many studies⁵⁹⁻⁶². These two homologous enzymes have high similarity yet significant differences in their biophysical parameters. A study by Bolaños *et al.*⁶³ showed that it is sufficient to mutate 13 amino acids on TbTIM to obtain TcTIM-like behaviour in reactivation experiments. Circular dichroism indicated that the chimeric proteins had a

TIM fold, however, the role that these mutations have on the structure and dynamics of the proteins is not well understood.

1.5 Methyl CpG binding protein 2 - MeCP2

The protein described in this section was studied in this work.

Methyl CpG binding protein 2 (MeCP2) is a transcriptional regulator essential for growth and synaptic activity of neurons⁶⁴. The malfunction of this protein is associated to the Rett syndrome, one of the most common causes of mental retardation in females^{65,66}. This X-linked neurologic disorder often causes death in infancy or severe neonatal encephalopathy in males⁶⁷.

MeCP2 is composed of six different domains: the N-terminal domain (NTD), the methyl-CpG binding domain (MBD), the intervening domain (ID), the transcriptional repression domain (TRD) and the C-terminal domain (CTD), which is subdivided into CTD- α and CTD- β ⁶⁸. MeCP2 is an IDP and its physical characteristics make its structural characterization a challenge. Out of the six domains, only the MBD domain has structural information available, and it only accounts for $\sim 17\%$ of the 486 amino acids in the protein (Figure 1.8)⁶⁹.

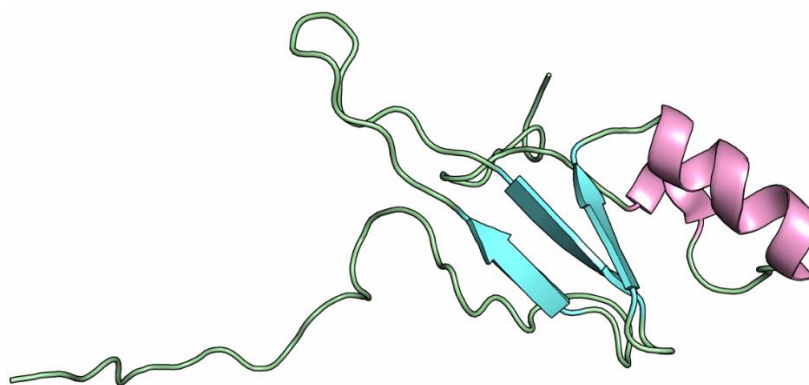


Figure 1.8 Three-dimensional structure of the MBD domain, the only ordered domain in MeCP2 and with a known structure. PDBid: 1QK9⁶⁹. Colored by secondary structure: α -helix (pink), β -sheet (cyan) and random coil (green).

Circular dichroism (CD) of recombinant human MeCP2 showed that the protein consists of ~35% β -strand/turn, 5% α -helix and almost 60% is unstructured⁷⁰. Further CD studies of isolated NTD, ID, TRD and CTD domains confirmed their lack of stable secondary structure⁷¹. Additionally, it was experimentally demonstrated that the NTD, CTD and TRD domains can undergo a coil to helix transition, with the TRD showing the greatest tendency for helix formation⁷¹.

Due to the lack of a three-dimensional structure of the full-length protein, only two computational studies have been reported, and they both focus on the MBD domain^{72,73}. A better understanding of the tertiary structure of MeCP2 is needed in order to discern the molecular links between MeCP2 domain organization, the multifunctionality of the protein, and the cellular pathogenesis of the Rett syndrome.

1.6 References

- (1) Kessel, A.; Ben-Tal, N. *Introduction to Proteins: Structure, Function and Motion*; CRC Press: Boca Raton, USA, 2010.
- (2) Patterson, S. How Much of the Proteome Do We See with Discovery-Based Proteomics Methods and How Much Do We Need to See? *Curr. Proteomics* **2006**, *1* (1), 3–12. <https://doi.org/10.2174/1570164043488306>.
- (3) Dias, C. L.; Ala-Nissila, T.; Wong-ekkabut, J.; Vattulainen, I.; Grant, M.; Karttunen, M. The Hydrophobic Effect and Its Role in Cold Denaturation. *Cryobiology* **2010**, *60* (1), 91–99. <https://doi.org/10.1016/j.cryobiol.2009.07.005>.
- (4) Talley, K.; Alexov, E. On the PH-Optimum of Activity and Stability of Proteins. *Proteins* **2010**, *78* (12), 2699–2706. <https://doi.org/10.1002/prot.22786>.
- (5) Anfinsen, C. B. Principles That Govern the Folding of Protein Chains. *Science* **1973**, *181* (4096), 223–230. <https://doi.org/10.1126/science.181.4096.223>.
- (6) Parui, S.; Jana, B. Cold Denaturation Induced Helix-to-Helix Transition and Its

- Implication to Activity of Helical Antifreeze Protein. *J. Mol. Liq.* **2021**, 338, 116627. <https://doi.org/10.1016/j.molliq.2021.116627>.
- (7) Killian, M. S.; Krebs, H. M.; Schmuki, P. Protein Denaturation Detected by Time-of-Flight Secondary Ion Mass Spectrometry. *Langmuir* **2011**, 27 (12), 7510–7515. <https://doi.org/10.1021/la200704s>.
- (8) Wikimedia Commons. Protein structure [https://en.wikipedia.org/wiki/File:Protein_structure_\(full\).png](https://en.wikipedia.org/wiki/File:Protein_structure_(full).png) (accessed Oct 23, 2021).
- (9) Branden, C. I.; Tooze, J. *Introduction to Protein Structure*, Second.; Garland Science: New York, USA, 2012.
- (10) Wikimedia Commons. A peptide bond generated in PyMOL <https://commons.wikimedia.org/wiki/File:Peptide-Figure-Revised.png#filelinks> (accessed Jul 29, 2021).
- (11) Champe, P. C.; Harvey, R. A. *Lippincott's Illustrated Reviews: Biochemistry*, Third.; Lippincott Williams & Wilkins: Baltimore, USA, 2005.
- (12) Schmidt, R. L.; Simonović, M. Synthesis and Decoding of Selenocysteine and Human Health. *Croat. Med. J.* **2012**, 53 (6), 535–550. <https://doi.org/10.3325/cmj.2012.53.535>.
- (13) Wikimedia Commons. Amino Acids https://commons.wikimedia.org/wiki/File:Molecular_structures_of_the_21_proteinogenic_amino_acids.svg (accessed Sep 25, 2021).
- (14) Bellissent-Funel, M. C.; Hassanali, A.; Havenith, M.; Henschman, R.; Pohl, P.; Sterpone, F.; Van Der Spoel, D.; Xu, Y.; Garcia, A. E. Water Determines the Structure and Dynamics of Proteins. *Chem. Rev.* **2016**, 116 (13), 7673–7697. <https://doi.org/10.1021/acs.chemrev.5b00664>.
- (15) Baldwin, R. L. Dynamic Hydration Shell Restores Kauzmann's 1959 Explanation

- of How the Hydrophobic Factor Drives Protein Folding. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111* (36), 13052–13056. <https://doi.org/10.1073/pnas.1414556111>.
- (16) Snyder, P. W.; Mecinović, J.; Moustakas, D. T.; Thomas, S. W.; Harder, M.; Mack, E. T.; Lockett, M. R.; Héroux, A.; Sherman, W.; Whitesides, G. M. Mechanism of the Hydrophobic Effect in the Biomolecular Recognition of Arylsulfonamides by Carbonic Anhydrase. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108* (44), 17889–17894. <https://doi.org/10.1073/pnas.1114107108>.
- (17) Jaskolski, M.; Dauter, Z.; Wlodawer, A. A Brief History of Macromolecular Crystallography, Illustrated by a Family Tree and Its Nobel Fruits. *FEBS J.* **2014**, *281* (18), 3985–4009. <https://doi.org/10.1111/febs.12796>.
- (18) Zwanzig, R.; Szabo, A.; Bagchi, B. Levinthal's Paradox. *Proc. Natl. Acad. Sci. U. S. A.* **1992**, *89* (1), 20–22. <https://doi.org/10.1073/pnas.89.1.20>.
- (19) Dobson, C. M. Protein Folding: Solid Evidence for Molten Globules. *Curr. Biol.* **1994**, *4* (7), 636–640. [https://doi.org/10.1016/S0960-9822\(00\)00141-X](https://doi.org/10.1016/S0960-9822(00)00141-X).
- (20) Frauenfelder, H.; Sligar, S. G.; Wolynes, P. G. The Energy Landscapes and Motions of Proteins. *Science* **1991**, *254* (5038), 1598–1603.
- (21) DeForte, S.; Uversky, V. N. Order, Disorder, and Everything in Between. *Molecules* **2016**, *21* (8). <https://doi.org/10.3390/molecules21081090>.
- (22) Wright, P. E.; Dyson, H. J. Intrinsically Unstructured Proteins: Re-Assessing the Protein Structure-Function Paradigm. *J. Mol. Biol.* **1999**, *293* (2), 321–331. <https://doi.org/10.1006/jmbi.1999.3110>.
- (23) Wright, P. E.; Dyson, H. J. Intrinsically Disordered Proteins in Cellular Signalling and Regulation. *Nat. Rev. Mol. Cell Biol.* **2015**, *16* (1), 18–29. <https://doi.org/10.1038/nrm3920>.
- (24) Dunker, a K.; Obradovic, Z.; Romero, P.; Garner, E. C.; Brown, C. J. Intrinsic Protein Disorder in Complete Genomes. *Genome Inform. Ser. Workshop Genome*

- Inform.* **2000**, *11*, 161–171. <https://doi.org/10.11234/gi1990.11.161>.
- (25) Tompa, P. Intrinsically Disordered Proteins: A 10-Year Recap. *Trends Biochem. Sci.* **2012**, *37* (12), 509–516. <https://doi.org/10.1016/j.tibs.2012.08.004>.
- (26) Eliezer, D. Biophysical Characterization of Intrinsically Disordered Proteins. *Curr. Opin. Struct. Biol.* **2009**, *19* (1), 23–30. <https://doi.org/10.1016/j.sbi.2008.12.004>.
- (27) Milles, S.; Salvi, N.; Blackledge, M.; Jensen, M. R. Characterization of Intrinsically Disordered Proteins and Their Dynamic Complexes: From in Vitro to Cell-like Environments. *Prog. Nucl. Magn. Reson. Spectrosc.* **2018**, *109*, 79–100. <https://doi.org/10.1016/j.pnmrs.2018.07.001>.
- (28) Moul, J.; Pedersen, J. T.; Judson, R.; Fidelis, K. A Large-Scale Experiment to Assess Protein Structure Prediction Methods. *Proteins Struct. Funct. Bioinforma.* **1995**, *23* (3), ii–iv. <https://doi.org/https://doi.org/10.1002/prot.340230303>.
- (29) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Židek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596*, 583–589. <https://doi.org/10.1038/s41586-021-03819-2>.
- (30) Ruff, K. M.; Pappu, R. V. AlphaFold and Implications for Intrinsically Disordered Proteins. *J. Mol. Biol.* **2021**, *433* (20), 167208. <https://doi.org/10.1016/j.jmb.2021.167208>.
- (31) Chong, S. H.; Chatterjee, P.; Ham, S. Computer Simulations of Intrinsically Disordered Proteins. *Annu. Rev. Phys. Chem.* **2017**, *68*, 117–134. <https://doi.org/10.1146/annurev-physchem-052516-050843>.
- (32) Piana, S.; Lindorff-larsen, K.; Shaw, D. E. How Robust Are Protein Folding

Simulations with Respect to Force Field Parameterization? *Biophys. J.* **2011**, *100* (9), L47–L49. <https://doi.org/10.1016/j.bpj.2011.03.051>.

- (33) Piana, S.; Klepeis, J. L.; Shaw, D. E. Assessing the Accuracy of Physical Models Used in Protein-Folding Simulations: Quantitative Evidence from Long Molecular Dynamics Simulations. *Curr. Opin. Struct. Biol.* **2014**, *24*, 98–105. <https://doi.org/10.1016/j.sbi.2013.12.006>.
- (34) Cino, E. A.; Choy, W.; Karttunen, M. Comparison of Secondary Structure Formation Using 10 Different Force Fields in Microsecond Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2012**, *8* (8), 2725–2740. <https://doi.org/10.1021/ct300323g>.
- (35) Sickmeier, M.; Hamilton, J. A.; LeGall, T.; Vacic, V.; Cortese, M. S.; Tantos, A.; Szabo, B.; Tompa, P.; Chen, J.; Uversky, V. N.; Obradovic, Z.; Dunker, A. K. DisProt: The Database of Disordered Proteins. *Nucleic Acids Res.* **2007**, *35* (Database issue), 786–793. <https://doi.org/10.1093/nar/gkl893>.
- (36) Necci, M.; Piovesan, D.; Hoque, M. T.; Walsh, I.; Iqbal, S.; Vendruscolo, M.; Sormanni, P.; Wang, C.; Raimondi, D.; Sharma, R.; Zhou, Y.; Litfin, T.; Galzitskaya, O. V.; Lobanov, M. Y.; Vranken, W.; Wallner, B.; Mirabello, C.; Malhis, N.; Dosztányi, Z.; Erdős, G.; Mészáros, B.; Gao, J.; Wang, K.; Hu, G.; Wu, Z.; Sharma, A.; Hanson, J.; Paliwal, K.; Callebaut, I.; Bitard-Feildel, T.; Orlando, G.; Peng, Z.; Xu, J.; Wang, S.; Jones, D. T.; Cozzetto, D.; Meng, F.; Yan, J.; Gsponer, J.; Cheng, J.; Wu, T.; Kurgan, L.; Promponas, V. J.; Tamana, S.; Marino-Buslje, C.; Martínez-Pérez, E.; Chasapi, A.; Ouzounis, C.; Dunker, A. K.; Kajava, A. V.; Leclercq, J. Y.; Aykac-Fas, B.; Lambrugh, M.; Maiani, E.; Papaleo, E.; Chemes, L. B.; Álvarez, L.; González-Foutel, N. S.; Iglesias, V.; Pujols, J.; Ventura, S.; Palopoli, N.; Benítez, G. I.; Parisi, G.; Bassot, C.; Elofsson, A.; Govindarajan, S.; Lamb, J.; Salvatore, M.; Hatos, A.; Monzon, A. M.; Bevilacqua, M.; Mičetić, I.; Minervini, G.; Paladin, L.; Quaglia, F.; Leonardi, E.; Davey, N.; Horvath, T.; Kovacs, O. P.; Murvai, N.; Pancsa, R.; Schad, E.; Szabo, B.; Tantos, A.; Macedo-Ribeiro, S.; Manso, J. A.; Pereira, P. J. B.; Davidović, R.; Veljkovic, N.; Hajdu-

- Soltész, B.; Pajkos, M.; Szaniszló, T.; Guharoy, M.; Lazar, T.; Macossay-Castillo, M.; Tompa, P.; Tosatto, S. C. E. Critical Assessment of Protein Intrinsic Disorder Prediction. *Nat. Methods* **2021**, *18* (5), 472–481. <https://doi.org/10.1038/s41592-021-01117-3>.
- (37) Best, R. B.; Buchete, N. V.; Hummer, G. Are Current Molecular Dynamics Force Fields Too Helical? *Biophys. J.* **2008**, *95* (1), 7–9. <https://doi.org/10.1529/biophysj.108.132696>.
- (38) Best, R. B. Computational and Theoretical Advances in Studies of Intrinsically Disordered Proteins. *Curr. Opin. Struct. Biol.* **2017**, *42*, 147–154. <https://doi.org/10.1016/j.sbi.2017.01.006>.
- (39) Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E. M.; Mittal, J.; Feig, M.; MacKerell, Alexander D., J. Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone Phi, Psi and Side-Chain Chi(1) and Chi(2) Dihedral Angles. *J. Chem. Theory Comput.* **2012**, *8* (9), 3257–3273. <https://doi.org/10.1021/ct3004000x>.
- (40) Best, R. B.; Hummer, G. Optimized Molecular Dynamics Force Fields Applied to the Helix-Coil Transition of Polypeptides. *J. Phys. Chem. B* **2009**, *113* (26), 9004–9015. <https://doi.org/10.1021/jp901540t>.
- (41) Aliev, A. E.; Kulke, M.; Khaneja, H. S.; Chudasama, V.; Sheppard, T. D.; Lanigan, R. M. Motional Timescale Predictions by Molecular Dynamics Simulations: Case Study Using Proline and Hydroxyproline Sidechain Dynamics. *Proteins* **2014**, *82* (2), 195–215. <https://doi.org/10.1002/prot.24350>.
- (42) Best, R. B.; Zheng, W.; Mittal, J. Balanced Protein-Water Interactions Improve Properties of Disordered Proteins and Non-Specific Protein Association. *J. Chem. Theory Comput.* **2014**, *10* (11), 5113–5124. <https://doi.org/10.1021/ct500569b>.
- (43) Piana, S.; Donchev, A. G.; Robustelli, P.; Shaw, D. E. Water Dispersion Interactions Strongly Influence Simulated Structural Properties of Disordered Protein States. *J.*

- Phys. Chem. B* **2015**, *119* (16), 5113–5123. <https://doi.org/10.1021/jp508971m>.
- (44) Rauscher, S.; Gapsys, V.; Gajda, M. J.; Zweckstetter, M.; De Groot, B. L.; Grubmüller, H. Structural Ensembles of Intrinsically Disordered Proteins Depend Strongly on Force Field: A Comparison to Experiment. *J. Chem. Theory Comput.* **2015**, *11* (11), 5513–5524. <https://doi.org/10.1021/acs.jctc.5b00736>.
- (45) Chang, M.; Wilson, C. J.; Karunatileke, N. C.; Moselhy, M. H.; Karttunen, M.; Choy, W. Y. Exploring the Conformational Landscape of the Neh4 and Neh5 Domains of Nrf2 Using Two Different Force Fields and Circular Dichroism. *J. Chem. Theory Comput.* **2021**, *17* (5), 3145–3156. <https://doi.org/10.1021/acs.jctc.0c01243>.
- (46) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; De Groot, B. L.; Grubmüller, H.; MacKerell, A. D. CHARMM36m: An Improved Force Field for Folded and Intrinsically Disordered Proteins. *Nat. Methods* **2016**, *14* (1), 71–73. <https://doi.org/10.1038/nmeth.4067>.
- (47) Lindorff-larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Improved Side-Chain Torsion Potentials for the Amber Ff99SB Protein Force Field. *Proteins Struct. Funct. Bioinforma.* **2010**, 1950–1958. <https://doi.org/10.1002/prot.22711>.
- (48) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters. *Proteins Struct. Funct. Bioinforma.* **2006**, *725* (June), 712–725. <https://doi.org/10.1002/prot>.
- (49) Knowles, J. R.; Albery, W. J. Perfection in Enzyme Catalysis: The Energetics of Triosephosphate Isomerase. *Acc. Chem. Res.* **1977**, *10* (4), 105–111. <https://doi.org/10.1021/cen-v046n004.p005>.
- (50) Blacklow, S. C.; Raines, R. T.; Lim, W. A.; Zamore, P. D.; Knowles, J. R. Triosephosphate Isomerase Catalysis Is Diffusion Controlled. *Biochemistry* **1988**,

27, 1158–1167. <https://doi.org/10.1021/bi00404a013>.

- (51) Wierenga, R. K.; Kapetaniou, E. G.; Venkatesan, R. Triosephosphate Isomerase: A Highly Evolved Biocatalyst. *Cell. Mol. Life Sci.* **2010**, *67* (23), 3961–3982. <https://doi.org/10.1007/s00018-010-0473-9>.
- (52) Wierenga, R. K. The TIM-Barrel Fold: A Versatile Framework for Efficient Enzymes. *FEBS Lett.* **2001**, *492*, 193–198. [https://doi.org/10.1016/s0014-5793\(01\)02236-0](https://doi.org/10.1016/s0014-5793(01)02236-0).
- (53) Roland, B. P.; Stuchul, K. A.; Larsen, S. B.; Amrich, C. G.; VanDemark, A. P.; Celotto, A. M.; Palladino, M. J. Evidence of a Triosephosphate Isomerase Non-Catalytic Function Crucial to Behavior and Longevity. *J. Cell Sci.* **2013**, *126* (14), 3151–3158. <https://doi.org/10.1242/jcs.124586>.
- (54) Liao, Q.; Kulkarni, Y.; Sengupta, U.; Petrović, D.; Mulholland, A. J.; Van Der Kamp, M. W.; Strodel, B.; Kamerlin, S. C. L. Loop Motion in Triosephosphate Isomerase Is Not a Simple Open and Shut Case. *J. Am. Chem. Soc.* **2018**, *140* (46), 15889–15903. <https://doi.org/10.1021/jacs.8b09378>.
- (55) Borchert, T. V.; Kishan, K. R.; Zeelen, J. P.; Schliebs, W.; Thanki, N.; Abagyan, R.; Jaenicke, R.; Wierenga, R. K. Three New Crystal Structures of Point Mutation Variants of Mono TIM: Conformational Flexibility of Loop-1, Loop-4 and Loop-8. *Structure* **1995**, *3* (7), 669–679. [https://doi.org/10.1016/S0969-2126\(01\)00202-7](https://doi.org/10.1016/S0969-2126(01)00202-7).
- (56) Gómez-Puyou, A.; Saavedra-Lira, E.; Becker, I.; Zubillaga, R. A.; Rojo-Domínguez, A.; Perez-Montfort, R. Using Evolutionary Changes to Achieve Species-Specific Inhibition of Enzyme Action - Studies with Triosephosphate Isomerase. *Chem. Biol.* **1995**, *2* (12), 847–855. [https://doi.org/10.1016/1074-5521\(95\)90091-8](https://doi.org/10.1016/1074-5521(95)90091-8).
- (57) Velanker, S. S.; Ray, S. S.; Gokhale, R. S.; Suma, S.; Balaram, H.; Balaram, P.; Murthy, M. R. N. Triosephosphate Isomerase from Plasmodium Falciparum: The Crystal Structure Provides Insights into Antimalarial Drug Design. *Structure* **1997**,

- 5 (6), 751–761. [https://doi.org/10.1016/S0969-2126\(97\)00230-X](https://doi.org/10.1016/S0969-2126(97)00230-X).
- (58) Téllez-Valencia, A.; Olivares-Illana, V.; Hernández-Santoyo, A.; Pérez-Montfort, R.; Costas, M.; Rodríguez-Romero, A.; López-Calahorra, F.; Tuena De Gómez-Puyou, M.; Gómez-Puyou, A. Inactivation of Triosephosphate Isomerase from *Trypanosoma Cruzi* by an Agent That Perturbs Its Dimer Interface. *J. Mol. Biol.* **2004**, *341* (5), 1355–1365. <https://doi.org/10.1016/j.jmb.2004.06.056>.
- (59) Garza-Ramos, G.; Cabrera, N.; Saavedra-Lira, E.; Tuena De Gómez-Puyou, M.; Ostoa-Saloma, P.; Pérez-Montfort, R.; Gómez-Puyou, A. Sulfhydryl Reagent Susceptibility in Proteins with High Sequence Similarity: Triosephosphate Isomerase from *Trypanosoma Brucei*, *Trypanosoma Cruzi* and *Leishmania Mexicana*. *Eur. J. Biochem.* **1998**, *253* (3), 684–691. <https://doi.org/10.1046/j.1432-1327.1998.2530684.x>.
- (60) García-Torres, I.; Cabrera, N.; Torres-Larios, A.; Rodríguez-Bolaños, M.; Díaz-Mazariegos, S.; Gómez-Puyou, A.; Perez-Montfort, R. Identification of Amino Acids That Account for Long-Range Interactions in Two Triosephosphate Isomerases from Pathogenic Trypanosomes. *PLoS One* **2011**, *6* (4). <https://doi.org/10.1371/journal.pone.0018791>.
- (61) Zomosa-Signoret, V.; Hernández-Alcántara, G.; Reyes-Vivas, H.; Martínez-Martínez, E.; Garza-Ramos, G.; Pérez-Montfort, R.; De Gómez-Puyou, M. T.; Gómez-Puyou, A. Control of the Reactivation Kinetics of Homodimeric Triosephosphate Isomerase from Unfolded Monomers. *Biochemistry* **2003**, *42* (11), 3311–3318. <https://doi.org/10.1021/bi0206560>.
- (62) Reyes-Vivas, H.; Martínez-Martínez, E.; Mendoza-Hernández, G.; López-Velázquez, G.; Pérez-Montfort, R.; Tuena De Gómez-Puyou, M.; Gómez-Puyou, A. Susceptibility to Proteolysis of Triosephosphate Isomerase from Two Pathogenic Parasites: Characterization of an Enzyme with an Intact and a Nicked Monomer. *Proteins Struct. Funct. Genet.* **2002**, *48* (3), 580–590. <https://doi.org/10.1002/prot.10179>.

- (63) Rodríguez-Bolaños, M.; Cabrera, N.; Perez-Montfort, R. Identification of the Critical Residues Responsible for Differential Reactivation of the Triosephosphate Isomerases of Two Trypanosomes. *Open Biol.* **2016**, *6* (10). <https://doi.org/10.1098/rsob.160161>.
- (64) Chahrour, M.; Jung, S. Y.; Shaw, C.; Zhou, X.; Wong, S. T. C.; Qin, J.; Zoghbi, H. Y. MeCP2, a Key Contributor to Neurological Disease, Activates and Represses Transcription. *Science* **2008**, *320* (May), 1224–1230. <https://doi.org/10.1126/science.1153252>.
- (65) Aimer, R.; Van der Veyver, I.; Wan, M.; Tran, C.; Francke, U.; Zoghbi, H. Rett Syndrome Is Caused by Mutations in X-Linked MECP2 Encoding Methyl-CpG-Binding Protein 2. *Nat. Genet* **1999**, *23* (october), 185–188.
- (66) Hagberg, B. Rett's Syndrome: Prevalence and Impact on Progressive Severe Mental Retardation in Girls. *Acta Paediatrica* **1985**, *74* (3), 405–408. <https://doi.org/10.1111/j.1651-2227.1985.tb10993.x>.
- (67) Villard, L. MECP2 Mutations in Males. *J. Med. Genet.* **2007**, *44* (7), 417–423. <https://doi.org/10.1136/jmg.2007.049452>.
- (68) Adams, V. H.; McBryant, S. J.; Wade, P. A.; Woodcock, C. L.; Hansen, J. C. Intrinsic Disorder and Autonomous Domain Function in the Multifunctional Nuclear Protein, MeCP2. *J. Biol. Chem.* **2007**, *282* (20), 15057–15064. <https://doi.org/10.1074/jbc.M700855200>.
- (69) Wakefield, R. I. D.; Smith, B. O.; Nan, X.; Free, A.; Soteriou, A.; Uhrin, D.; Bird, A. P.; Barlow, P. N. The Solution Structure of the Domain from MeCP2 That Binds to Methylated DNA. *J. Mol. Biol.* **1999**, *291* (5), 1055–1065. <https://doi.org/10.1006/jmbi.1999.3023>.
- (70) Hite, K. C.; Adams, V. H.; Hansen, J. C. Recent Advances in MeCP2 Structure and Function. *Biochem. Cell Biol.* **2009**, *87* (1), 219–227. <https://doi.org/10.1139/O08-115>.

- (71) Hite, K. C.; Kalashnikova, A. A.; Hansen, J. C. Coil-to-Helix Transitions in Intrinsically Disordered Methyl CpG Binding Protein 2 and Its Isolated Domains. *Protein Sci.* **2012**, *21* (4), 531–538. <https://doi.org/10.1002/pro.2037>.
- (72) Kucukkal, T. G.; Alexov, E. Structural, Dynamical, and Energetical Consequences of Rett Syndrome Mutation R133C in MeCP2. *Comput. Math. Methods Med.* **2015**, *2015*.
- (73) Yang, Y.; Kucukkal, T. G.; Li, J.; Alexov, E.; Cao, W. Binding Analysis of Methyl-CpG Binding Domain of MeCP2 and Rett Syndrome Mutations. *ACS Chem. Biol.* **2016**, *11* (10), 2706–2715. <https://doi.org/10.1021/acscchembio.6b00450>.

2 Molecular dynamics simulations

In this work, molecular dynamics simulations were used to study two different proteins. The following chapter gives an overview of this method.

2.1 Introduction

The field of molecular dynamics (MD) simulations began with the work of Alder and Wainwright on hard-sphere liquids in the late 1950's¹, followed in 1964 by Rahman's work on a MD simulation of liquid argon with a Lennard-Jones potential². Stillinger and Rahman's study of liquid water³, published in 1971, finished preparing the stage, and in 1975 the first simulation of a macromolecule of biological interest was published⁴. The simulation concerned BPTI, a small, highly stable protein, whose X-ray structure became available the same year⁵. Although this simulation was done in vacuum with a crude molecular mechanics potential and lasted for only 9.2 ps, the results were essential in changing our view of proteins as relatively rigid structures. This work is part of what led to the Nobel Prize in Chemistry to Karplus, Warshel and Levitt in 2013. As a result of methodology improvements as well as the ever increasing computing power, MD simulations have become a standard tool in the study of biomolecules⁶.

An MD simulation produces a dynamical trajectory for a system composed of N atoms by integrating Newton's equations of motion. A set of initial conditions, a model to represent the forces acting between atoms, and boundary conditions are needed. The most common choice is periodic boundary conditions (PBC). Then, one needs to solve the classical equations of motion:

$$m_i \ddot{\mathbf{r}}_i = \mathbf{f}_i = \sum_{\substack{j=1 \\ (j \neq i)}}^N \mathbf{f}_{ij} \qquad \mathbf{f}_i = -\frac{\partial}{\partial \mathbf{r}_i} u(\mathbf{r}), \qquad (1)$$

where m is the mass of atom i , $\ddot{\mathbf{r}}_i$ is its acceleration, and the sum is over all N atoms, excluding i itself. We need to calculate the forces \mathbf{f}_i acting on each atom, which are in turn derived from a potential energy $u = u(\mathbf{r})$ where $\mathbf{r} = (\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N)$ represents the

complete set of $3N$ atomic coordinates. Newton's third law implies that the force exerted by the atom j on atom i is $\mathbf{f}_{ij} = -\mathbf{f}_{ji}$, so each atom pair needs to be examined only once. The potential energy function can be written as

$$\begin{aligned}
 u(\mathbf{r}) = & \sum_{bonds} k_i^{bond} (b_i - b_0)^2 + \sum_{angles} k_i^{angle} (\theta_i - \theta_0)^2 \\
 & + \sum_{dihedrals} k_i^{dih} [1 + \cos(n_i \varphi_i - \varphi_0)] + \sum_{impropers} k_i^{imp} (\psi_i - \psi_0)^2 \\
 & + \sum_i \sum_{j \neq i} 4 \epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] + \sum_i \sum_{j \neq i} \frac{q_i q_j}{r_{ij}}.
 \end{aligned} \tag{2}$$

The exact form of Equation 2 is determined by the force field, which contains all the necessary strength parameters k_i and constants therein ($b_0, \theta_0, \varphi_0, \psi_0$, etc). Charges are usually determined by quantum chemical calculations by fitting partial atomic charges to the quantum electrostatic potential, while force constants and idealized bond lengths and angles are often taken from crystal structures and adapted to match normal mode frequencies for certain peptide fragments⁷. All common force fields group these terms into bonded (first four terms) and non-bonded (last two terms) interactions. These are illustrated in Figure 2.1⁷⁻⁹.

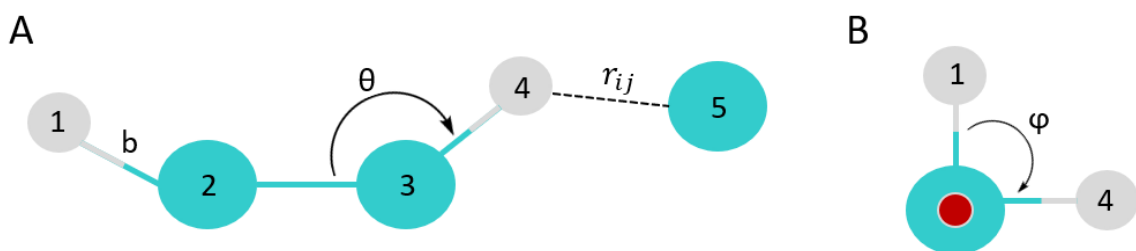


Figure 2.1 Schematic view of force field interactions. Covalent bonds are indicated by heavy solid lines and nonbonded interactions by a light, dashed line. A) Atoms 1 and 2 represent the bond length, atoms 2, 3 and 4 the angle θ , and atoms 4 and 5 show a nonbonded interaction with a distance r_{ij} . B) Atoms 1 and 4 represent the dihedral angle φ around the central bond between atoms 2 and 3. This dihedral represents the angle at which two adjacent planes meet.

In Equation 2, the first term represents the oscillations about the equilibrium bond length b_0 . It is represented with a harmonic potential where k_i^{bond} represents the force constant of the bond. This constant is generally very high, indicating that it takes a large amount of energy to stretch or compress a chemical bond. The harmonic potential is a good approximation for small deviations (smaller than 10%) from the reference bond length. Additionally, the use of the harmonic function implies that the bond cannot be broken, so no chemical processes can be studied. Since large deviations from the reference bond length are rare in simulations of biological macromolecules, other potential energy functions are rarely used⁸.

The second term in Equation 2 represents the oscillations of three atoms about an equilibrium bond angle θ_0 . The value of the force constants k_i^{angle} is typically lower than those for bond stretching, indicating that it takes less energy to deviate a bond angle from its reference value⁹.

The third term is the dihedral term (also known as the torsional term) and it represents the torsional rotation of four atoms around a central bond. Torsional motions are typically hundreds of times less stiff than bond stretching motions. They play a crucial role in determining the local structure of a macromolecule and the relative stability of different molecular conformations. In the potential energy function

$$u(\varphi_i) = \sum_{dihedrals} k_i^{dih} [1 + \cos(n_i\varphi_i - \varphi_0)] \quad (3)$$

k_i^{dih} determines the height of the potential energy barrier, n_i the number of minima between 0 and 2π , and φ_0 is the phase factor which determines their position. There is no unique way to determine the balance between the torsional, and the van der Waals and Coulomb components in the potential energy profile observed upon rotation of a dihedral angle. It is common practice between force field developers to combine Equation 3 with non-bonded energy terms to produce the desired torsion profile⁹.

The term $\sum_{improper} k_i^{imp} (\Psi_i - \Psi_0)^2$ is introduced in order to preserve the planarity of groups with flat geometry, such as sp^2 hybridized carbons in carbonyl groups or in aromatic rings. It provides a penalty function for bending out-of-plane⁸.

The fifth term in Eq. 2 represents the van der Waals component of the potential and is also known as the Lennard-Jones 12-6 potential. Here, ϵ_{ij} represents the depth of the well, and σ_{ij} the distance at which the potential energy between atoms i and j becomes zero. It is possible to define a set of parameters (ϵ_{ij} and σ_{ij}) for each different pair of atoms, but for convenience, most force fields give individual atomic parameters (ϵ_i and σ_i) together with rules on how to combine them (see section 2.5). The Lennard-Jones potential has an attractive and a repulsive term. The attractive one originates from the dispersion forces generated by instantaneous dipoles, which arise from fluctuations in the electronic charge distributions of all atoms. The repulsive term is due to the Pauli exclusion principle⁹.

The molecular electronic density can be obtained with high accuracy by means of high-level quantum mechanics calculations, however, reducing such density to a manageable description to be used in MD simulations is not trivial. The usual choice is to assign a partial atomic charge to each nucleus. This is a convenient representation as it allows the use of Coulomb's law to compute their contribution to the total energy,

$$u(\mathbf{r}) = \sum_i \sum_{j \neq i} \frac{q_i q_j}{r_{ij}}, \quad (4)$$

where r_{ij} is the distance between nuclei i and nuclei j , and q_i and q_j are the partial atomic charges. The most common way to calculate them consists in performing an *ab initio* calculation and then derive them from the quantum mechanical potential. Unfortunately, they cannot be derived unambiguously because atomic charges are not experimental observables and the methods developed to determine them do not always produce the same distribution of partial charges⁸.

2.2 Integration algorithms

Having computed all forces between the particles, one can integrate Newton's equations of motion. There are many methods to perform step-by-step numerical integration of these equations, but many of them are too costly or not stable enough for long simulations¹⁰. There are several requirements for a good integrator. The speed of the algorithm is relevant but not crucial because the fraction of time spent integrating the equations of motion, compared to computing the interactions, is relatively small. Accuracy for large time steps is more important, because the longer the time step one can use, the fewer evaluations of the forces are needed per unit of simulation time. Algorithms that use large time steps achieve this by storing information in increasingly higher-order derivatives of the particle coordinates and consequently, they require more memory storage. Another important criterion is energy conservation, which can occur short-term and long-term. The sophisticated higher-order algorithms have very good energy conservation for short times (i.e., during a small number of time steps). However, they often have energy drifts for long times. In contrast, Verlet-style algorithms tend to have only moderate short-term energy conservation but little long-term drift¹¹.

Newton's equations of motion are time reversible but many algorithms are not. As a consequence, if one were to reverse the momenta of all particles at a given instant, the system would not trace back its trajectory even if the calculation was done with infinite numerical precision. Many seemingly reasonable algorithms differ in another crucial aspect from Hamilton's equation of motion. True Hamiltonian dynamics leave the magnitude of any volume element in phase space unchanged (a property known as symplecticity), but many numerical schemes, in particular those that are not time-reversible, do not preserve the area in phase space. This in turn is not compatible with energy preservation and thus non-area-preserving algorithms will have serious long-term energy drift problems¹¹.

The Verlet algorithm¹² is fast, requires very little memory, has a fair short-term energy conservation and exhibits little long-term energy drift. This is due to the fact that this algorithm is time reversible and area preserving. Its disadvantage is that it is not particularly accurate for long time steps and so one needs to compute the forces on all particles rather

frequently. To derive it, we start with a Taylor expansion of the coordinate of a particle around time t ,

$$r(t + \Delta t) = r(t) + v(t)\Delta t + \frac{f(t)}{2m}\Delta t^2 + \frac{\Delta t^3}{3!}\ddot{r} + O(\Delta t^4) \quad (5)$$

similarly,

$$r(t - \Delta t) = r(t) - v(t)\Delta t + \frac{f(t)}{2m}\Delta t^2 - \frac{\Delta t^3}{3!}\ddot{r} + O(\Delta t^4) \quad (6)$$

Adding these two equations we get:

$$r(t + \Delta t) + r(t - \Delta t) = 2r(t) + \frac{f(t)}{m}\Delta t^2 + O(\Delta t^4) \quad (7)$$

which gives

$$r(t + \Delta t) \approx 2r(t) - r(t - \Delta t) + \frac{f(t)}{m}\Delta t^2 \quad (8)$$

The estimate of the new position has an error that is of the order of Δt^4 , where Δt is the time step used in the simulation. The Verlet algorithm does not use the velocity to compute the new position. However, one can derive the velocity from knowledge of the trajectory using:

$$r(t + \Delta t) - r(t - \Delta t) = 2v(t)\Delta t + O(\Delta t^3) \quad (9)$$

which then gives,

$$v(t) = \frac{r(t+\Delta t) - r(t-\Delta t)}{2\Delta t} + O(\Delta t^2). \quad (10)$$

The expression for the velocity is only accurate to order Δt^2 . Having the new positions, those at time $t - \Delta t$ may be discarded. The current positions become the old ones and the new positions become the current ones¹¹.

The Verlet algorithm as defined above is hardly ever used in practise. Instead, the velocity-Verlet algorithm is employed. It has the same properties as the Verlet algorithm but evolves explicitly the velocities with an accuracy in the order of $O(\Delta t^4)$. In this algorithm, the new positions are calculated from

$$r(t + \Delta t) = r(t) + v(t)\Delta t + \frac{f(t)}{2m}\Delta t^2 \quad (11)$$

and the velocity is updated by:

$$v(t + \Delta t) = v(t) + \frac{f(t+\Delta t)+f(t)}{2m}\Delta t. \quad (12)$$

Note that in this algorithm, we can compute the new velocities only after we have computed the new positions, and from these, the new forces.

There are still ongoing efforts to develop new algorithms that allow the use of longer time steps while preserving the dynamics. In particular, several multiple step algorithms have been proposed in recent years¹³⁻¹⁵. These algorithms exploit the time scale separation between faster (which can be integrated more frequently) and slower (updated at a lower rate) degrees of freedom.

2.3 Thermostats and Barostats

Integration of Newton's equations of motion produces the microcanonical ensemble. In this ensemble the number of particles N , the volume V and the total energy E are constant (NVE ensemble). Experiments, however, are typically conducted at constant temperature and/or constant pressure and, therefore, it is desirable to perform MD simulations in ensembles such as the canonical (NVT) or the isothermal-isobaric ensemble (NPT). This can be achieved with the use of thermostats and barostats⁹.

Thermostats can be roughly divided into global and local. Global ones act instantaneously with the same strength on all particles of the system. Local thermostats, on the other hand, act on individual atoms or pairs and dissipate energy on a spatially localized scale.

Nevertheless, the addition of one may significantly affect the thermal fluctuations in the system and cause energy drifts that sometimes have their origin in the accumulation of numerical errors¹⁶.

One of the best known artifacts in MD simulations is the so-called “flying ice cube”¹⁷. Over the simulation time there is a gradual loss of the vibrational, internal kinetic energy and an increase in translational external kinetic energy. Eventually the system freezes and becomes a flying ice cube. This artifact is a violation of the equipartition principle and is due to velocity rescaling, a technique applied by global thermostats including the commonly used Berendsen weak-coupling method¹⁸ and Nosé-Hoover^{19,20}. This problem can be solved in three different ways. The first one is velocity reassignment instead of rescaling. This essentially means using a local thermostat such as the Andersen²¹, Langevin^{22,23} or the DPD methods^{24,25}. The second way of addressing this problem is to remove the motion of the center of mass and the third one is to use better algorithms for rescaling. Modern simulation protocols remove the center of mass motion periodically throughout the simulation and use sufficiently large coupling times for global thermostats. The velocity rescale algorithm of Bussi, Donadio and Parrinello^{26,27} is a modification of the Berendsen weak coupling method, has been shown to perform well and is widely popular¹⁶.

Simulations using the NPT ensemble are typically needed for membrane systems. This ensemble is achieved through the use of a barostat. The most popular ones include the Langevin piston²⁸, the Parrinello-Rahman method²⁹, the Martyna-Tuckerman-Tobias-Klein algorithm^{30,31} and the Berendsen barostat¹⁸. The Berendsen method is conceptually simple and is very commonly used but is not entirely correct since it does not produce the correct canonical distribution³². In contrast, the Parrinello-Rahman barostat produces the correct distribution, but is computationally more expensive¹⁶.

2.4 Force fields

A force field is a mathematical expression of the potential energy of a system of particles. It consists of an analytical form of the interatomic potential energy (Equation 2) and its set of parameters. The fundamental assumption is the Born-Oppenheimer approximation³³, which neglects the motion of atomic nuclei when describing the electrons in a molecule. The physical basis of this approximation is that the electrons and the nucleons have more than a 1000-fold mass difference, which in turn causes the nuclei to move ~1000 times more slowly than electrons. The Born-Oppenheimer approximation makes possible to separate the motion of the nuclei and the electrons, and therefore to write the system as a function of the nuclear coordinates only and thus enables the use of classical mechanics.

There are two other assumptions at work: additivity and transferability. Additivity means that the potential energy of any system can be written as a sum of potentials with a simple physical interpretation (bond stretching, angle bending, van der Waals interactions, etc.). Additive force fields are characterized by point charges in each of the atoms, centered on the atomic nucleus. Many popular atomistic force fields are additive, the most widely used are AMBER^{34,35}, CHARMM^{36,37}, GROMOS^{38,39} and OPLS⁴⁰. Additive force fields are generally parametrized using a mean-field approximation such that the electronic distribution of molecules cannot change their response to variations in the local electric field. Polarizable force fields have three primary methods to treat polarization classically via the induced point dipole, fluctuating charge and the Drude oscillator. The requirement to solve the magnitude of all the induced dipoles self-consistently is computationally demanding, and it is typically the limiting factor in the efficiency of all polarizable models⁴¹. The inclusion of explicit electronic polarization allows one to transfer gas-phase *ab initio* potentials to condensed phases, and is recommended in cases where permanent polarization does not account for most of the simulated effect, such as in evaluating the redox potential of proteins and to study proton transport in membrane channels⁴². Recent efforts in developing new polarizable force fields have been made to improve water⁴³, RNA⁴⁴ and urea crystals and aqueous solutions⁴⁵ simulations.

The term transferable means that the force field parameters derived from a small set of molecules can be applied to molecules with similar chemical structures, and that the force

field is transferable to different state points (e.g., pressure, temperature) and to different properties (e.g., thermodynamic, structural)^{9,46}. While a force field that uses special types of interactions for specific molecules may be very accurate for a particular application, it could be considered not transferable and would therefore have very limited predictive power for unrelated applications. There are two types of parameter transfers. The first one is internal, in which the force field parameters are transferred within a molecule, such as using the parameter derived for a residue, in a protein. This type of transfer is valid in most cases. The other one is external, parameters derived from a molecule are used in a similar but different molecule. For example, parameters derived for alkanes may be used for halogen-substituted alkanes. External transfers can introduce considerable errors simply because some of the molecular properties may not be strictly transferable⁴⁷.

Beyond the three main assumptions made by force fields, they can differ in the way they treat the interactions between particles. There are two main differences in the bonded contributions of the force fields. The first one is the use of “improper” dihedrals, which can be used to maintain chirality or planarity at an atom center with bonds to three other atoms. While AMBER^{34,35} and OPLS⁴⁰ apply the dihedral term in Equation 2 to planar groups, CHARMM^{36,37} adds a separate term for improper dihedral energy that has a quadratic dependence on the value of the improper dihedral. The second difference is that the CHARMM^{36,37} force field adds an Urey-Bradley angle term. This term treats the two terminal atoms in an angle with a quadratic term that depends on the distance between the atoms⁴⁸.

Similar to the bonded terms, there are two key differences between force fields in the treatment of non-bonded interactions. The first one lies in the combination rules for the determination of the Lennard-Jones parameters ϵ_{ij} and σ_{ij} between dissimilar atoms. The subscript ij associated with these parameters is used to make explicit their dependence on the atom types for both atom i and atom j . OPLS⁴⁰ uses the geometric mean to calculate ϵ_{ij} and σ_{ij} , *i.e.* $k_{ij} = (k_i k_j)^{1/2}$, whereas AMBER^{34,35} and CHARMM^{36,37} use the geometric mean for ϵ_{ij} and the arithmetic mean, $\frac{1}{2}(\sigma_i + \sigma_j)$, for σ_{ij} , also known as the Lorentz-Berthelot rules of mixing⁴⁹. The Lorentz-Berthelot rules work reasonably well in most

cases but there are some instances in which they fail in a rather striking manner, see for example the discussion in Refs. (16) and (50). For example, they may lead to incorrect thermodynamic properties for simple binary mixtures and they give surface-gas interactions that are 10 times stronger than they should be¹⁶. The second difference is the handling of 1,4 non-bonded interactions. These are the interactions between atoms 1 and 4 in the 1-2-3-4 dihedral. Amber scales 1,4 Lennard-Jones interactions by $\frac{1}{2}$ and Coulomb interactions by $\frac{1}{1.2}$, OPLS⁴⁰ applies a factor of $\frac{1}{2}$ to both interactions, and CHARMM^{36,37} does not scale them except for a few atom type pairs⁴⁸.

The lack of consensus on the best way to treat both bonded and non-bonded interactions suggests that there is no single best solution to this problem. Both OPLS⁴⁰ and AMBER^{34,35} have undergone revisions of their Φ and Ψ dihedral parameters, giving rise to a number of descendants from their original force field^{40,51}. CHARMM^{36,37} has followed a different approach, by adding a new “correction map” (CMAP) term to the potential energy equation⁴⁸. This term is a grid-based correction for the Φ -, Ψ -angular dependence of the energy³⁷.

2.5 Time scales

In MD simulations one would ideally like to choose a time step as large as possible, in order to sample phase space rapidly and save on computer expense. However, if a too large time step is chosen, the motion of particles becomes unstable due to the truncation error in the integration process and the total energy of the system may increase rapidly with time. This behavior is called exploding and is caused when a large time step propagates the positions of atoms to be partially overlapping, creating a strong repulsive force between them⁵². In order to have numerical stability and accuracy in the conservation of energy, the time step must be chosen at least an order of magnitude smaller than the smallest vibrational period of the system⁵².

In protein simulations, the fastest intramolecular vibrational motions are the C-H, N-H and O-H stretches (Figure 2.2). For these, a time step of $\Delta t = 0.5 - 1$ fs is needed. In most

simulations, these bond vibrations are not of interest per se and constraints can be applied to the bond lengths and angles, making it possible to extend the time step to 2 fs⁴⁹. SHAKE⁵³ and LINCS^{54,55} are the constraint algorithms most widely used.

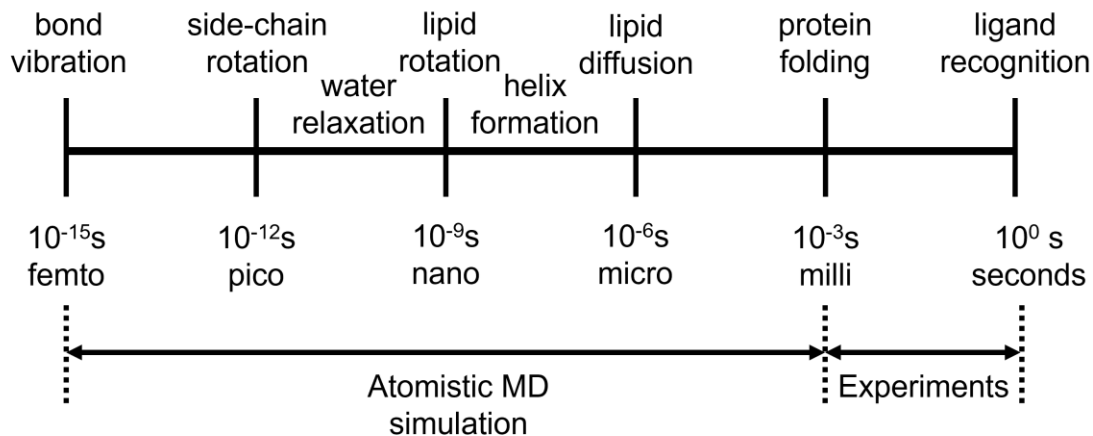


Figure 2.2 Range of timescale for atomistic MD simulations

Having fixed the time step, the timescales accessible in MD simulations will depend on the size of the system, and the hardware and software used to run the simulations. Roughly one billion arithmetic operations are needed at each time step for a system with one hundred thousand atoms. A single high-end processor core would take thousands of years to complete a millisecond long simulation. Fortunately, software that parallelizes MD force calculations across multiple computer processors as well as GPUs has existed for nearly three decades, becoming much more efficient and scalable in the past several years. The widely used MD codes NAMD⁵⁶, GROMACS⁵⁷ and AMBER⁵⁸ can now deliver a performance of over 100 ns/day on commodity computer clusters in systems of ~10,000 atoms, with the number of processors needed scaling roughly linearly with the number of atoms in the system⁵⁹. A typical simulation (a moderately sized, solvated, globular protein) has ~50,000 atoms.

There is a special-purpose supercomputer named Anton, designed by D. E. Shaw Research, that has made possible to reach the millisecond-scale in all-atom simulations. Anton is able to achieve this speed because it was specifically designed for MD simulations. It is the result of an algorithm/hardware co-design process in which the choice of algorithms

impacted the design of the hardware, and vice versa⁶⁰. While time on an older (2016) Anton machine has been generously granted to the scientific community, access is still extremely limited and is only available to some faculty at a U.S. academic or non-profit research institution.

2.6 Coarse-graining methods

One way to substantially accelerate simulations at the cost of reduced accuracy, is the use of coarse-grained models. In these, the original system is replaced by a simpler one, with less degrees of freedom, effectively averaging over some chosen properties of microscopic entities to form larger basic units⁶¹. Reducing the number of interactions that must be computed smoothens the energy landscape, making them faster than all-atom simulations (Figure 2.3)⁴⁹. The specific acceleration attained by a CG model depends on the details of the model used.

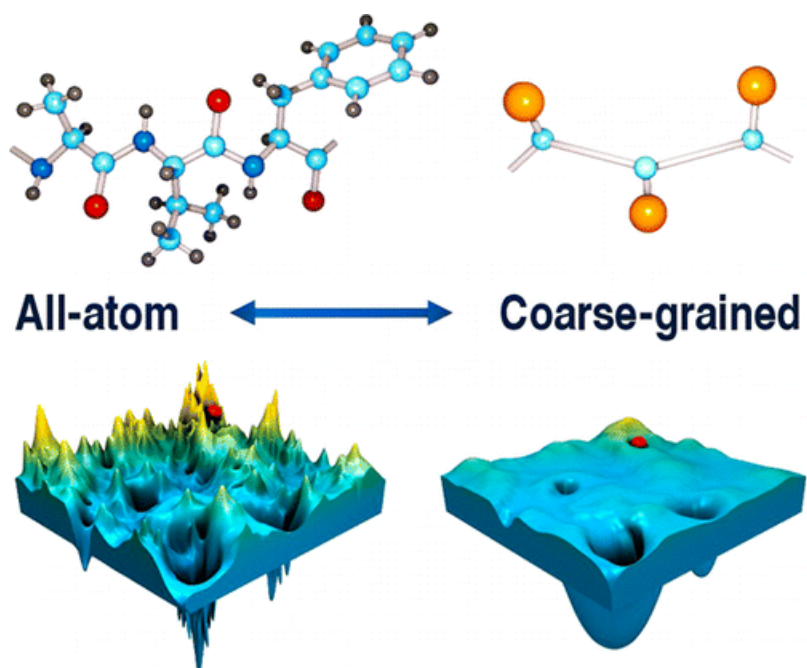


Figure 2.3 All-atom vs coarse-grained energy landscape. The flattening of the energy surface in the coarse-grained model enables its efficient exploration. From Chem Rev, 2016 (Open Access article under an ACS AuthorChoice License)⁶².

There are four conditions that degrees of freedom must meet in order to be eliminated in a physically correct manner, such that a computationally efficient yet accurate CG model is obtained:

1. they must be non-essential to the property of interest,
2. they must be large in number or computationally costly, so that the computational gain is substantial enough to offset the loss in accuracy,
3. the interactions governing the degrees of freedom to be eliminated must be largely decoupled from those which will be kept,
4. their elimination should allow a simple and efficient representation of the interactions governing the remaining degrees of freedom⁶³.

There are two fundamentally different approaches to designing a force field for CG models for particle simulations, one is “physics-based” and the other is “knowledge-based”. In general, a physics-based CG force field can be described by a similar formula as an all-atom force field (Equation 2). However, during the coarse-graining process, some atoms are removed and their degrees of freedom are averaged out. One must then introduce explicitly the internal correlations between groups of atoms (now represented as united atoms) in the form of multibody terms. Most approaches keep the distinction between local energy terms and so-called contact potentials⁶². This philosophy of modeling is also known as structure-based, systematic, hierarchical or bottom-up^{64,65}. It often requires a time-consuming parametrization procedure and has complex potential forms, resulting in lower performance and thus less sampling⁶⁶. Nevertheless, force fields such as the PLUM model⁶⁷ allow an accurate sampling of local conformations and can achieve a realistic α/β content. The theoretical justification of structure-based coarse-graining is the Henderson theorem⁶⁸, which defines a one-to-one relationship between a set of radial distribution functions and a set of pair potentials for CG sites.

In contrast, knowledge-based force fields are derived from the statistical analysis of structural features observed in databases of experimental structures. Depending on the level of coarse-grained representation, definition of the model force field and the complexity of the databases being used, the final formula that defines the force field will be composed from a significantly larger number of terms than an all-atom force field. Moreover, some terms describe specific conditional combinations of bonds, angular and non-bonded

interactions⁶². This method of modeling can also be found in the literature as building-block, thermodynamics-based or top-down. These models are often cheaper, due to simpler potential forms and requiring only partial parametrization. However, their structural accuracy is limited as the representation of the atomistic detail is suboptimal⁶⁶.

One of the most widely used CG models is the knowledge-based Martini⁶⁹⁻⁷¹ force field. It was initially developed for lipids but since then, it has been extended to include proteins, carbohydrates, DNA, RNA and small molecules. The Martini⁶⁹⁻⁷¹ model relies on a four-to-one mapping scheme, where on average four non-hydrogen atoms are mapped to a single CG bead (Figure 2.4). It has four main types of particles: polar, nonpolar, apolar and charged. These types are in turn divided into subtypes according to their hydrogen-bonding capabilities or their degree of polarity. This gives a total of 18 particle types⁷¹. The model was reparametrized in April 2021, and in contrast with previous versions, it defined a new particle type specific for water. Although this new bead type enables the optimization of water properties independently from other targets, structure-based CG models are more suitable for applications that require finer details⁷⁰.

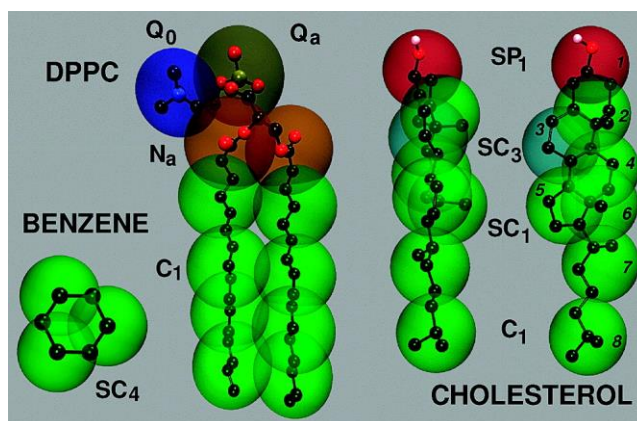


Figure 2.4 Mapping between the chemical structure and the coarse grained model for DPPC, cholesterol and benzene. The coarse-grained bead types which determine their relative hydrophilicity are indicated. Reprinted with permission from Marrink *et al*⁷¹. Copyright 2021 American Chemical Society.

Coarse-grained representations have been successfully used to study protein folding mechanisms⁷², conformational changes upon ligand binding, membrane proteins⁶⁹ and the self-assembly of protein/membrane and protein/detergent complexes⁷³.

2.7 References

- (1) Alder, B. J.; Wainwright, T. E. Phase Transition for a Hard Sphere System. *J. Chem. Phys.* **1957**, *27* (5), 1208–1209. <https://doi.org/10.1063/1.1743957>.
- (2) Rahman, A. Correlations in the Motion of Atoms in Liquid Argon. *Phys. Rev.* **1964**, *136* (2A), A405–A411. <https://doi.org/10.1103/PhysRev.136.A405>.
- (3) Rahman, A.; Stillinger, F. H. Molecular Dynamics Study of Liquid Water. *J. Chem. Phys.* **1971**, *55* (7), 3336–3359. <https://doi.org/10.1063/1.447358>.
- (4) Levitt, M.; Washel, A. Computational Simulation of Protein Folding. *Nature* **1975**, *253*, 694–698. <https://doi.org/10.1038/253694a0>.
- (5) Deisenhofer, J.; Steigemann, W. Crystallographic Refinement of the Structure of Bovine Pancreatic Trypsin Inhibitor at 1.5 Å Resolution. *Acta Crystallogr. Sect. B* **1975**, *31* (1), 238–250. <https://doi.org/10.1107/S0567740875002415>.
- (6) Karplus, M. Molecular Dynamics of Biological Macromolecules: A Brief History and Perspective. *Biopolymers* **2003**, *68* (3), 350–358. <https://doi.org/10.1002/bip.10266>.
- (7) Ponder, J. W.; Case, D. A. Force Fields for Protein Simulations. *Adv. Protein Chem.* **2003**, *66*, 27–85. [https://doi.org/10.1016/S0065-3233\(03\)66002-X](https://doi.org/10.1016/S0065-3233(03)66002-X).
- (8) González, M. A. Force Fields and Molecular Dynamics Simulations. *Collect. SFN* **2011**, *12*, 169–200. <https://doi.org/10.1051/sfn/201112009>.
- (9) Monticelli, L.; Tieleman, D. P. Force Fields for Classical Molecular Dynamics. In

- Biomolecular Simulations: Methods and Protocols*; Monticelli, L., Salonen, E., Eds.; Humana Press: Totowa, USA, 2013; pp 197–213. https://doi.org/10.1007/978-1-62703-017-5_8.
- (10) Leimkuhler, B. J.; Reich, S.; Skeel, R. D. Integration Methods for Molecular Dynamics. In *Mathematical Approaches to Biomolecular Structure and Dynamics*; Mesirov, J. P., Schulten, K., Sumners, D. W., Eds.; Springer New York: New York, USA, 1996; pp 161–185. https://doi.org/10.1007/978-1-4612-4066-2_10.
- (11) Frenkel, D.; Smit, B. Chapter 4 - Molecular Dynamics Simulations. In *Understanding Molecular Simulation*; Frenkel, D., Smit, B., Eds.; Academic Press: San Diego, 2002; pp 63–107. <https://doi.org/https://doi.org/10.1016/B978-012267351-1/50006-7>.
- (12) Verlet, L. Computer “Experiments” on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Phys. Rev.* **1967**, *159* (1), 98–103. <https://doi.org/10.1103/PhysRev.159.98>.
- (13) Liberatore, E.; Meli, R.; Rothlisberger, U. A Versatile Multiple Time Step Scheme for Efficient Ab Initio Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2018**, *14* (6), 2834–2842. <https://doi.org/10.1021/acs.jctc.7b01189>.
- (14) Lagardère, L.; Aviat, F.; Piquemal, J. P. Pushing the Limits of Multiple-Time-Step Strategies for Polarizable Point Dipole Molecular Dynamics. *J. Phys. Chem. Lett.* **2019**, *10* (10), 2593–2599. <https://doi.org/10.1021/acs.jpcclett.9b00901>.
- (15) Monmarché, P.; Weisman, J.; Lagardère, L.; Piquemal, J. P. Velocity Jump Processes: An Alternative to Multi-Timestep Methods for Faster and Accurate Molecular Dynamics Simulations. *J. Chem. Phys.* **2020**, *153*, 024101. <https://doi.org/10.1063/5.0005060>.
- (16) Wong-ekkabut, J.; Karttunen, M. The Good, the Bad and the User in Soft Matter Simulations. *Biochim. Biophys. Acta - Biomembr.* **2016**, *1858* (10), 2529–2538. <https://doi.org/10.1016/j.bbamem.2016.02.004>.

- (17) Harvey, S. C.; Tan, R. K.-Z.; Cheatham III, T. E. The Flying Ice Cube: Velocity Rescaling in Molecular Dynamics Leads to Violation of Energy Equipartition. *J. Comput. Chem.* **1997**, *19* (7), 726–740. [https://doi.org/10.1002/\(SICI\)1096-987X\(199805\)19:7<726::AID-JCC4>3.0.CO;2-S](https://doi.org/10.1002/(SICI)1096-987X(199805)19:7<726::AID-JCC4>3.0.CO;2-S).
- (18) Berendsen, H. J. C. C.; Postma, J. P. M. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular Dynamics with Coupling to an External Bath. *J. Chem. Phys.* **1984**, *81* (8), 3684–3690. <https://doi.org/10.1063/1.448118>.
- (19) Nosé, S. A Unified Formulation of the Constant Temperature Molecular Dynamics Methods. *J. Chem. Phys.* **1984**, *81* (1), 511–519. <https://doi.org/10.1063/1.447334>.
- (20) Hoover, W. G. Canonical Dynamics: Equilibrium Phase-Space Distributions. *Phys. Rev. A* **1985**, *31* (3), 1695–1697. <https://doi.org/10.1103/PhysRevA.31.1695>.
- (21) Andersen, H. C. Molecular Dynamics Simulations at Constant Pressure and/or Temperature. *J. Chem. Phys.* **1980**, *72* (4), 2384–2393. <https://doi.org/10.1063/1.439486>.
- (22) Schneider, T.; Stoll, E. Molecular-Dynamics Study of a Three-Dimensional One-Component Model for Distortive Phase Transitions. *Phys. Rev. B* **1978**, *17* (3), 1302–1322. <https://doi.org/10.1103/PhysRevB.17.1302>.
- (23) Kremer, K.; Grest, G. S. Dynamics of Entangled Linear Polymer Melts: A Molecular-dynamics Simulation. *J. Chem. Phys.* **1990**, *92* (8), 5057–5086. <https://doi.org/10.1063/1.458541>.
- (24) Español, P.; Warren, P. Statistical Mechanics of Dissipative Particle Dynamics. *Epl* **1995**, *30* (4), 191–196. <https://doi.org/10.1209/0295-5075/30/4/001>.
- (25) Vattulainen, I.; Karttunen, M.; Besold, G.; Polson, J. M. Integration Schemes for Dissipative Particle Dynamics Simulations: From Softly Interacting Systems towards Hybrid Models. *J. Chem. Phys.* **2002**, *116* (10), 3967–3979. <https://doi.org/10.1063/1.1450554>.

- (26) Bussi, G.; Donadio, D.; Parrinello, M. Canonical Sampling through Velocity Rescaling. *J. Chem. Phys.* **2007**, *126* (1), 1–7. <https://doi.org/10.1063/1.2408420>.
- (27) Bussi, G.; Zykova-Timan, T.; Parrinello, M. Isothermal-Isobaric Molecular Dynamics Using Stochastic Velocity Rescaling. *J. Chem. Phys.* **2009**, *130* (7), 74101. <https://doi.org/10.1063/1.3073889>.
- (28) Feller, S. E.; Zhang, Y.; Pastor, R. W.; Brooks, B. R. Constant Pressure Molecular Dynamics Simulation: The Langevin Piston Method. *J. Chem. Phys.* **1995**, *103* (11), 4613–4621. <https://doi.org/10.1063/1.470648>.
- (29) Parrinello, M.; Rahman, A. Polymorphic Transitions in Single Crystals : A New Molecular Dynamics Method. *J. Appl. Phys.* **1981**, *52* (12), 7182–7190. <https://doi.org/10.1063/1.328693>.
- (30) Martyna, G. J.; Tuckerman, M. E.; Tobias, D. J.; Klein, M. L. Explicit Reversible Integrators for Extended Systems Dynamics. *Mol. Phys.* **1996**, *87* (5), 1117–1157. <https://doi.org/10.1080/00268979600100761>.
- (31) Tuckerman, M. E.; Alejandre, J.; López-Rendón, R.; Jochim, A. L.; Martyna, G. J. A Liouville-Operator Derived Measure-Preserving Integrator for Molecular Dynamics Simulations in the Isothermal-Isobaric Ensemble. *J. Phys. A. Math. Gen.* **2006**, *39* (19), 5629–5651. <https://doi.org/10.1088/0305-4470/39/19/s18>.
- (32) Shirts, M. R. Simple Quantitative Tests to Validate Sampling from Thermodynamic Ensembles. *J. Chem. Theory Comput.* **2013**, *9* (2), 909–926. <https://doi.org/10.1021/ct300688p>.
- (33) Born, M.; Oppenheimer, R. Zur Quantentheorie Der Molekeln. *Ann. Phys.* **1927**, *389* (20), 457–484. <https://doi.org/https://doi.org/10.1002/andp.19273892002>.
- (34) Aliev, A. E.; Kulke, M.; Khaneja, H. S.; Chudasama, V.; Sheppard, T. D.; Lanigan, R. M. Motional Timescale Predictions by Molecular Dynamics Simulations: Case Study Using Proline and Hydroxyproline Sidechain Dynamics. *Proteins* **2014**, *82* (2), 195–215. <https://doi.org/10.1002/prot.24350>.

- (35) Svozil, D.; Sponer, J.; Iii, T. E. C.; Laughton, C. A.; Orozco, M. Refinement of the AMBER Force Field for Nucleic Acids: Improving the Description of α/γ Conformers. *Biophys. J.* **2007**, *92* (June), 3817–3829. <https://doi.org/10.1529/biophysj.106.097782>.
- (36) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; De Groot, B. L.; Grubmüller, H.; MacKerell, A. D. CHARMM36m: An Improved Force Field for Folded and Intrinsically Disordered Proteins. *Nat. Methods* **2016**, *14* (1), 71–73. <https://doi.org/10.1038/nmeth.4067>.
- (37) Buck, M.; Bouguet-Bonnet, S.; Pastor, R. W.; MacKerell, A. D. Importance of the CMAP Correction to the CHARMM22 Protein Force Field: Dynamics of Hen Lysozyme. *Biophys. J.* **2006**, *90* (4), 36–38. <https://doi.org/10.1529/biophysj.105.078154>.
- (38) Daura, X.; Mark, A. E.; Van Gunsteren, W. F. Parametrization of Aliphatic CHN United Atoms of GROMOS96 Force Field. *J. Comput. Chem.* **1998**, *19* (5), 535–547. [https://doi.org/10.1002/\(SICI\)1096-987X\(19980415\)19:5<535::AID-JCC6>3.0.CO;2-N](https://doi.org/10.1002/(SICI)1096-987X(19980415)19:5<535::AID-JCC6>3.0.CO;2-N).
- (39) Soares, T. A.; Hünenberger, P. H.; Kastenholz, M. A.; Kräutler, V.; Lenz, T.; Lins, R. D.; Oostenbrink, C.; Van Gunsteren, W. F. An Improved Nucleic Acid Parameter Set for the GROMOS Force Field. *J. Comput. Chem.* **2005**, *26* (7), 725–737. <https://doi.org/10.1002/jcc.20193>.
- (40) Kaminski, G. A.; Friesner, R. A.; Tirado-rives, J.; Jorgensen, W. L. Evaluation and Reparametrization of the OPLS-AA Force Field for Proteins via Comparison with Accurate Quantum Chemical Calculations on Peptides. *J. Phys. Chem. B* **2001**, *105* (28), 6474–6487.
- (41) Lemkul, J. A.; Huang, J.; Roux, B.; Mackerell, A. D. An Empirical Polarizable Force Field Based on the Classical Drude Oscillator Model: Development History and Recent Applications. *Chem. Rev.* **2016**, *116* (9), 4983–5013. <https://doi.org/10.1021/acs.chemrev.5b00505>.

- (42) Warshel, A.; Kato, M.; Pislakov, A. V. Polarizable Force Fields: History, Test Cases, and Prospects. *J. Chem. Theory Comput.* **2007**, *3* (6), 2034–2045. <https://doi.org/10.1021/ct700127w>.
- (43) Naserifar, S.; Goddard, W. A. The Quantum Mechanics-Based Polarizable Force Field for Water Simulations. *J. Chem. Phys.* **2018**, *149* (17). <https://doi.org/10.1063/1.5042658>.
- (44) Lemkul, J. A.; MacKerell, A. D. Polarizable Force Field for RNA Based on the Classical Drude Oscillator. *J. Comput. Chem.* **2018**, *39* (32), 2624–2646. <https://doi.org/10.1002/jcc.25709>.
- (45) Jeong, K. J.; McDaniel, J. G.; Yethiraj, A. A Transferable Polarizable Force Field for Urea Crystals and Aqueous Solutions. *J. Phys. Chem. B* **2020**, *124* (34), 7475–7483. <https://doi.org/10.1021/acs.jpcc.0c05814>.
- (46) Mackerell, A. D. Empirical Force Fields for Biological Macromolecules: Overview and Issues. *J. Comput. Chem.* **2004**, *25* (13), 1584–1604. <https://doi.org/10.1002/jcc.20082>.
- (47) Sato, F.; Hojo, S.; Sun, H. On the Transferability of Force Field Parameters - With an Ab Initio Force Field Developed for Sulfonamides. *J. Phys. Chem. A* **2003**, *107* (2), 248–257. <https://doi.org/10.1021/jp026612i>.
- (48) Guvench, O.; MacKerell, A. D. Comparison of Protein Force Fields for Molecular Dynamics Simulations. In *Molecular Modeling of Proteins*; Kukol, A., Ed.; Humana Press: Totowa, USA, 2008; pp 63–88. https://doi.org/10.1007/978-1-59745-177-2_4.
- (49) Andreas Kukol. *Molecular Modeling of Proteins*, Second.; Hatfield, UK, 2015.
- (50) Daun, K. J.; Sipkens, T. A.; Titantah, J. T.; Karttunen, M. Thermal Accommodation Coefficients for Laser-Induced Incandescence Sizing of Metal Nanoparticles in Monatomic Gases. *Appl. Phys. B Lasers Opt.* **2013**, *112* (3), 409–420. <https://doi.org/10.1007/s00340-013-5508-0>.

- (51) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters. *Proteins Struct. Funct. Bioinforma.* **2006**, 725 (June), 712–725. <https://doi.org/10.1002/prot>.
- (52) Kim, S. Issues on the Choice of a Proper Time Step in Molecular Dynamics. *Phys. Procedia* **2014**, 53, 60–62. <https://doi.org/10.1016/j.phpro.2014.06.027>.
- (53) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes. *J. Comput. Phys.* **1977**, 23 (3), 327–341. [https://doi.org/https://doi.org/10.1016/0021-9991\(77\)90098-5](https://doi.org/https://doi.org/10.1016/0021-9991(77)90098-5).
- (54) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. LINCS : A Linear Constraint Solver for Molecular Simulations. *J. Comput. Chem.* **1997**, 18 (12), 1463–1472.
- (55) Hess, B. P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. *J. Chem. Theory Comput.* **2008**, 4 (1), 116–122. <https://doi.org/10.1021/ct700200b>.
- (56) Phillips, J. C.; Hardy, D. J.; Maia, J. D. C.; Stone, J. E.; Ribeiro, J. V.; Bernardi, R. C.; Buch, R.; Fiorin, G.; Hénin, J.; Jiang, W.; McGreevy, R.; Melo, M. C. R.; Radak, B. K.; Skeel, R. D.; Singharoy, A.; Wang, Y.; Roux, B.; Aksimentiev, A.; Luthey-Schulten, Z.; Kalé, L. V.; Schulten, K.; Chipot, C.; Tajkhorshid, E. Scalable Molecular Dynamics on CPU and GPU Architectures with NAMD. *J. Chem. Phys.* **2020**, 153 (4), 1–33. <https://doi.org/10.1063/5.0014475>.
- (57) James Abraham, M.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS : High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, 2, 19–25. <https://doi.org/10.1016/j.softx.2015.06.001>.
- (58) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber Biomolecular

- Simulation Programs. *J. Comput. Chem.* **2005**, *26* (16), 1668–1688. <https://doi.org/10.1002/jcc.20290>.
- (59) Dror, R. O.; Dirks, R. M.; Grossman, J. P.; Xu, H.; Shaw, D. E. Biomolecular Simulation: A Computational Microscope for Molecular Biology. *Annu. Rev. Biophys.* **2012**, *41* (1), 429–452. <https://doi.org/10.1146/annurev-biophys-042910-155245>.
- (60) Shaw, D. E.; Dror, R. O.; Salmon, J. K.; Grossman, J. P.; Mackenzie, K. M.; Bank, J. A.; Young, C.; Deneroff, M. M.; Batson, B.; Bowers, K. J.; Chow, E.; Eastwood, M. P.; Ierardi, D. J.; Klepeis, J. L.; Kuskin, J. S.; Larson, R. H.; Lindorff-Larsen, K.; Maragakis, P.; Moraes, M. A.; Piana, S.; Shan, Y.; Towles, B. Millisecond-Scale Molecular Dynamics Simulations on Anton. In *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis; SC '09*; Association for Computing Machinery: New York, NY, USA, 2009. <https://doi.org/10.1145/1654059.1654099>.
- (61) Murtola, T.; Bunker, A.; Vattulainen, I.; Deserno, M.; Karttunen, M. Multiscale Modeling of Emergent Materials: Biological and Soft Matter. *Phys. Chem. Chem. Phys.* **2009**, *11* (12), 1869–1892. <https://doi.org/10.1039/b818051b>.
- (62) Kmiecik, S.; Gront, D.; Kolinski, M.; Wieteska, L.; Dawid, A. E.; Kolinski, A. Coarse-Grained Protein Models and Their Applications. *Chem. Rev.* **2016**, *116* (14), 7898–7936. <https://doi.org/10.1021/acs.chemrev.6b00163>.
- (63) Riniker, S.; Allison, J. R.; Van Gunsteren, W. F. On Developing Coarse-Grained Models for Biomolecular Simulation: A Review. *Phys. Chem. Chem. Phys.* **2012**, *14* (36), 12423–12430. <https://doi.org/10.1039/c2cp40934h>.
- (64) Lyubartsev, A. P.; Laaksonen, A. Calculation of Effective Interaction Potentials from Radial Distribution Functions: A Reverse Monte Carlo Approach. *Phys. Rev. E* **1995**, *52* (4), 3730–3737.
- (65) Faller, R.; Schmitz, H.; Biermann, O.; Müller-Plathe, F. Automatic Parameterization

- of Force Fields for Liquids by Simplex Optimization. *J. Comput. Chem.* **1999**, *20* (10), 1009–1017. [https://doi.org/10.1002/\(SICI\)1096-987X\(19990730\)20:10<1009::AID-JCC3>3.0.CO;2-C](https://doi.org/10.1002/(SICI)1096-987X(19990730)20:10<1009::AID-JCC3>3.0.CO;2-C).
- (66) Alessandri, R.; Souza, P. C. T.; Thallmair, S.; Melo, M. N.; De Vries, A. H.; Marrink, S. J. Pitfalls of the Martini Model. *J. Chem. Theory Comput.* **2019**, *15* (10), 5448–5460. <https://doi.org/10.1021/acs.jctc.9b00473>.
- (67) Berau, T.; Deserno, M. Generic Coarse-Grained Model for Protein Folding and Aggregation. *J. Chem. Phys.* **2009**, *130* (23). <https://doi.org/10.1063/1.3152842>.
- (68) Henderson, R. L. A Uniqueness Theorem for Fluid Pair Correlation Functions. *Phys. Lett. A* **1974**, *49* (3), 197–198. [https://doi.org/10.1016/0375-9601\(74\)90847-0](https://doi.org/10.1016/0375-9601(74)90847-0).
- (69) Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S. J. The MARTINI Coarse-Grained Force Field: Extension to Proteins. *J. Chem. Theory Comput.* **2008**, *4* (5), 819–834. <https://doi.org/10.1021/ct700324x>.
- (70) Souza, P. C. T.; Alessandri, R.; Barnoud, J.; Thallmair, S.; Faustino, I.; Grünewald, F.; Patmanidis, I.; Abdizadeh, H.; Bruininks, B. M. H.; Wassenaar, T. A.; Kroon, P. C.; Melcr, J.; Nieto, V.; Corradi, V.; Khan, H. M.; Domański, J.; Javanainen, M.; Martinez-Seara, H.; Reuter, N.; Best, R. B.; Vattulainen, I.; Monticelli, L.; Periole, X.; Tieleman, D. P.; de Vries, A. H.; Marrink, S. J. Martini 3: A General Purpose Force Field for Coarse-Grained Molecular Dynamics. *Nat. Methods* **2021**, *18* (4), 382–388. <https://doi.org/10.1038/s41592-021-01098-3>.
- (71) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; Vries, A. H. De. The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations. *J. Phys. Chem. B* **2007**, *111* (June), 7812–7824. <https://doi.org/10.1021/jp071097f>.
- (72) Clementi, C. Coarse-Grained Models of Protein Folding: Toy Models or Predictive Tools? *Curr. Opin. Struct. Biol.* **2008**, *18* (1), 10–15. <https://doi.org/10.1016/j.sbi.2007.10.005>.
- (73) Bond, P. J.; Sansom, M. S. P. Insertion and Assembly of Membrane Proteins via

Simulation. *J. Am. Chem. Soc.* **2006**, *128* (8), 2697–2704.
<https://doi.org/10.1021/ja0569104>.

3 About this thesis

3.1 Significance and aims

The primary aim of this work was to improve the understanding of the applicability of different computational techniques to study globular and IDP proteins. To accomplish this, two proteins were chosen as study targets: Triosephosphate isomerase (TIM) and Methyl CpG binding protein 2 (MeCP2). Both proteins are targets for drug design.

TIM is an extensively studied globular protein and is often considered a “textbook” protein. Its first crystal structure revealed the TIM barrel topology for the first time, which is an eightfold repeat of ($\beta\alpha$) units¹. This structural motif is the most common enzyme fold and is present in ~10% of all known proteins¹⁻³. There is plenty of experimental data available on this protein as well as a few computational studies⁴⁻¹². TIM is essential for maintaining life under anaerobic conditions and has been used as target for antiparasitic drugs¹³⁻¹⁵.

In contrast, MeCP2 is a protein whose full-length structure is not even known. The only available structure contains solely ~17% of its amino acids¹⁶. MeCP2 was the first of the methylated DNA binding protein (MBP) family to be identified¹⁷. It selectively binds CpG dinucleotides and mediates transcriptional repression through interaction with histone deacetylase and the corepressor SIN3A^{18,19}. The malfunction of this protein causes the neurodevelopmental disorder Rett syndrome^{20,21}. Less is known about MeCP2 structure compared to its functions because it is an intrinsically disordered protein (IDP). Circular dichroism has shown that almost 60% of the protein is unstructured²², and five out of its six domains lack a stable tertiary structure²³. The knowledge gap between structure and function of this protein could potentially be bridged using computer simulations.

The findings should be useful for improving the general understanding of the use of computational techniques for the study of biomolecules, as well as providing new insights into the structure of TIM and MeCP2.

3.2 Thesis outline

Chapters 4 and 6 represent individual, first author publications, in their unmodified forms. While chapter 5 is not a first author publication, I carried out every computational component of the work.

Chapter 4 describes a study of the globular protein TIM. All-atom simulations of three TIM proteins (TcTIM, TbTIM and a chimeric protein) were performed. The residue interaction networks of the amino acids in these trajectories were analyzed, as well as their electrostatic interactions and the impact of simulation length on them. Our findings provide new insights on the mechanisms that give rise to the different biophysical behaviors of these highly similar proteins and underline the importance of long simulations.

The study of MeCP2 starts in chapter 5. We present simulations of the MBD domain in solution and in the presence of a surface, in order to compare them with the experimental setup of high-speed atomic force microscopy (HS-AFM). Chapter 6 contains the next step, which was to complete the rest of the protein by *ab initio* modelling. Since an all-atom simulation of this model is not enough to guarantee adequate sampling of its conformational space, coarse-grained modeling was used to complement the atomistic picture. The coarse-grained simulations sampled a conformation that had not been observed in the all-atom simulation but that was in good agreement with HS-AFM data. Together, chapters 5 and 6 provide a detailed conformational ensemble of MeCP2, which is compatible with experimental data and can be the basis of further studies.

The major conclusions from this thesis and possible future directions are discussed in chapter 7.

3.3 References

- (1) Wierenga, R. K.; Kapetaniou, E. G.; Venkatesan, R. Triosephosphate Isomerase: A Highly Evolved Biocatalyst. *Cell. Mol. Life Sci.* **2010**, *67* (23), 3961–3982. <https://doi.org/10.1007/s00018-010-0473-9>.

- (2) Wierenga, R. K. The TIM-Barrel Fold: A Versatile Framework for Efficient Enzymes. *FEBS Lett.* **2001**, *492*, 193–198. [https://doi.org/10.1016/s0014-5793\(01\)02236-0](https://doi.org/10.1016/s0014-5793(01)02236-0).
- (3) Roland, B. P.; Stuchul, K. A.; Larsen, S. B.; Amrich, C. G.; VanDemark, A. P.; Celotto, A. M.; Palladino, M. J. Evidence of a Triosephosphate Isomerase Non-Catalytic Function Crucial to Behavior and Longevity. *J. Cell Sci.* **2013**, *126* (14), 3151–3158. <https://doi.org/10.1242/jcs.124586>.
- (4) Garza-Ramos, G.; Cabrera, N.; Saavedra-Lira, E.; Tuena De Gómez-Puyou, M.; Ostoa-Saloma, P.; Pérez-Montfort, R.; Gómez-Puyou, A. Sulfhydryl Reagent Susceptibility in Proteins with High Sequence Similarity: Triosephosphate Isomerase from *Trypanosoma Brucei*, *Trypanosoma Cruzi* and *Leishmania Mexicana*. *Eur. J. Biochem.* **1998**, *253* (3), 684–691. <https://doi.org/10.1046/j.1432-1327.1998.2530684.x>.
- (5) García-Torres, I.; Cabrera, N.; Torres-Larios, A.; Rodríguez-Bolaños, M.; Díaz-Mazariegos, S.; Gómez-Puyou, A.; Perez-Montfort, R. Identification of Amino Acids That Account for Long-Range Interactions in Two Triosephosphate Isomerases from Pathogenic Trypanosomes. *PLoS One* **2011**, *6* (4). <https://doi.org/10.1371/journal.pone.0018791>.
- (6) Zomosa-Signoret, V.; Hernández-Alcántara, G.; Reyes-Vivas, H.; Martínez-Martínez, E.; Garza-Ramos, G.; Pérez-Montfort, R.; De Gómez-Puyou, M. T.; Gómez-Puyou, A. Control of the Reactivation Kinetics of Homodimeric Triosephosphate Isomerase from Unfolded Monomers. *Biochemistry* **2003**, *42* (11), 3311–3318. <https://doi.org/10.1021/bi0206560>.
- (7) Reyes-Vivas, H.; Martínez-Martínez, E.; Mendoza-Hernández, G.; López-Velázquez, G.; Pérez-Montfort, R.; Tuena De Gómez-Puyou, M.; Gómez-Puyou, A. Susceptibility to Proteolysis of Triosephosphate Isomerase from Two Pathogenic Parasites: Characterization of an Enzyme with an Intact and a Nicked Monomer. *Proteins Struct. Funct. Genet.* **2002**, *48* (3), 580–590.

<https://doi.org/10.1002/prot.10179>.

- (8) Rodríguez-Bolaños, M.; Cabrera, N.; Perez-Montfort, R. Identification of the Critical Residues Responsible for Differential Reactivation of the Triosephosphate Isomerases of Two Trypanosomes. *Open Biol.* **2016**, *6* (10). <https://doi.org/10.1098/rsob.160161>.
- (9) Vázquez-Raygoza, A.; Cano-González, L.; Velázquez-Martínez, I.; Trejo-Soto, P. J.; Castillo, R.; Hernández-Campos, A.; Hernández-Luis, F.; Oria-Hernández, J.; Castillo-Villanueva, A.; Avitia-Domínguez, C.; Sierra-Campos, E.; Valdez-Solana, M.; Téllez-Valencia, A. Species-Specific Inactivation of Triosephosphate Isomerase from *Trypanosoma Brucei*: Kinetic and Molecular Dynamics Studies. *Molecules* **2017**, *22* (12). <https://doi.org/10.3390/molecules22122055>.
- (10) Dantu, S. C.; Groenhof, G. Conformational Dynamics of Active Site Loops 5, 6 and 7 of Enzyme Triosephosphate Isomerase: A Molecular Dynamics Study. *bioRxiv* **2018**. <https://doi.org/10.1101/459198>.
- (11) Liao, Q.; Kulkarni, Y.; Sengupta, U.; Petrović, D.; Mulholland, A. J.; Van Der Kamp, M. W.; Strodel, B.; Kamerlin, S. C. L. Loop Motion in Triosephosphate Isomerase Is Not a Simple Open and Shut Case. *J. Am. Chem. Soc.* **2018**, *140* (46), 15889–15903. <https://doi.org/10.1021/jacs.8b09378>.
- (12) Cansu, S.; Doruker, P. Dimerization Affects Collective Dynamics of Triosephosphate Isomerase. *Biochemistry* **2008**, *47* (5), 1358–1368. <https://doi.org/10.1021/bi701916b>.
- (13) Gómez-Puyou, A.; Saavedra-Lira, E.; Becker, I.; Zubillaga, R. A.; Rojo-Domínguez, A.; Perez-Montfort, R. Using Evolutionary Changes to Achieve Species-Specific Inhibition of Enzyme Action - Studies with Triosephosphate Isomerase. *Chem. Biol.* **1995**, *2* (12), 847–855. [https://doi.org/10.1016/1074-5521\(95\)90091-8](https://doi.org/10.1016/1074-5521(95)90091-8).
- (14) Velanker, S. S.; Ray, S. S.; Gokhale, R. S.; Suma, S.; Balaram, H.; Balaram, P.;

- Murthy, M. R. N. Triosephosphate Isomerase from *Plasmodium Falciparum*: The Crystal Structure Provides Insights into Antimalarial Drug Design. *Structure* **1997**, *5* (6), 751–761. [https://doi.org/10.1016/S0969-2126\(97\)00230-X](https://doi.org/10.1016/S0969-2126(97)00230-X).
- (15) Téllez-Valencia, A.; Olivares-Illana, V.; Hernández-Santoyo, A.; Pérez-Montfort, R.; Costas, M.; Rodríguez-Romero, A.; López-Calahorra, F.; Tuena De Gómez-Puyou, M.; Gómez-Puyou, A. Inactivation of Triosephosphate Isomerase from *Trypanosoma Cruzi* by an Agent That Perturbs Its Dimer Interface. *J. Mol. Biol.* **2004**, *341* (5), 1355–1365. <https://doi.org/10.1016/j.jmb.2004.06.056>.
- (16) Wakefield, R. I. D.; Smith, B. O.; Nan, X.; Free, A.; Soteriou, A.; Uhrin, D.; Bird, A. P.; Barlow, P. N. The Solution Structure of the Domain from MeCP2 That Binds to Methylated DNA. *J. Mol. Biol.* **1999**, *291* (5), 1055–1065. <https://doi.org/10.1006/jmbi.1999.3023>.
- (17) Lewis, J. D.; Meehan, R. R.; Henzel, W. J.; Maurer-Fogy, I.; Jeppesen, P.; Klein, F.; Bird, A. Purification, Sequence, and Cellular Localization of a Novel Chromosomal Protein That Binds to Methylated DNA. *Cell* **1992**, *69* (6), 905–914. [https://doi.org/10.1016/0092-8674\(92\)90610-O](https://doi.org/10.1016/0092-8674(92)90610-O).
- (18) Nan, X.; Ng, H.; Johnson, C. A.; Laherty, C. D.; Turner, B. M.; Eisenman, R. N.; Bird, A. Transcriptional Repression by the Methyl-CpG-Binding Protein MeCP2 Involves a Histone Deacetylase Complex. **1998**, *393* (May), 386–389.
- (19) Jones, P. L.; Veenstra, G. J. C.; Wade, P. A.; Vermaak, D.; Kass, S. U.; Landsberger, N.; Strouboulis, J.; Wolffe, A. P. Methylated DNA and MeCP2 Recruit Histone Deacetylase to Repress Transcription. *Nat. Genet.* **1998**, *19* (2), 187–191. <https://doi.org/10.1038/561>.
- (20) Aimer, R.; Van der Veyver, I.; Wan, M.; Tran, C.; Francke, U.; Zoghbi, H. Rett Syndrome Is Caused by Mutations in X-Linked MECP2 Encoding Methyl-CpG-Binding Protein 2. *Nat. Genet* **1999**, *23* (october), 185–188.
- (21) Hagberg, B. Rett's Syndrome: Prevalence and Impact on Progressive Severe Mental

Retardation in Girls. *Acta Pædiatrica* **1985**, *74* (3), 405–408.
<https://doi.org/10.1111/j.1651-2227.1985.tb10993.x>.

- (22) Hite, K. C.; Adams, V. H.; Hansen, J. C. Recent Advances in MeCP2 Structure and Function. *Biochem. Cell Biol.* **2009**, *87* (1), 219–227. <https://doi.org/10.1139/O08-115>.
- (23) Hite, K. C.; Kalashnikova, A. A.; Hansen, J. C. Coil-to-Helix Transitions in Intrinsically Disordered Methyl CpG Binding Protein 2 and Its Isolated Domains. *Protein Sci.* **2012**, *21* (4), 531–538. <https://doi.org/10.1002/pro.2037>.

4 Highly similar sequence and structure yet different biophysical behaviour: A computational study of two triosephosphate isomerases

Cecilia Chávez-García^{1,2} and Mikko Karttunen^{*1,2,3}

¹Department of Chemistry, the University of Western Ontario, 1151 Richmond Street, London, Ontario N6A 5B7, Canada; ²The Centre of Advanced Materials and Biomaterials Research, the University of Western Ontario, 1151 Richmond Street, London, Ontario, N6A 5B7, Canada; ³Department of Physics and Astronomy, the University of Western Ontario, 1151 Richmond Street, London, Ontario, N6A 3K7, Canada

*Corresponding author: mikko.karttunen@uwo.ca

Link: <https://www.biorxiv.org/content/10.1101/2021.10.13.464197v1.full.pdf>

Submission ID: ci-2021-01245m

4.1 Abstract

Homodimeric triosephosphate isomerases (TIM) from *Trypanosoma cruzi* (TcTIM) and *Trypanosoma brucei* (TbTIM) have a markedly similar amino acid sequences and three-dimensional structures. However, several of their biophysical parameters, such as their susceptibility to sulfhydryl agents and their reactivation speed after being denatured, have significant differences. The causes of these differences were explored with microsecond-scale molecular dynamics (MD) simulations of three different TIM proteins: TcTIM, TbTIM and a chimeric protein, Mut1. We examined their electrostatic interactions and explored the impact of simulation length on them. The same salt bridge between catalytic residues Lys 14 and Glu 98 was observed in all three proteins, but key differences were found in other interactions that the catalytic amino acids form. In particular, a cation- π interaction between catalytic amino acids Lys 14 and His 96, and both a salt bridge and a hydrogen bond between catalytic Glu168 and residue Arg100, were only observed in TcTIM. Furthermore, although TcTIM forms less hydrogen bonds than TbTIM and Mut1, its hydrogen bond network spans almost the entire protein, connecting the residues in both monomers. This work provides new insight on the mechanisms that give rise to the different behaviour of these proteins. The results also show the importance of long simulations.

4.2 Introduction

One of the most common structural motifs in proteins is the triosephosphate isomerase (TIM) barrel, which is present in ~10% of all known proteins and is the most common enzyme fold in the Protein Data Bank (PDB) database¹⁻⁴. TIM is an enzyme which takes part in the fifth step of glycolysis by interconverting glyceraldehyde 3-phosphate into dihydroxyacetone phosphate. The TIM barrel consists of an eightfold repeat of ($\beta\alpha$) units in such a way that β -strands in the inside are surrounded by α -helices on the outside. TIM is present in almost all organisms and is usually found as a dimer, although it can form a tetramer in some extremophile organisms⁵⁻⁸. It is completely active only in the dimeric form even though each monomer contains the catalytic residues (N12, K14, H96 and E168). The catalytic residues are strictly conserved throughout the whole TIM family^{1,9,10}.

TIM is essential for maintaining life under anaerobic conditions and, consequently, it has been used as a target for drug design when dealing with human parasites¹¹⁻¹³.

There are multiple instances in which homologous enzymes with high similarity have significant differences in their biophysical parameters¹⁴⁻¹⁷. This is exemplified by the triosephosphate isomerases of *Trypanosoma cruzi* (TcTIM), the parasite that causes Chagas' disease, and *Trypanosoma brucei* (TbTIM), causative agent of the African sleeping sickness. They have a sequence identity of 73.9% and a sequence similarity of 92.4%¹⁸ (Fig. 4.1). Previous works have found significant differences in their susceptibility to sulfhydryl agents, reactivation speed after being denatured with chemical agents such as guanidine hydrochloride, and their proteolysis susceptibility with subtilisin^{16,18-20}. Interestingly, a study of TcTIM and TbTIM by Rodríguez-Bolaños et al.²¹ showed that it is sufficient to mutate 13 amino acids on TbTIM to obtain TcTIM-like behaviour in reactivation experiments. Circular dichroism indicated that the chimeric proteins had the same fold as the native, however, the role that these mutations have on the structure and dynamics of the proteins is not well understood.

For many years, studies have focused on the interaction between TIM proteins and benzothiazoles, which have been found to deactivate the enzyme^{13,22,23}. While there is a plethora of experimental data, there are very few MD simulations of TcTIM and TbTIM in the absence of ligands. Some of the simulations are only 40-60 ns long, and thus cannot capture phenomena that occur in longer timescales²⁴⁻²⁶. There is only one study in the microsecond scale. In that Dantu and Groenhof use a combination of QM/MM and crystal unit cell simulations²⁷. However, this study focuses on the effect of binding of substrates in loops 5, 6 and 7. The most comprehensive study on TIM so far used conventional MD and enhanced sampling techniques to characterize the motion of loops 6 and 7²⁸. This study showed that loop 6 does not follow a simple two-state rigid-body transition as previously thought. However, it did not explore in detail the interactions between the two monomers. The simulations that report the root-mean-square fluctuation (RMSF) of TIM proteins have consistently found the highest RMSF values in loops 5 and 6^{22,24,25}, which is in good agreement with our work.

Here we present microsecond-scale all-atom simulations of TcTIM, TbTIM and a chimeric TbTIM with the 13 mutations as identified by Rodríguez-Bolaños *et al*²¹. A conserved salt bridge between catalytic residues Lys 14 and Glu 98^{29,30} was observed in all three proteins. In contrast, a cation- π interaction between catalytic amino acids Lys 14 and His 96, and both a salt bridge and a hydrogen bond between catalytic Glu 168 and Arg 100 were only observed in TcTIM. Furthermore, TcTIM and TbTIM exhibited different hydrogen bond networks, with the chimeric protein behaving similar to TbTIM. The hydrogen bond network observed in these proteins helps to explain why regions 1, 4 and 8 become more rigid when the dimer is formed.

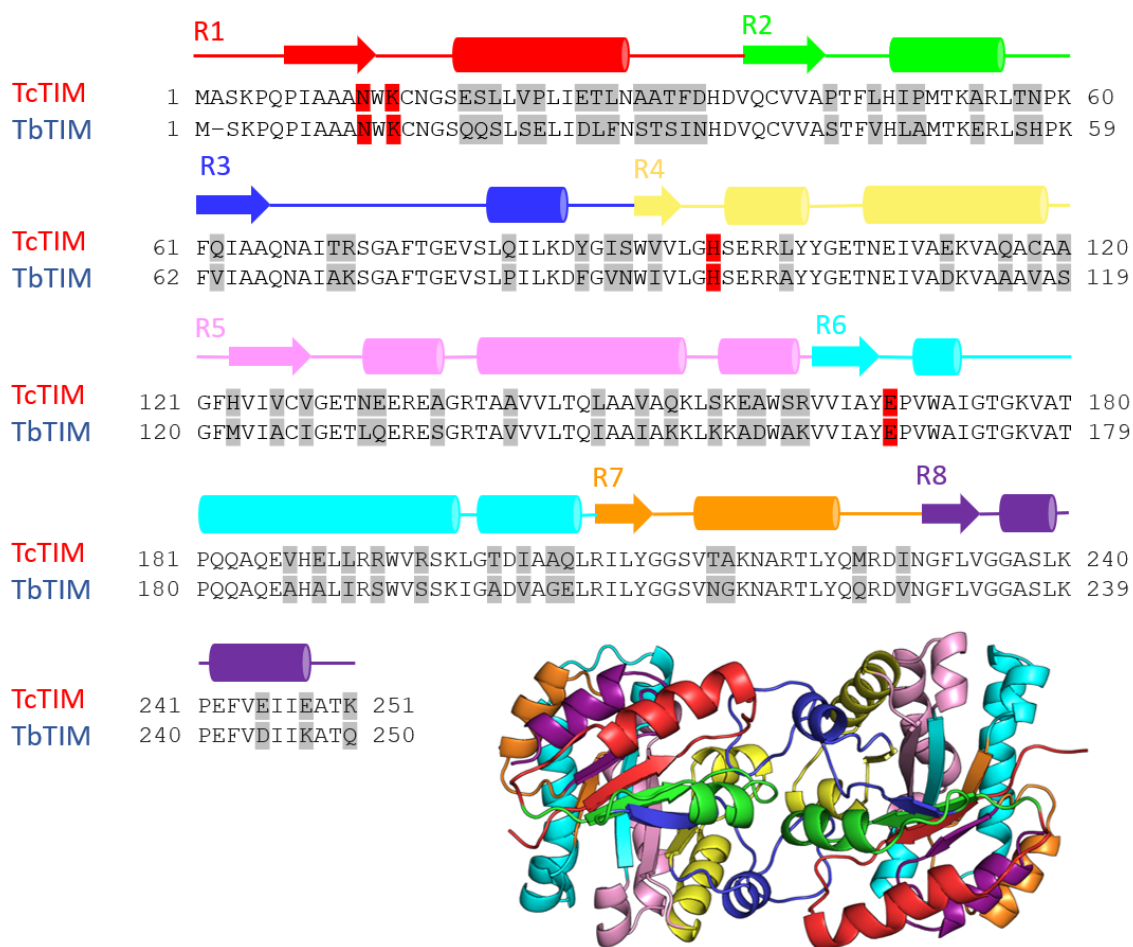


Figure 4. 1 Alignment of TcTIM and TbTIM sequences with the secondary structure elements marked as lines (loops), β -strands (arrows) and α -helices (cylinders). The colors in each motif correspond to the 3D structure below it. The differences in the amino acids are highlighted in gray and the catalytic residues in red.

4.3 Materials and methods

All-atom MD simulations were performed on three TIM proteins: 1) the TIM from *T. brucei* (TbTIM), 2) the TIM from *T. cruzi* (TcTIM), and 3) a chimeric protein: [TcTIM:2; TbTIM:1,3-8; Q18E, E23P, D26E, S32T, I33F, N34D] henceforth referred to as Mut1. Each protein was simulated as a dimer. The initial structures were taken from the Protein Data Bank (PDB ids: 1TCD³¹ and 5TIM³²) and any missing side chains were completed using the whatif web server³³. Mut1 was built using TbTIM as template and mutating the required amino acids using Pymol³⁴.

Each protein was placed in a dodecahedral box in which the distance from the edges of the box to every atom in the protein was at least 1 nm. The box was solvated with explicit water and 150 mM of NaCl was added to reproduce physiological conditions. Counterions were added to keep the overall charge neutrality of the systems, 6 Cl⁻ ions for the TcTIM and 10 Cl⁻ ions for both 5TIM and Mut1. All simulations were performed using GROMACS 2016.3³⁵ with the TIP3P water model³⁶ and the CHARMM36m force field³⁷.

Each system was first energy minimized using the method of steepest descents and pre-equilibrated at constant particle number, temperature and volume, for 100 ps. The pre-equilibration was followed by a production run with a time step of 2 fs. The Lennard-Jones potential was truncated using a shift function between 1.0 and 1.2 nm. Electrostatic interactions were calculated using the particle-mesh Ewald method (PME)^{38,39} with a real space cut-off of 1.2 nm. The temperature was set to 310 K with the V-rescale algorithm⁴⁰ and pressure was kept at 1 atm using the Parrinello-Rahman barostat⁴¹. Bonds involving hydrogens were constrained using the Parallel Linear Constraint Solver (P-LINCS) algorithm⁴². The root-mean-square deviation (RMSD) was used to monitor equilibration. Since the RMSD for TcTIM kept increasing during the first microsecond, the simulations were extended to 3 μ s. We will return to this issue later in Results.

Trajectory analyses were performed using Gromacs built-in tools³⁵, MDAnalysis^{43,44} and the VMD plug-ins Salt bridges⁴⁵ and RIP-MD⁴⁶. RIP-MD generates residue interaction networks (RINs) from MD trajectory files. In a RIN, the nodes of the network represent amino-acid residues and the connections between them depict non-covalent interactions.

These include hydrogen bonds, salt bridges, cation- π , π - π , arginine-arginine, and Coulomb interactions. RIP-MD starts with a MD trajectory and the parameters defining the interactions as input. It then searches for interactions between all atoms in each snapshot of the trajectory. Finally, it generates a consensus RIN where edges exist if they are present in at least a given percentage of the snapshots. For our study, we used a 30% threshold.

4.4 Results

We performed MD simulations on three different TIM proteins: TcTIM, TbTIM and Mut1. Figure 4.2 shows that the RMSD of all three systems increase throughout the first 1500 ns, after which they stabilized. For this reason, the simulations were extended to 3 μ s and all reported averages were calculated during the last microsecond of the trajectories.

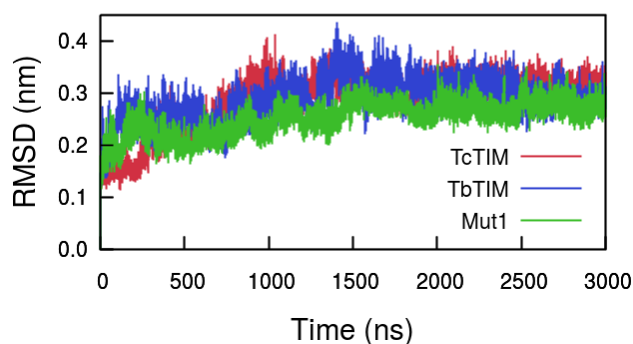


Figure 4.2 Root-mean-square deviation with respect to the crystal structure. RMSD increases throughout the first 1500 ns before stabilizing.

The root-mean-square fluctuations (RMSF) of most of the residues in the three proteins are very similar (Fig. 4.3). The main differences appear at the highest peaks, which are located at residues 133-137 (loop 5) and 173-178 (loop 6) in both monomers. The amino acids in loop 6 correspond to the catalytic loop. This loop has a “phosphate gripper” motif⁴⁷ which is likely engaged in substrate binding and product release, as its opening and closing motion has a rate constant that closely matches the turnover time for catalysis^{48,49}. TcTIM is the only protein with peaks at loops 5 and 6 in monomer A and TbTIM is the only protein with a peak in loop 5 monomer B. All three proteins have a peak in loop 6 monomer B, albeit

the peak in TbTIM is almost two times larger than in the other two systems. TcTIM is the only protein in which the first residues of monomer B have a very low RMSF value, indicating an interaction with monomer A.

In a previous computational study of TcTIM²⁵, loops 5 and 6 were reported to have the largest fluctuations. This study used GROMOS96 (43a2)⁵⁰, a united-atom force field and the SPC water model⁵¹. The fact that the same results were obtained with two very different force fields (GROMOS and CHARMM) underlines their robustness and their independence of the chosen force field. All the other main peaks correspond to amino acids located in loops with the exception of residues 236-239, which span the short helix in region 8.

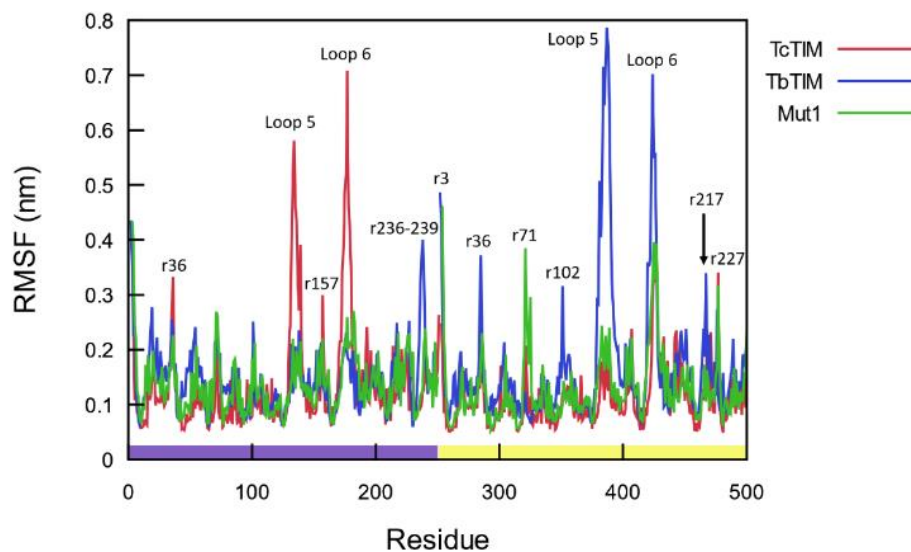


Figure 4.3 Root mean square fluctuations for the last microsecond. Since each protein was simulated as a dimer, the color bar at the bottom distinguishes the residues in monomer A (purple) from those in monomer B (yellow). The main peaks are located at loops 5 and 6 in each monomer.

The number of contacts between monomers, defined as amino acids whose C β atoms (C α for glycine) are within 0.8 nm distance, is shown in Figure 4.4. Even though a protein residue-residue contact is not uniquely defined, this definition captures all possible interactions between two residues and it has been used in a number of previous studies⁵²⁻⁵⁵. The number of contacts changes for both TcTIM and Mut1 during the first microsecond

before stabilizing. In contrast, the number of contacts between TbTIM's monomers fluctuated throughout the entire simulation. The average number of contacts between monomers during the last microsecond is 119 for TcTIM, 89 for TbTIM and 116 for Mut1. This is in good agreement with previous experiments, since the number of contacts between monomers can be related to its thermal stability and TcTIM has higher thermal stability than TbTIM²¹.

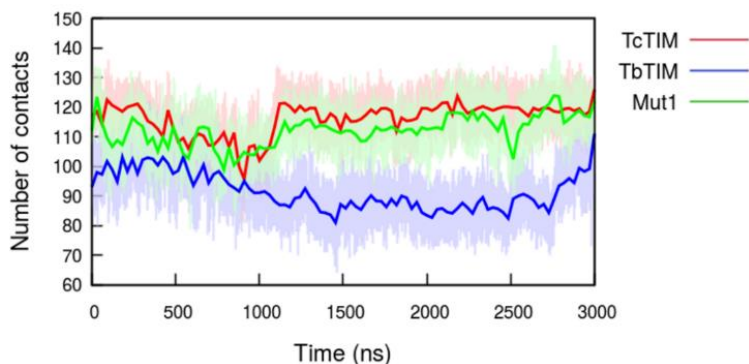


Figure 4.4 Number of contacts between monomers. Residues are considered to be in contact if their respective C β atoms (C α for glycine) are less than 0.8 nm apart. The number of contacts in TcTIM and Mut1 decreases over the first microsecond, increases over the next 200 ns and then stabilizes. In TbTIM, they decrease during the first half of the simulation and increase again after 1500 ns. Solid lines are B \acute{e} zier curves that interpolate the data.

In order to identify if the systems were in the open or closed conformations, the minimum distance between loop 6 (residues 170-180) and loop 7 (residues 211-216) was measured (Fig. 4.5, S4.1). Five different states were sampled in TcTIM, with distances 0.18, 0.28, 0.50, 0.67 and 0.80 nm. Only the first three states were observed in TbTIM and Mut1. It was previously reported that loop 6 can sample multiple conformational states with the tip of the loop moving \sim 0.7 nm between the fully open and fully closed conformations²⁸. This is in good agreement with the difference between the two extremes observed in the current TcTIM simulation. TbTIM fluctuated the most between states, and while Mut1 also showed many fluctuations, its loop in monomer B remained in the fully closed conformation for the last microsecond of the simulation. The cross-correlation between the open-close conformation of the two monomers decayed to zero during the simulation, indicating that the movement of these loops is independent between monomers.

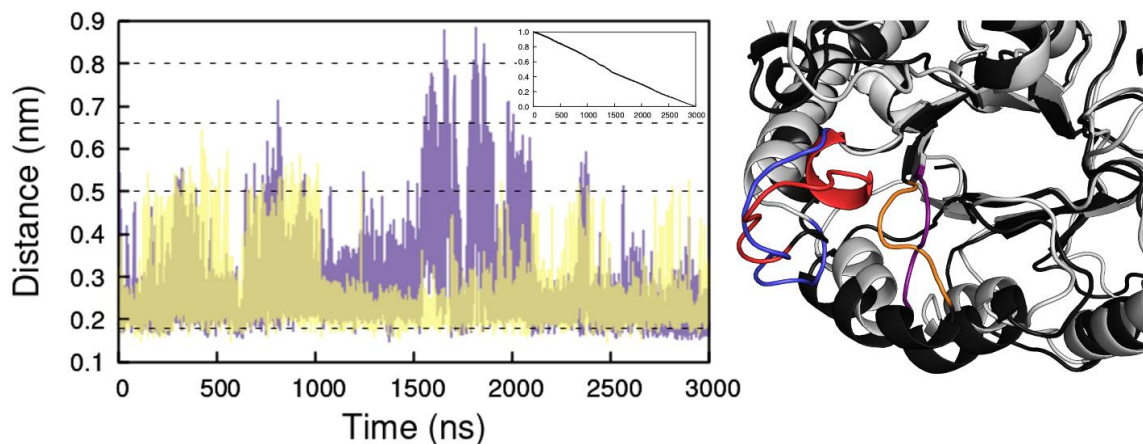


Figure 4.5 Minimum distance between loops 6 and 7 in TcTIM for monomer A (purple) and monomer B (yellow). Dashed lines mark the five different states sampled by the loops. Inset: cross-correlation between the loop state of the two monomers. Right: alignment of the open (gray) and closed (black) conformations. Loop 6 is shown in red for the closed conformation and in blue for the open conformation, loop 7 is shown in orange (closed) and purple (open).

Electrostatic interactions

We used the RIP-MD46 plugin for VMD⁴⁵ to analyze the last 2 μ s of each trajectory and to compute the following electrostatic interactions between all residues: salt bridges, cation- π interactions, π - π interactions and hydrogen bonds. No arginine-arginine interactions were found in the simulations. We will describe these interactions in the section below in more detail, but Table 1 summarizes the parameters defining them.

Table 4.1 Summary of interactions defined in RIP-MD

Hydrogen bonds	dist (donor, acceptor) \leq d $\theta(\overrightarrow{C-H}, \overrightarrow{acceptor}) \geq a$	d = 3 Å a = 120°
Salt bridges	Contacts between NH/NZ groups of Arg/Lys	d = 6 Å

and OE/OD in Asp/Glu $\leq d$		
Cation-π	dist (aromatic ring, cation) $\leq d$	$d = 6 \text{ \AA}$
interactions	$\theta (\overrightarrow{\text{normal vector ring}}, \overrightarrow{\text{ring center} - \text{cation}}) = a$	$a \in [0^\circ, 60^\circ]$ or $a \in [120^\circ, 180^\circ]$
π-π	dist (aromatic ring, aromatic ring) $\leq d$	$d = 7 \text{ \AA}$
interactions		
Arg-Arg	dist (guanidine, guanidine) $\leq d$	$d = 5 \text{ \AA}$

From the three proteins, only TcTIM had cation- π interactions (Fig. S4.2). These interactions were defined between the geometric center of the ring in the aromatic residue and the charged atom in the second residue, with a cutoff distance of 0.6 nm. This threshold was chosen because 99% of significant cation- π interactions occur within a distance of 0.6 nm⁵⁶. The only cation- π interaction between monomers occurs between amino acids Tyr 103 in monomer A and Arg 99 in monomer B. Figure 4.6 shows the distance that defines this cation- π interaction throughout the simulation. The distance fluctuates throughout time but remains within the limits that define the cation- π interaction during most frames in the last 1,500 ns of the simulation.

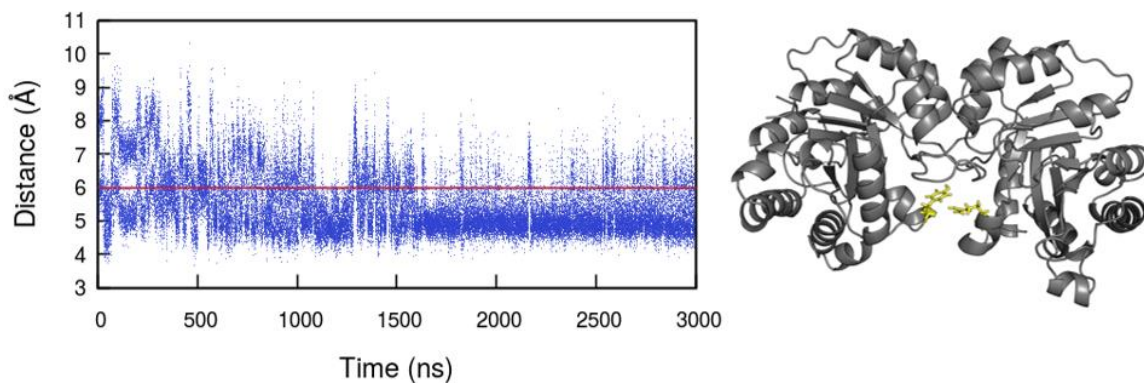


Figure 4.6 Cation- π interaction between Tyr 103 in monomer A and Arg 99 in monomer B in TcTIM. This interaction is defined by the distance between the geometric center of the aromatic residue in tyrosine and the charged atom in arginine, with a cutoff distance of 0.6 nm. This interaction was only observed in TcTIM. Right: these amino acids are located in region 4. A red line marks the threshold that defines the cation- π interaction.

Π - π interactions were defined with the distance between the geometric centers of the rings in the aromatic amino acids, with a cutoff distance of 0.7 nm. The distance between two interacting aromatic rings is geometry dependent and varies between 0.45 and 0.7 nm⁵⁷. All three systems presented π - π interactions (Fig. S4.3). Out of the 14 π - π interactions in TcTIM, four occurred at the interface between monomers. TbTIM only presented five π - π interactions and the mutant had four, two of which occurred at the interface. Interestingly, two of the π - π interactions in the mutant correspond to interactions in TcTIM and the other two to interactions in TbTIM.

Residue interaction networks

In RIP-MD⁴⁶, salt bridges are treated as a contact between two heavy atoms of opposite charge with a distance threshold of 0.6 nm^{46,58}. The same salt bridges were observed in the three proteins: between Glu 78 (Glu 77 in TbTIM) and Arg 99 (Arg 98 in TbTIM) of both monomers (Fig. S4.4). Residue Arg 55 forms a salt bridge with residue 27 from the same monomer, in both monomers in TcTIM. Since RIP-MD⁴⁶ does not provide the time dependence of the salt bridges, we used the Salt Bridge VMD plugin⁴⁵ to calculate them.

This plugin uses a different cutoff distance (0.32 nm) to define a salt bridge, however, as long as the interacting atoms are within the threshold in one frame, the program outputs the distance between them as a function of time. Figure 4.7 illustrates the fluctuations of the distance between atoms that form salt bridges. The interaction between Glu 27 and Arg 55 in TcTIM fluctuates considerably and the salt bridge is defined in only a portion of the frames. In contrast, the salt bridge between Glu 98 and Lys 14 stays well within the limit that defines this interaction throughout the whole simulation.

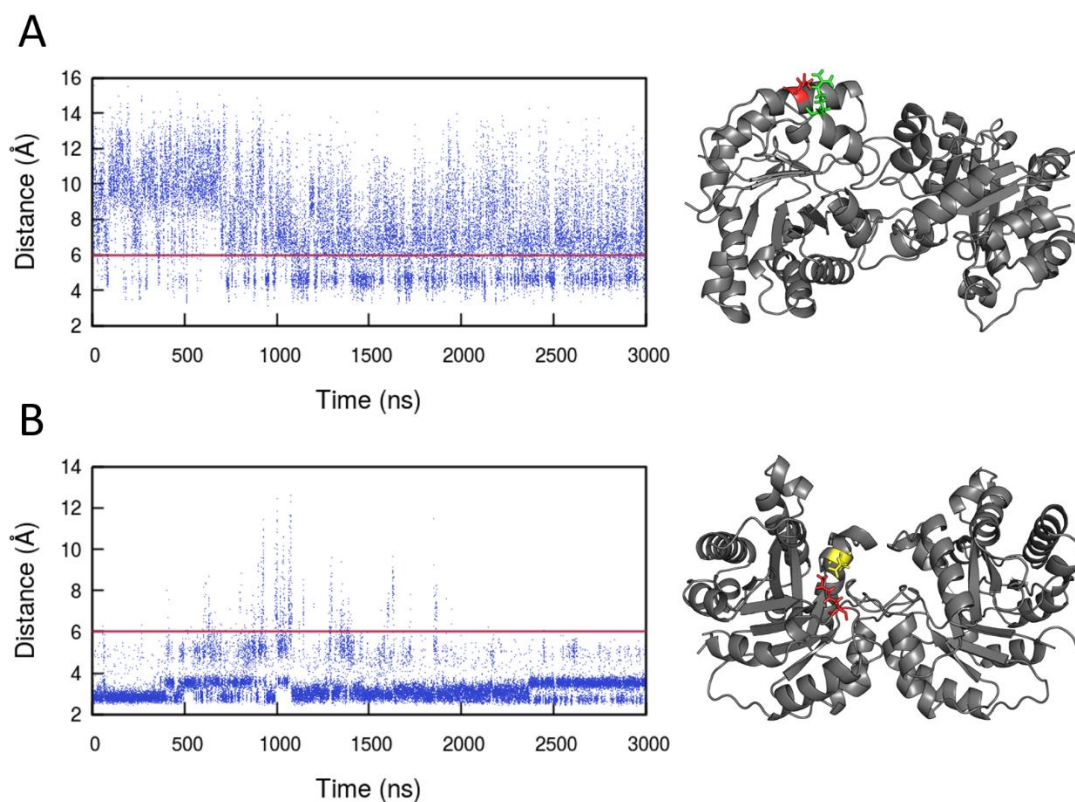


Figure 4.7 Salt bridges in TcTIM between Glu27 monomer A and Arg55 monomer A (A), and between Glu98 monomer A and Lys14 monomer A (B). These interactions were computed using the Salt Bridges plugin for VMD⁴⁵. Right: 3D location of the amino acids. Residues Glu27 and Lys14 are shown in red (region 1), Arg55 in green (region 2) and Glu98 in yellow (region 4). A red line marks the threshold that defines the salt bridge interaction.

Hydrogen bonds were defined using a cutoff radius of 0.3 nm for atoms whose acceptor-hydrogen-donor angle is greater than 120° ⁵⁹. The average number of hydrogen bonds over the last microsecond was 345 for TcTIM and $369 \pm 1\%$ for TbTIM and Mut1 (Fig. 4.8).

Figures S4.5-S4.15 show the hydrogen bond networks in the three proteins. For clarity, hydrogen bonds formed between neighboring amino acids (less than 4 residues apart) have been removed from the graphs. Even though TcTIM forms less hydrogen bonds than TbTIM, they connect amino acids in a network that involve more interactions between monomers and extends throughout the whole protein (Fig. S4.16).

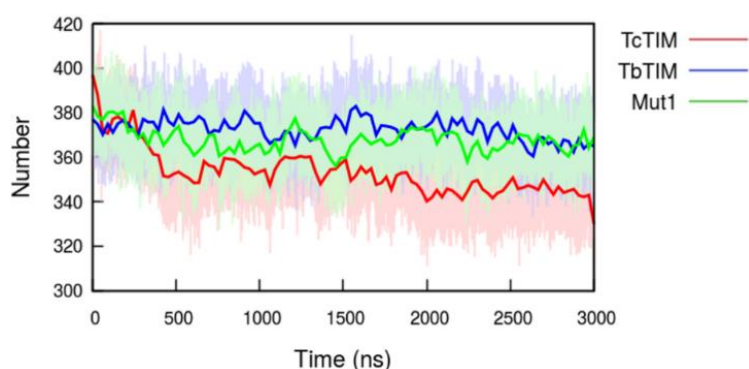


Figure 4.8 Total number of intramolecular hydrogen bonds in each simulation. Solid lines are Bézier curves that interpolate the data. Hydrogen bonds were defined with a cutoff radius of 0.3 nm for atoms whose acceptor-hydrogen-donor angle is greater than 120° .

Species-specific inhibition of TIMs can be achieved by targeting a non-conserved amino acid (Cys15) that lies at the dimer interface and which is important for catalysis¹¹. The susceptibility of TcTIM to thiol agents is approximately 40 times higher than that of TbTIM⁶⁰. Figure S4.17 shows the amino acids that form hydrogen bonds with Cys15 (Cys14 in TbTIM). Hydrogen bonds were determined with a cutoff angle of 30° for the hydrogen-donor-acceptor angle and a cutoff radius of 0.3 nm; OH and NH groups were regarded as donors, and O and N as acceptors. The residues that participate in hydrogen bonds with this cysteine change after the first microsecond in all three systems. In particular, those formed with residues Gly 73 and Ala 74 in TcTIM, with residue Ala 236 in TbTIM and with residues Phe 75 and Ser 80 in Mut1. Interestingly, TbTIM's Cys14 interacts with residues in region 8 from the same monomer, while TcTIM and Mut1's Cys15 interact with residues from region 3 in the other monomer. Fluctuations in the hydrogen bond network for Cys14/15 can only be noticed in simulations longer than 1 μ s.

Only the hydrogen bond between Cys15 monomer B and Phe75 monomer A in TcTIM was identified as such by RIP-MD⁴⁵, the other interactions were identified as C α contacts.

4.5 Discussion

Biological relevance

TIM has four catalytic residues: Asn12, Lys 14, His 96 and Glu 168^{2,10}. TcTIM is the only protein with cation- π interactions and one of them is between two of the catalytic residues: His 96 and Lys 14 (Fig. S4.2). Cation- π interactions can enhance binding energies by 2–5 kcal/mol, making them competitive with hydrogen bonds⁶¹. Another interaction only found in this protein occurs between catalytic Glu 168 and residue Arg 100, they form both a salt bridge and a hydrogen bond (Fig. S4.4, S4.5). In contrast, a salt bridge between catalytic Lys 14 (13 in TbTIM) and residue Glu 98 (97 in TbTIM) was observed in all three proteins (Fig. S4.4). This is a conserved salt bridge that has been observed in several crystal structures^{29,30}. Lys 14 is also involved in the main network of hydrogen bonds in TcTIM. It forms a hydrogen bond with its neighbor, catalytic Asn 12, which in turn forms a hydrogen bond with residue Thr 76 from the other monomer and with the other catalytic residue, His 96 (Fig. S4.5). Similarly, Mut1 forms a hydrogen bond between two of the catalytic residues, Lys 14 and Asn 12, which in turn forms a hydrogen bond with residue 76 from the other monomer. However, unlike TcTIM, these residues are not connected to others in the network (Fig. S4.9). In both Mut1 and TbTIM, catalytic Asn 12 (Asn 11 for TbTIM) forms a hydrogen bond with residue Val 234 (Val 233 in TbTIM) and Lys 14 (Lys 13 in TbTIM) forms one with Gly 236 (Gly 235 in TbTIM) (Fig. S4.10, S4.11, S4.13, S4.15). Catalytic Glu 168 forms three hydrogen bonds in TcTIM: with Arg 100, Val 128 and Glu 130, but it only forms the last two bonds in TbTIM and Mut1 (Fig. S4.5, S4.11, S4.12, S4.14, S4.15).

Regions 1, 4 and 8 are known to become more rigid when the dimer is formed³⁰. Region 1 forms several hydrogen bonds with region 3 of the opposite monomer. TcTIM forms six of these bonds at the interface, while TbTIM and Mut1 form only two (Fig. S4.5, S4.6,

S4.9). Of all regions, region 4 had the highest number of salt bridges, comprising 13 residues in TcTIM, 11 in TbTIM and 10 in Mut1 (Fig. S4.4). Salt bridges between Arg 99 (Arg 98 in TbTIM) and Glu 78 (Glu 77 in TbTIM) involving both monomers were found at the interface of all three proteins. Only TcTIM formed salt bridges in region 8, two in each monomer.

The same salt bridge between residues Arg 192 (Arg 191 in TbTIM) and Asp 228 (Asp 227 in TbTIM) was observed in the two native proteins but not in Mut1 (Fig. S4.4). Since it has been shown that this conserved bridge is important for the efficient folding of TIM⁶², we expect that Mut1 will have a low recovery of activity in denaturation and refolding experiments.

TcTIM had a high RMSF value in loops 5 and 6 monomer A, but a low value in the same loops in monomer B (Fig. 4.3). Five amino acids in monomer B are involved in salt bridges, but only one in monomer A. Three amino acids from region 6 form salt bridges in monomer A, and four in monomer B. One of them is catalytic Glu 168. Similarly, TbTIM forms more salt bridges in monomer A loop 5, than in monomer B. Six residues from loop 5 monomer A form salt bridges but only three in monomer B. Only one residue in loop 6 monomer B forms a salt bridge (Fig. S4.4). Mut1 has a similar number of residues from loop 5 involved in salt bridges in both monomers, which explains why neither one of them has a high RMSF value. Salt bridges can vary in strength from weak (0.5 kcal/mol) to strong (3–5 kcal/mol) and play an important role in structure stabilization⁶³. This may explain why the chimeric protein has a lower catalytic efficiency than its parent protein¹⁴.

The RMSF of TcTIM residues (Fig. 4.3) showed low values for the end of each monomer. This is explained by the network of hydrogen bonds (Fig. S4.7, S4.8) Amino acids at the end of the chain are connected to other residues forming a chain of hydrogen bonds that is not observed in the other two proteins. Furthermore, TcTIM has more interactions between monomers than the other two proteins. Some of these interactions were found in the mutant but not on TbTIM. This may explain why TcTIM has higher thermal stability even though it forms less hydrogen bonds than TbTIM²⁶. Based on this, we expect that Mut1 would have a thermal stability higher than TbTIM but lower than TcTIM.

The need for long simulations

While the RMSD helped to identify the need to increase the simulation time of all three systems, it is not the only quantity where this issue can be noticed. When the RMSF of the first 200 ns of the simulation is compared with the RMSF of the last 200 ns of each simulation, noticeable differences can be observed (Fig. 4.9, S4.18, S4.19). One of the two main peaks in TcTIM and in the RMSF of TbTIM do not even appear at the beginning of the simulations. Furthermore, fluctuations in many amino acids decrease, which help to highlight the relevance of the main peaks. In contrast, when one compares the last 200 ns of the simulation with the last microsecond of the trajectory, differences are considerably smaller. An exception to this appears in loop 6 monomer B of TbTIM, which is only observed when the average is calculated over the last microsecond of the simulation.

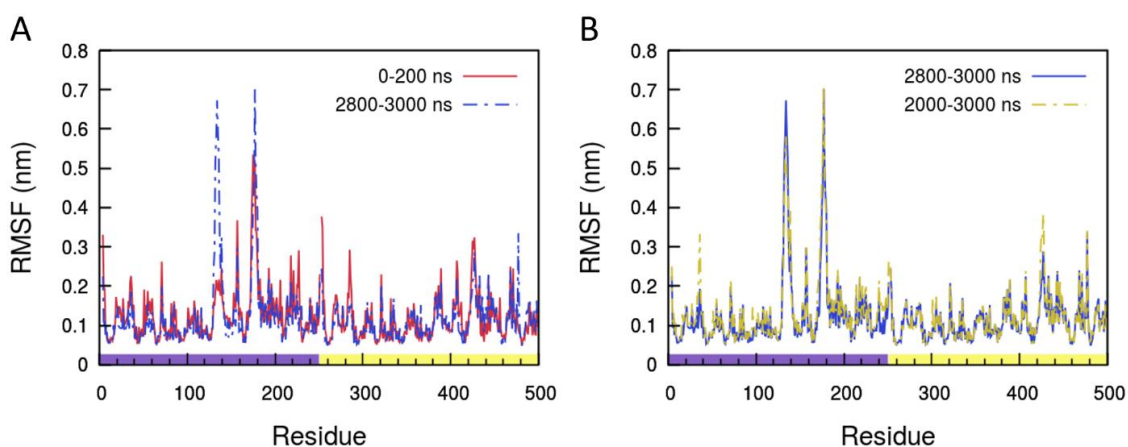


Figure 4.9 Changes over time in the root mean square fluctuations of the residues of TcTIM. A) RMSF of the first 200 ns of the simulation vs the last 200 ns, and B) RMSF of the last microsecond of the trajectory vs the last 200 ns. The peak in loop 5 of monomer A does not appear at the beginning of the trajectory and fluctuations in the minor peaks decrease at longer times. The color bar at the bottom of figure B distinguishes the residues in monomer A (purple) from those in monomer B (yellow).

One way to monitor convergence of simulations is to plot the average of a quantity over different time intervals. If the average changes with different time windows, then the system is not properly equilibrated. Figure 4.10 shows the number of intramolecular hydrogen bonds for the TcTIM trajectory. Horizontal lines mark the averages taken over

different time windows. The average for the first 500 ns is 7% higher than that of the last 500 ns. The average over the last microsecond of the trajectory equals the average over the last 500 ns (345 hydrogen bonds), which indicates that the trajectory has stabilized over the last microsecond.

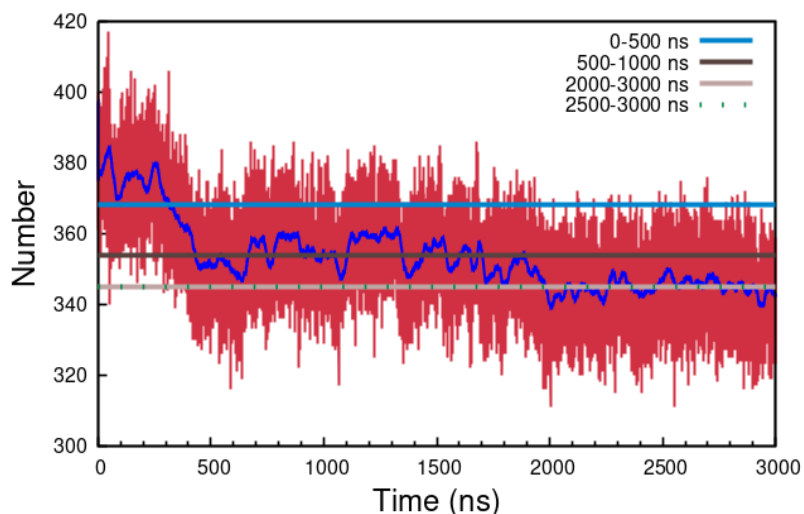


Figure 4.10 Number of hydrogen bonds in TcTIM. Convergence (horizontal lines) is reached around 2 μ s. Averaging over the last 500 ns or over the last microsecond of the simulation produces the same average (345 hydrogen bonds; the lines overlap). Blue: moving average.

In order to find how the hydrogen bond networks changed with time, we used the Gromacs built-in cluster tool to generate clusters for the first and last 500 ns of the simulations. The clusters were generated using the gromos method with a RMSD cutoff of 0.2 nm. We then used RIP-MD⁴⁶ to analyze the RINs in the representative structure of the most populated cluster. Figures S4.20 and S4.21 show the most significant changes in TcTIM. At the beginning of the simulation there is little connectivity between the main hydrogen bond networks of each monomer, when residues 75, 78 and 99 in monomer B interact with residues 15 and 103 in monomer A, these two networks merge into one (Fig. 4.11).

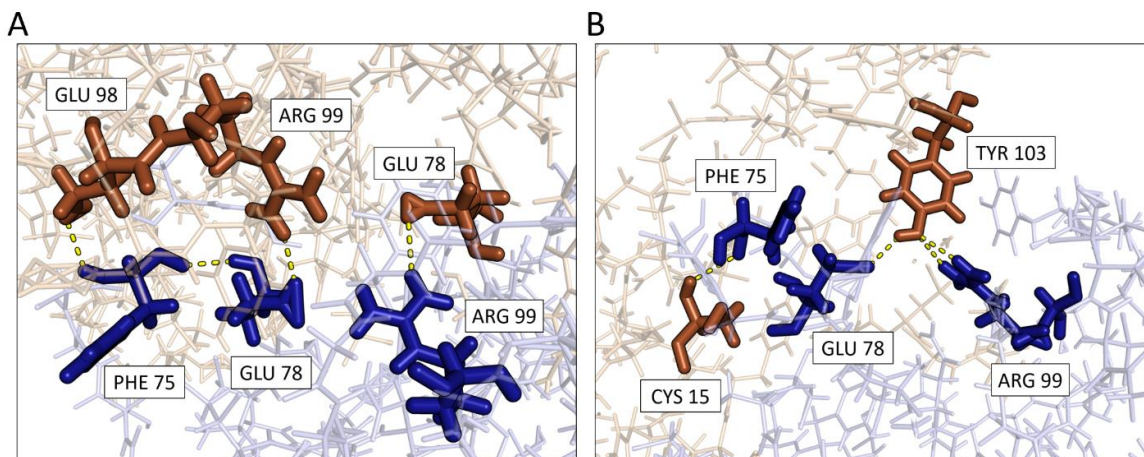


Figure 4.11 Changes in the hydrogen bond network of TcTIM. Interface between monomers (A) at the beginning and (B) at the end of the simulation. Residues Phe 75, Glu 78 and Arg 99 in monomer B (blue) change their interaction with residues in monomer A (orange).

Changes in protein dynamics are in turn reflected in the interactions between amino acids. Some interactions appeared after the first 500 ns of the simulation, e.g., the π - π interaction between residues 187 and 210 in monomer A of TbTIM, and others became more stable only after the first microsecond of the simulation, e.g., the π - π interaction between residues 36 and 224 of monomer B in TcTIM (Fig. 4.12).

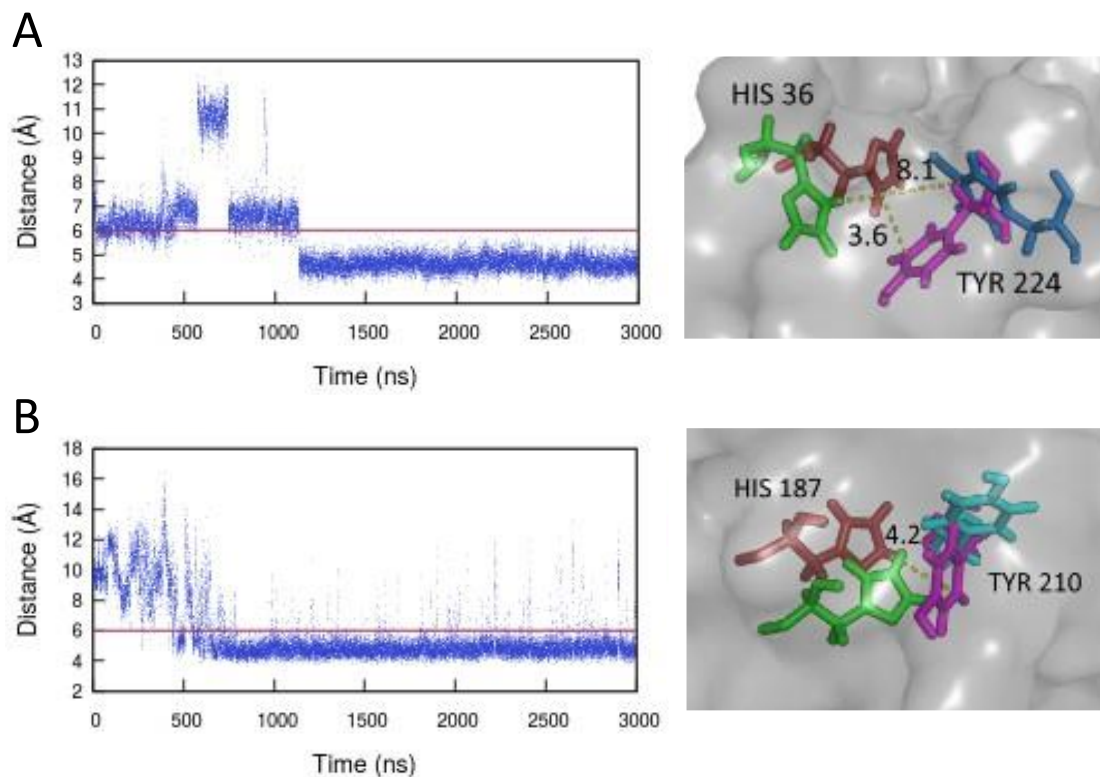


Figure 4.12 π - π interactions in TcTIM between residues His 36 and Tyr 224 of monomer B (A) and in TbTIM between residues His 187 and Tyr 210 in monomer A (B). A red line marks the threshold that defines the π - π interaction. At the right side of each graph, conformation changes between the first frame (green and blue) and the last frame (red and purple) of each simulation are shown. The interaction in TcTIM was not observed during the first 500 ns of the simulation. The interaction in TbTIM was observed since the beginning but it only become stable after the first microsecond.

4.6 Conclusions

We have performed molecular dynamics simulations on three different TIM proteins: TcTIM, TbTIM and a chimeric protein, Mut1. We examined the different electrostatic interactions that occur in these proteins: salt bridges, cation- π interactions, π - π interactions and hydrogen bonds, and also explored the impact of simulation length on them. Some of these interactions appeared only after the first microsecond of the simulation, and convergence of the number of hydrogen bonds was only reached in the last of the 3 μ s of

the simulation. Although TcTIM forms less hydrogen bonds than TbTIM and Mut1, they form a network that spans almost the entire protein, connecting the residues in both monomers. Key differences were found in the interactions that the catalytic amino acids form, such as a cation- π interaction between catalytic amino acids Lys 14 and His 96, only observed in TcTIM, but a salt bridge between catalytic residue Lys 14 and Glu 98 was observed in all three proteins. Further experiments will be required to confirm our hypothesis on the thermal stability of Mut1.

4.7 Supplemental information

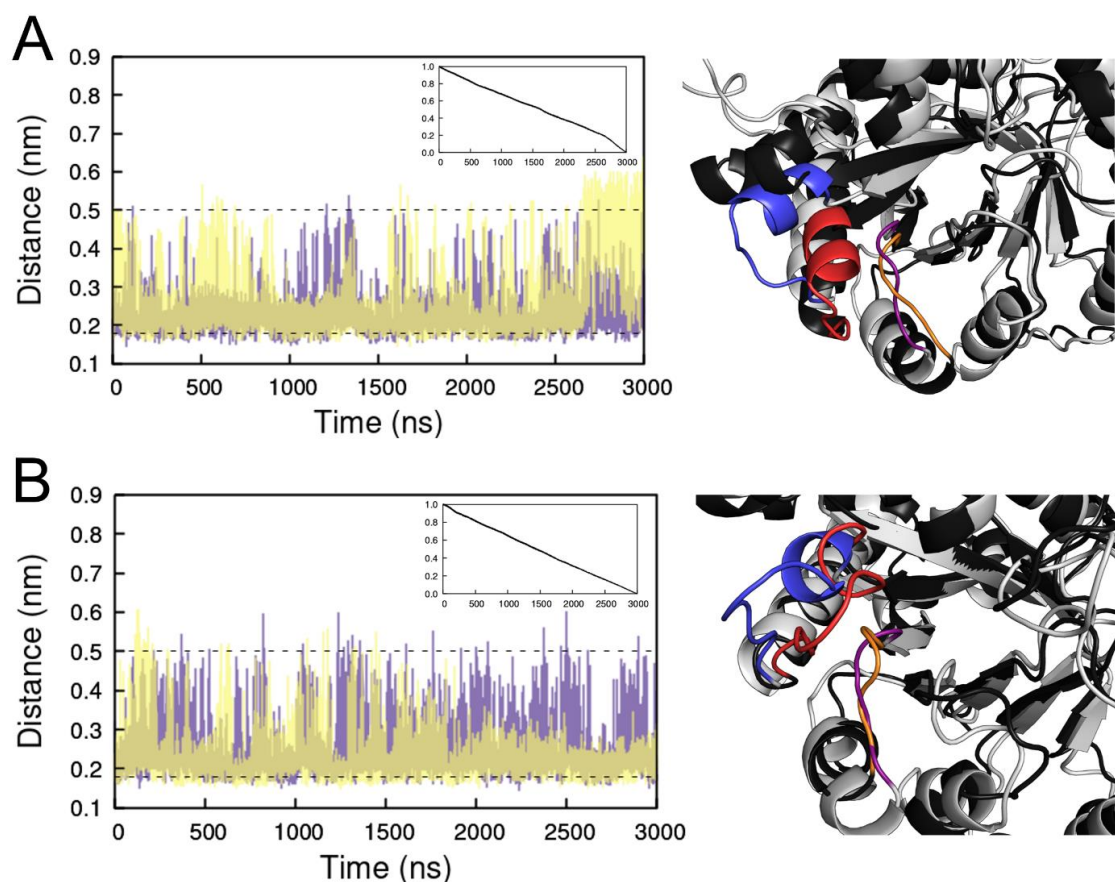


Figure S4.1 Minimum distance between loops 6 and 7 in TbTIM (A) and Mut1 (B) for monomer A (purple) and monomer B (yellow). Dashed lines mark the two different states sampled by the loops. Inside: cross-correlation between the loop state of the two monomers. Right: alignment of the open (gray) and closed (black) conformations. Loop 6 is shown in red

for the closed conformation and blue for the open conformation, loop 7 is shown in orange (closed) and purple (open).

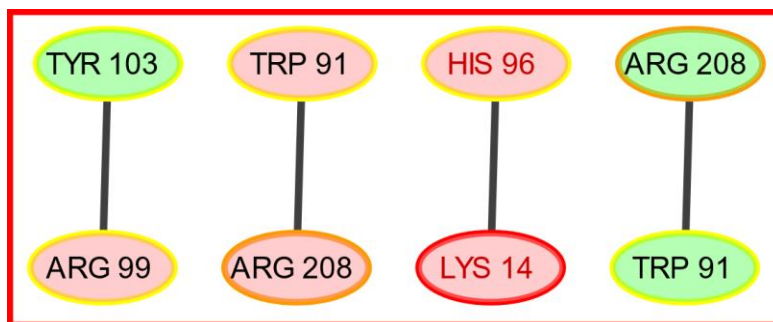


Figure S4.2 Cation- π interactions for TcTIM throughout the last 2 μ s of the simulation. Amino acids in monomer A are shown in green and residues in monomer B in red. Each node is coloured according to the color scheme for regions in Fig. 4.1 of the main text. The catalytic residues are written in red text.

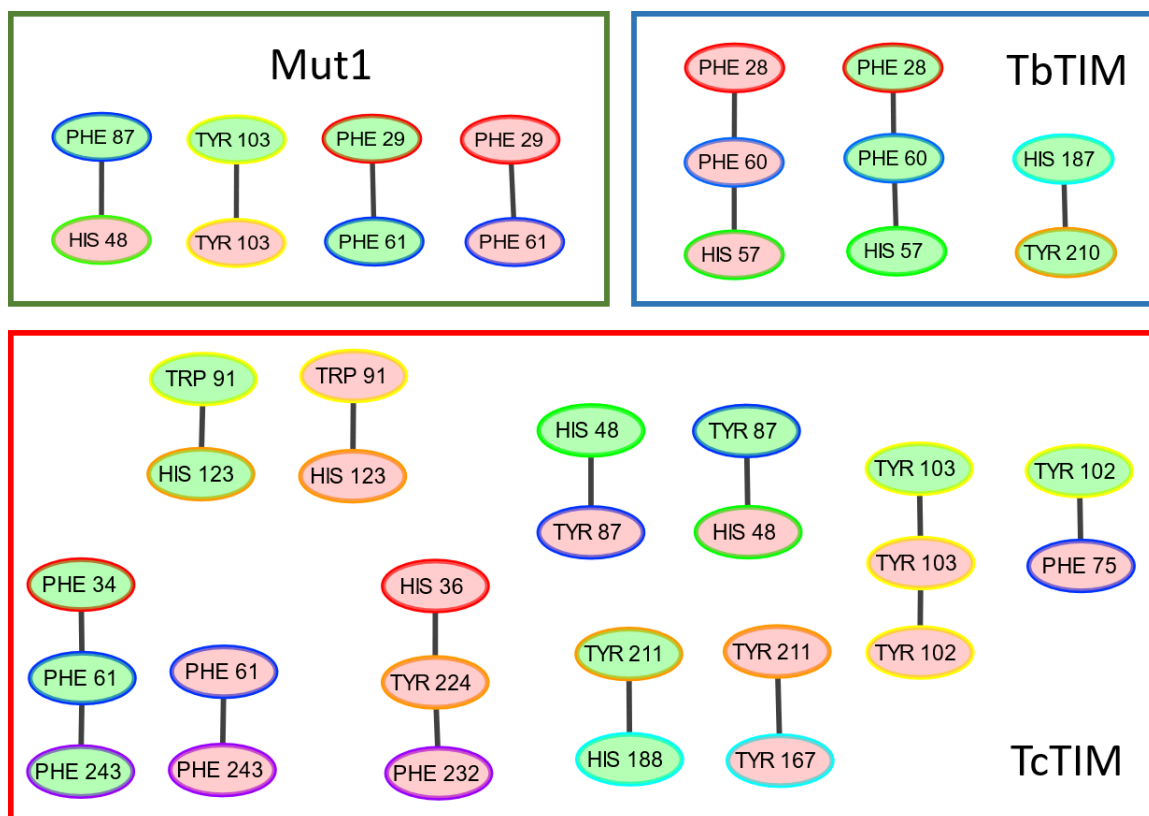


Figure S4.3 π - π interactions for TcTIM, TbTIM and Mut1 throughout the last 2 μ s of the simulations. Amino acids in monomer A are shown in green and residues in monomer B in

red. Each node is coloured according to the color scheme for regions in Fig. 4.1 of the main text.

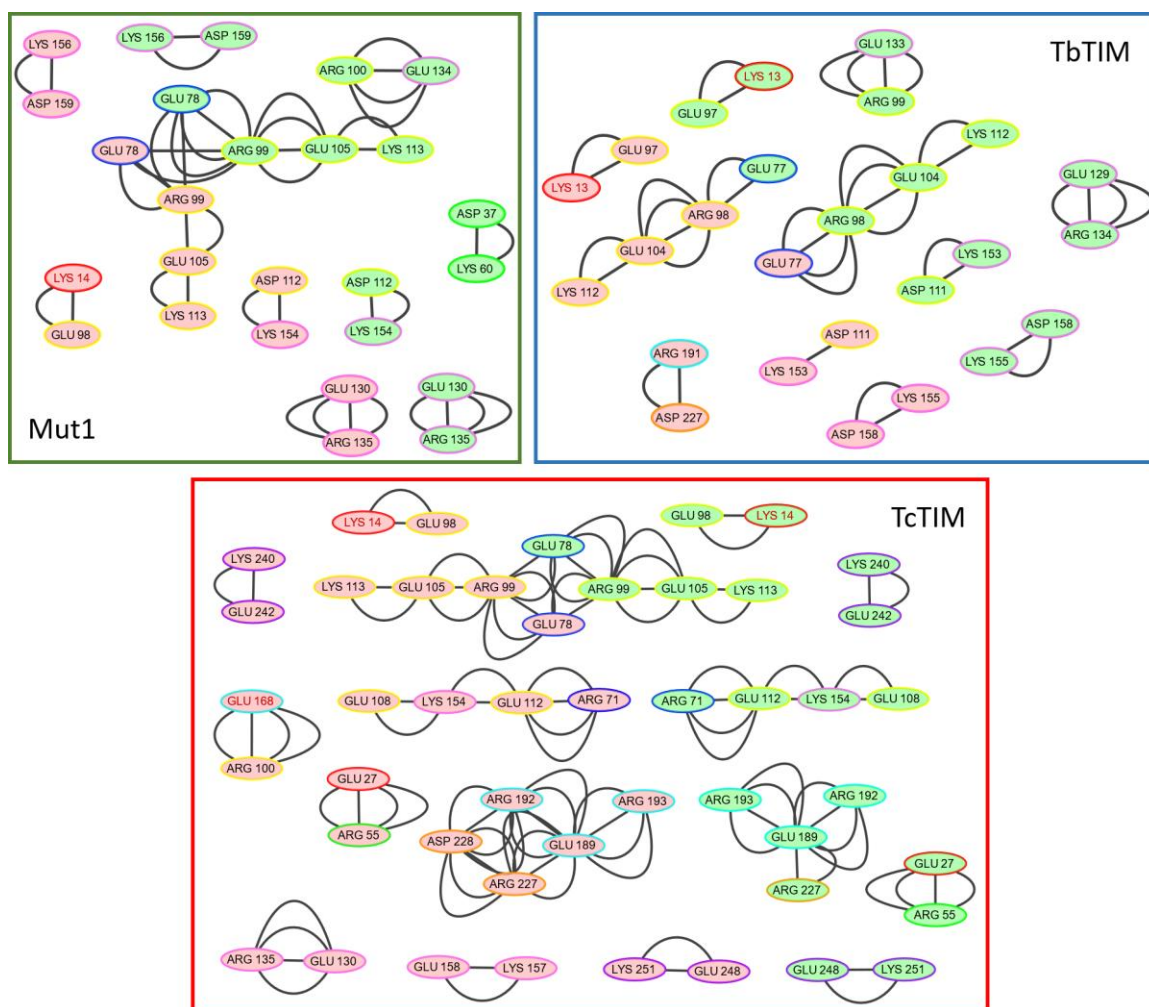


Figure S4.4 Salt bridges for TcTIM, TbTIM and Mut1 throughout the last 2 μ s of the simulations. TcTIM forms more salt bridges than the other two proteins. Amino acids in monomer A are shown in green and residues in monomer B in red. Each node is coloured according to the color scheme for regions in Fig. 4.1 of the main text. The catalytic residues are written in red text.

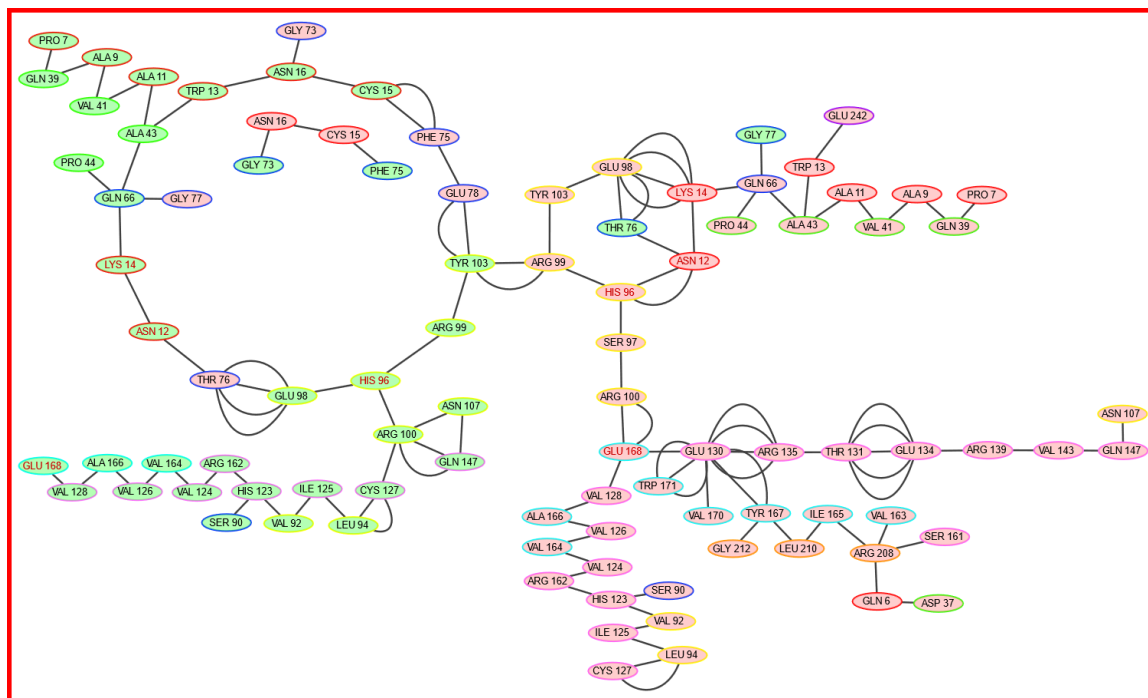


Figure S4.5 Main hydrogen bond network in TcTIM throughout the last 2 μ s of the simulation. Each node is coloured according to the color scheme for regions in Fig. 4.1 of the main text. The catalytic residues are written in red text.

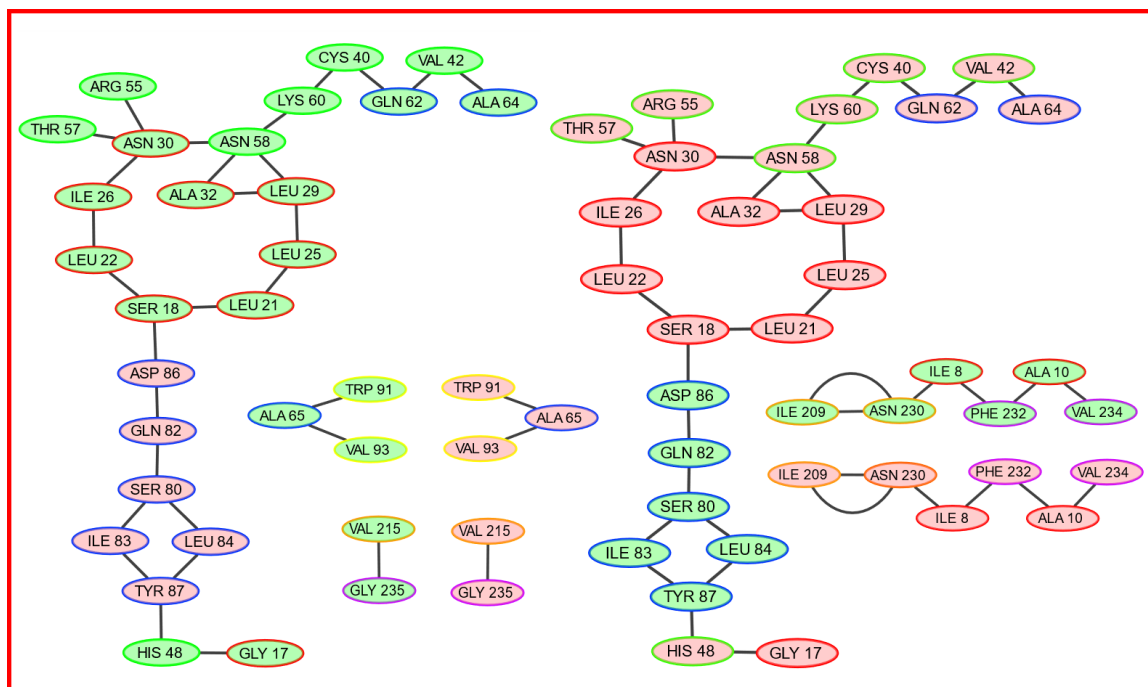


Figure S4.6 Hydrogen bonds in TcTIM throughout the last 2 μ s of the simulation. These hydrogen bonds are found in both monomers. Amino acids in monomer A are shown in green and residues in monomer B in red. Each node is coloured according to the color scheme for regions in Fig. 4.1 of the main text.

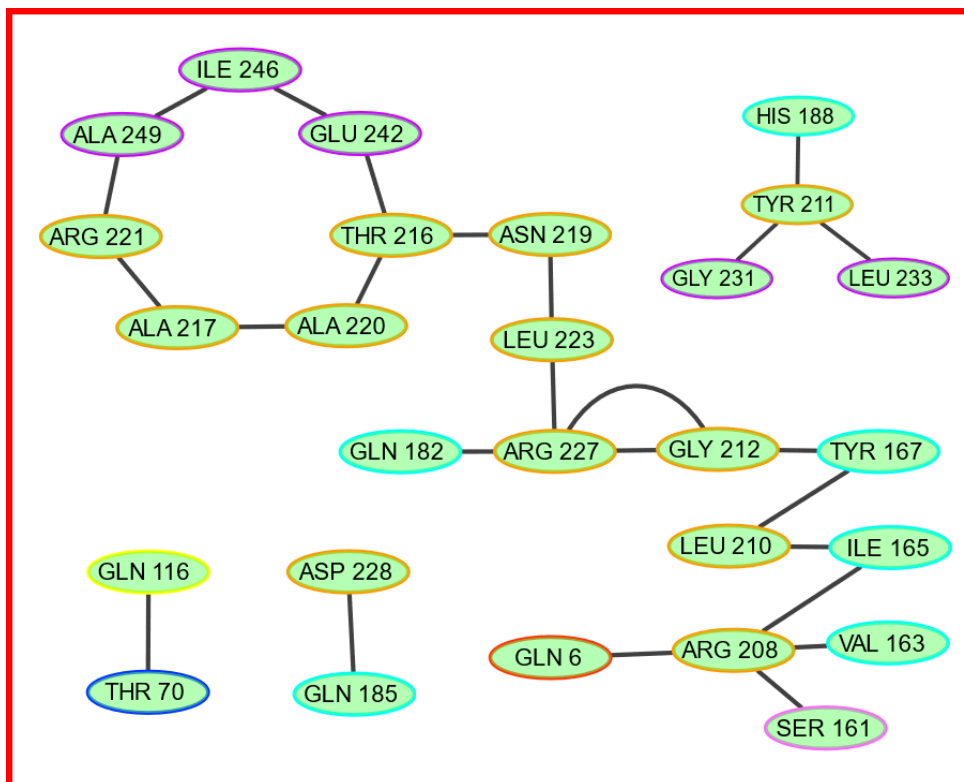


Figure S4.7 Hydrogen bonds in TcTIM throughout the last 2 μ s of the simulation. These hydrogen bonds are found in both monomers. Amino acids in monomer A are shown in green and residues in monomer B in red. Each node is coloured according to the color scheme for regions in Fig. 4.1 of the main text.

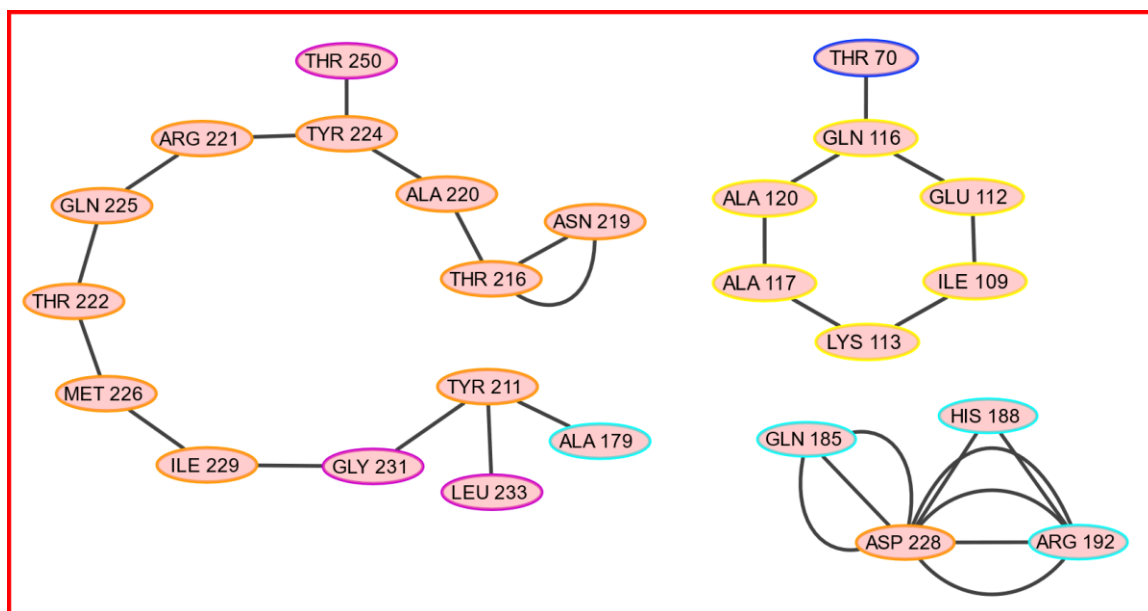


Figure S4.8 Hydrogen bonds in monomer B throughout the last 2 μ s of the TcTIM simulation. Each node is coloured according to the color scheme for regions in Fig. 4.1 of the main text.

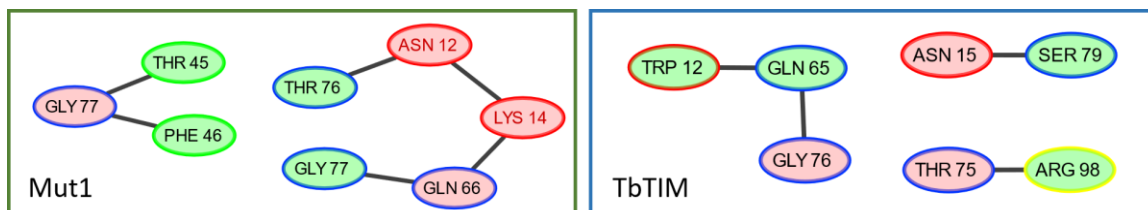


Figure S4.9 Hydrogen bonds involving amino acids at the interface between monomers in Mut1 and TbTIM throughout the last 2 μ s of the simulations. Amino acids in monomer A are shown in green and residues in monomer B in red. Each node is coloured according to the color scheme for regions in Fig. 4.1 of the main text. The catalytic residues are written in red text.

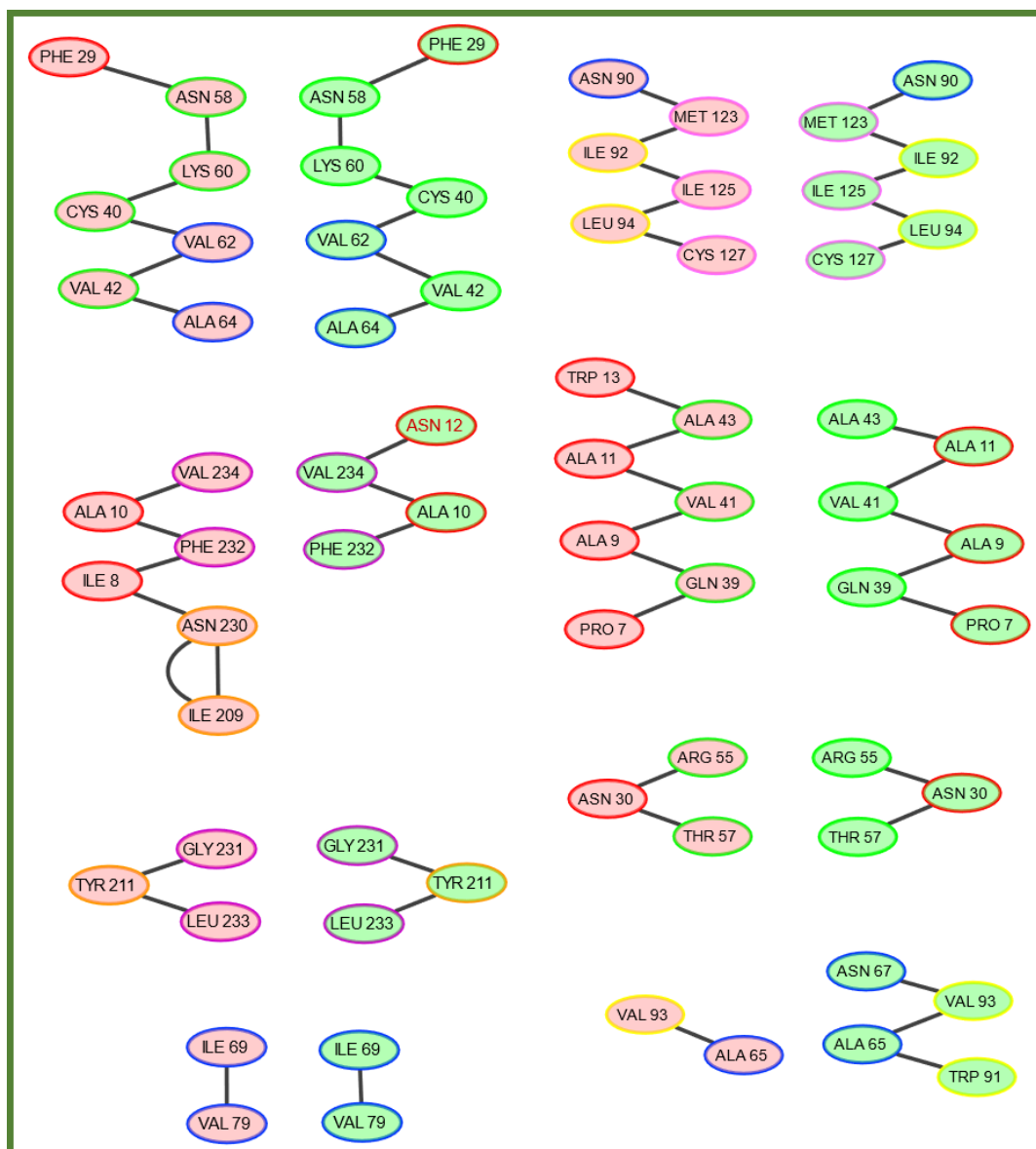


Figure S4.10 Hydrogen bonds in Mut1 throughout the last 2 μ s of the simulation. These hydrogen bonds are found in both monomers. Amino acids in monomer A are shown in green and residues in monomer B in red. Each node is coloured according to the color scheme for regions in Fig. 4.1 of the main text. The catalytic residue is written in red text.

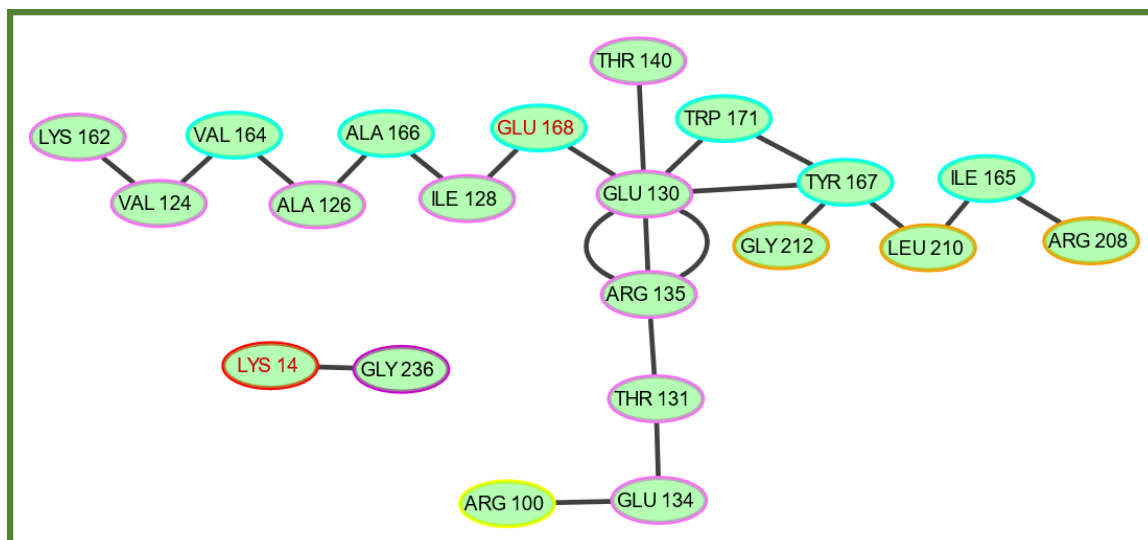


Figure S4.11 Hydrogen bonds in monomer A throughout the last 2 μ s of the Mut1 simulation. Each node is coloured according to the color scheme for regions in Fig. 4.1 of the main text. The catalytic residues are written in red text.

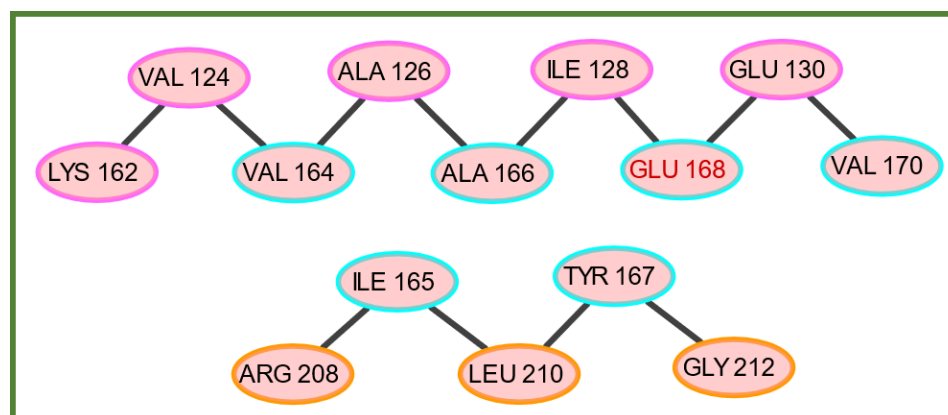


Figure S4.12 Hydrogen bonds in monomer B throughout the last 2 μ s of the Mut1 simulation. Each node is coloured according to the color scheme for regions in Fig. 4.1 of the main text. The catalytic residue is written in red text.

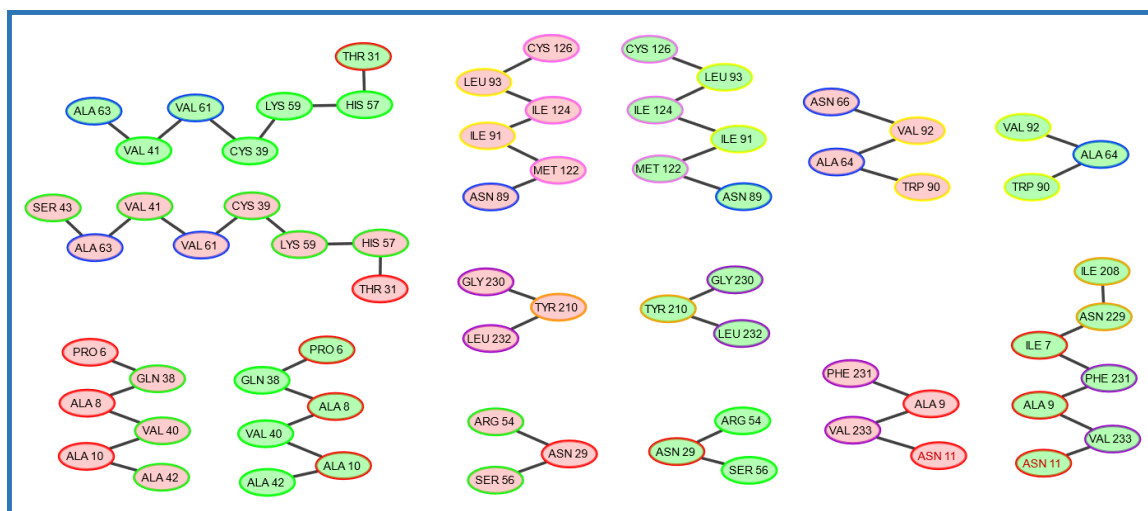


Figure S4.13 Hydrogen bonds in TbTIM throughout the last 2 μ s of the simulation. These hydrogen bonds are found in both monomers. Amino acids in monomer A are shown in green and residues in monomer B in red. Each node is coloured according to the color scheme for regions in Fig. 4.1 of the main text. The catalytic residues are written in red text.

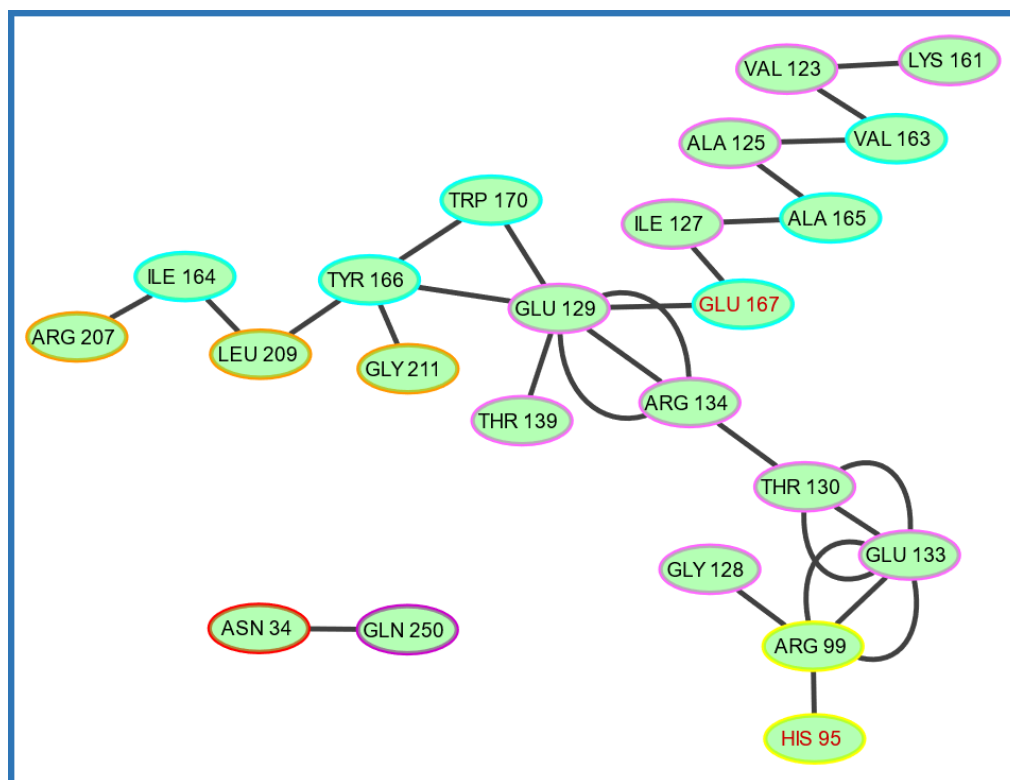


Figure S4.14 Hydrogen bonds in monomer A throughout the last 2 μ s of the TbTIM simulation. Each node is coloured according to the color scheme for regions in Fig. 4.1 of the main text. The catalytic residue is written in red text.

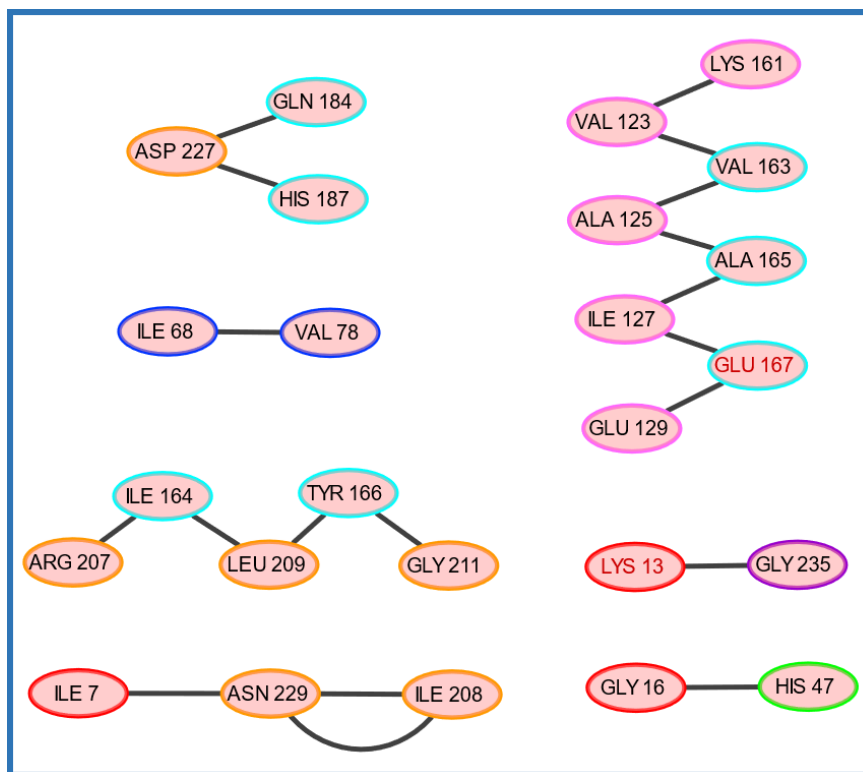


Figure S4.15 Hydrogen bonds in monomer B throughout the last 2 μ s of the TbTIM simulation. Each node is coloured according to the color scheme for regions in Fig. 4.1 of the main text. The catalytic residues are written in red text.

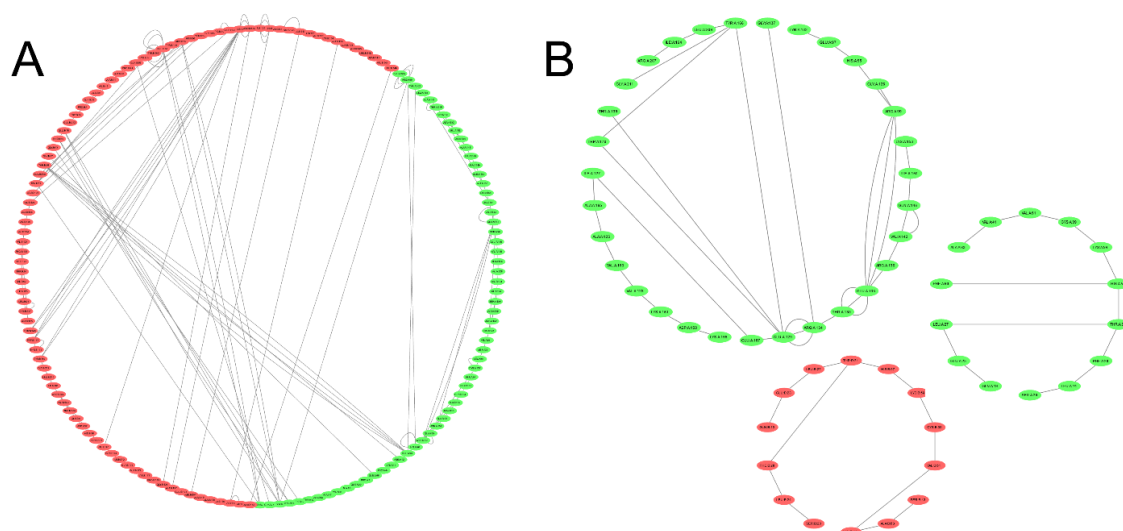


Figure S4.16 Main hydrogen bond networks throughout the last 2 μ s for A) TcTIM and B) TbTIM. Amino acids in monomer A are shown in green and residues in monomer B in red. Hydrogen bonds in TcTIM connect amino acids in a network that involves many interactions

between monomers and extends throughout the whole protein, in contrast with TbTIM, whose networks are contained within each monomer and involve fewer residues.

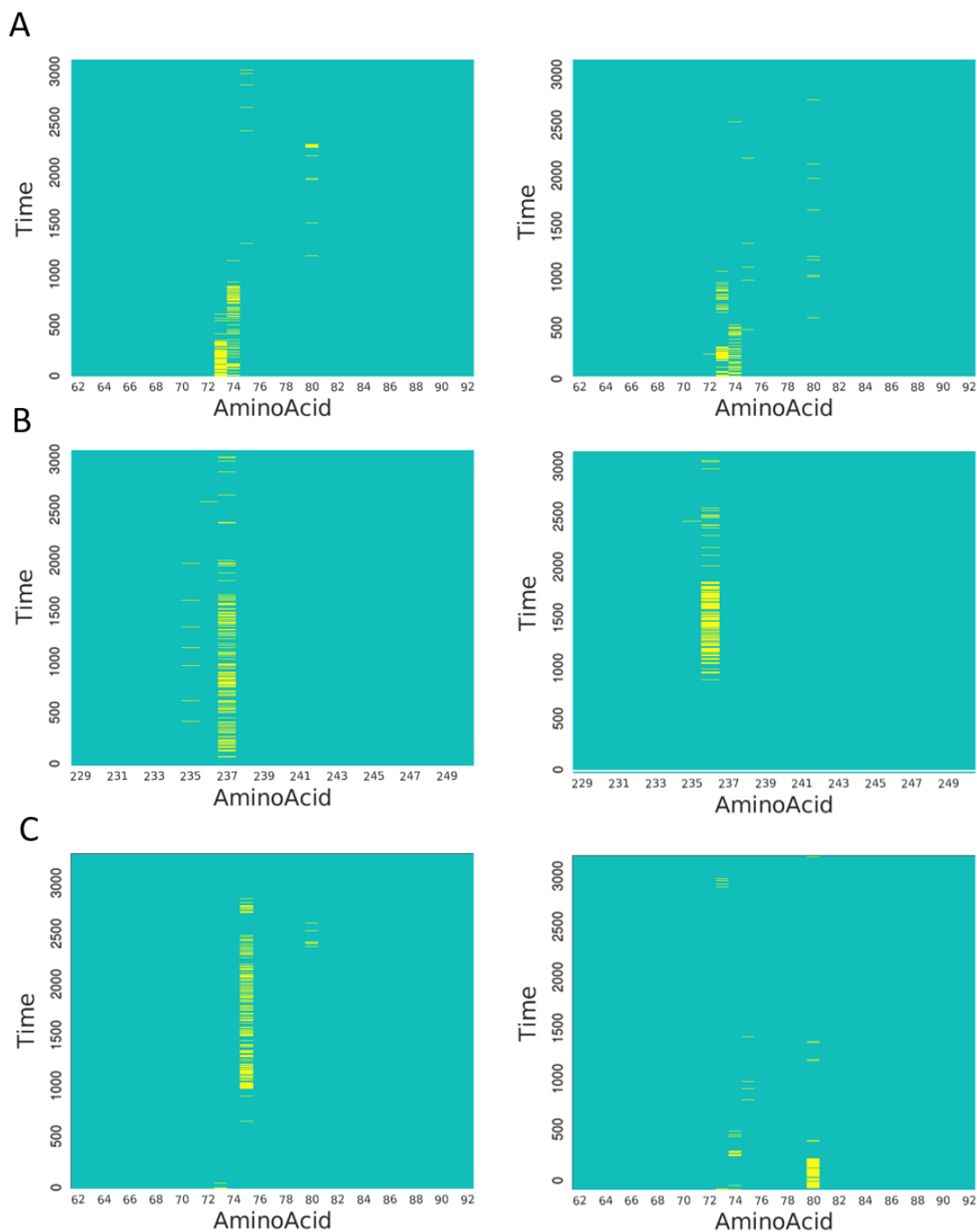


Figure S4.17 Hydrogen bonds in Cys 14/15 monomer A (left) and monomer B (right) in: A) TcTIM, B) TbTIM and C) Mut1. This cysteine forms hydrogen bonds with region 3 of the

other monomer in TcTIM and Mut1 and forms hydrogen bonds in region 8 of the same monomer in TbTIM.

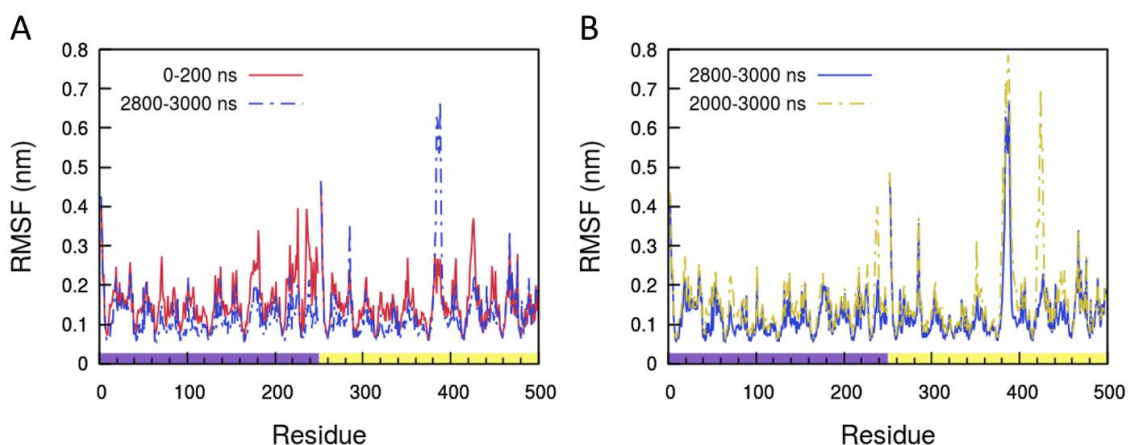


Figure S4.18 Changes over time in the root mean square fluctuation (RMSF) of the TbTIM simulation. RMSF of the first 200 ns of the simulation vs the last 200 ns (A), and RMSF of the last microsecond of the trajectory vs the last 200 ns (B). The color bar at the bottom of figure B distinguishes the residues in monomer A (purple) from those in monomer B (yellow). There are no significant peaks in the RMSF of the beginning of the trajectory. The last 200 ns of the simulation failed to capture the peak at loop 6 monomer B.

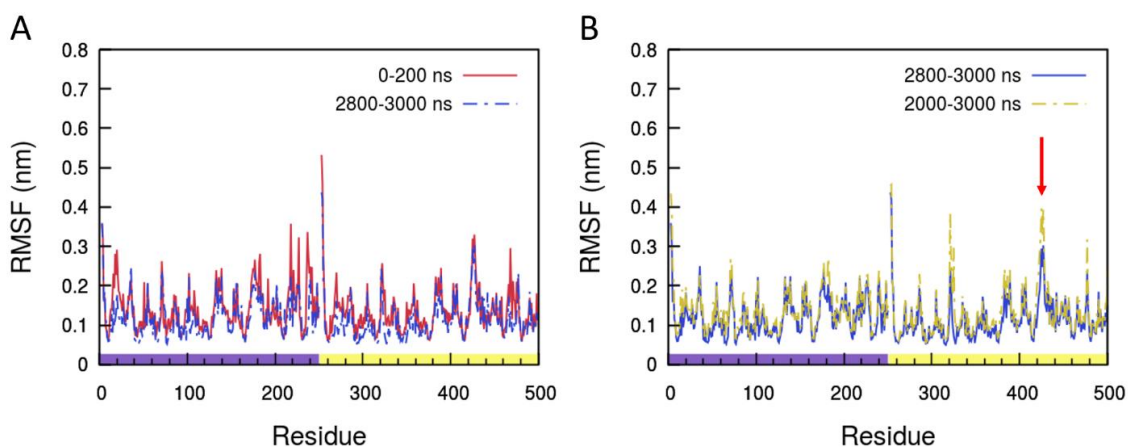


Figure S4.19 Changes over time in the root mean square fluctuation (RMSF) of the Mut1 simulation. RMSF of the first 200 ns of the simulation vs the last 200 ns (A), and RMSF of the last microsecond of the trajectory vs the last 200 ns (B). The color bar at the bottom of figure B distinguishes the residues in monomer A (purple) from those in monomer B (yellow).

Fluctuations at the minor peaks decreased with time and the peak at loop 6 monomer B (red arrow) increased.

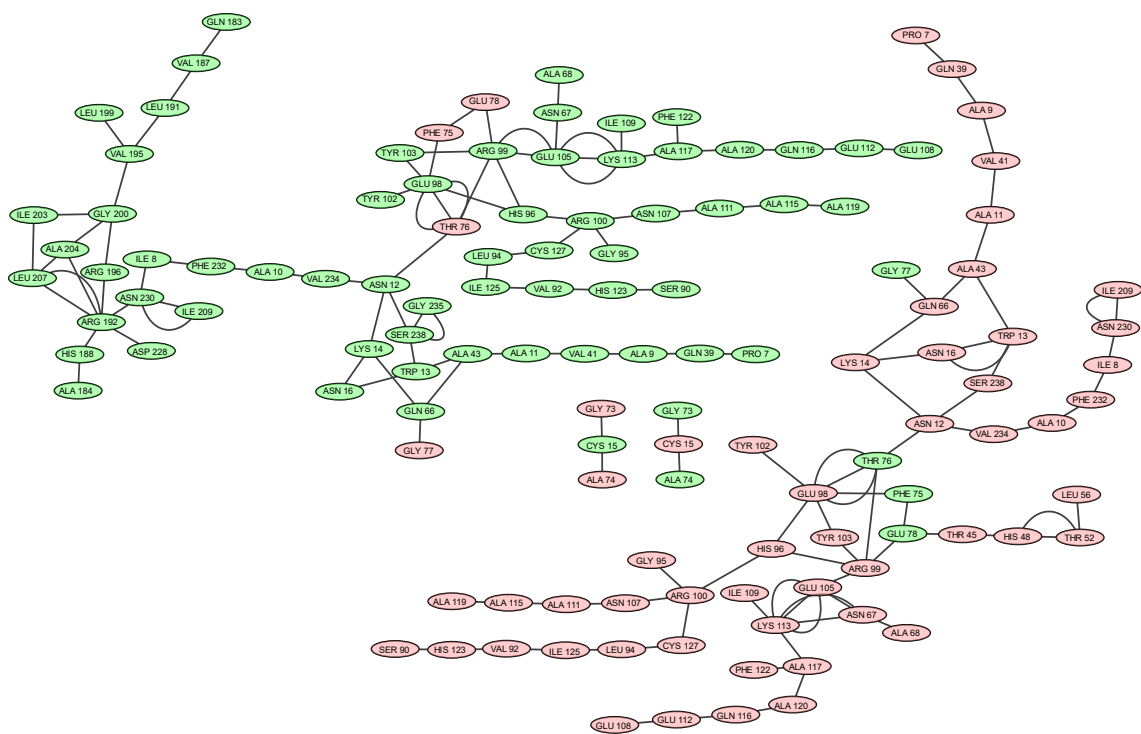


Figure S4.20 Main hydrogen bond networks for TcTIM in the most populated cluster of the first 500 ns of the simulation. There is a different network for each monomer. Amino acids in monomer A are shown in green and residues in monomer B in red.

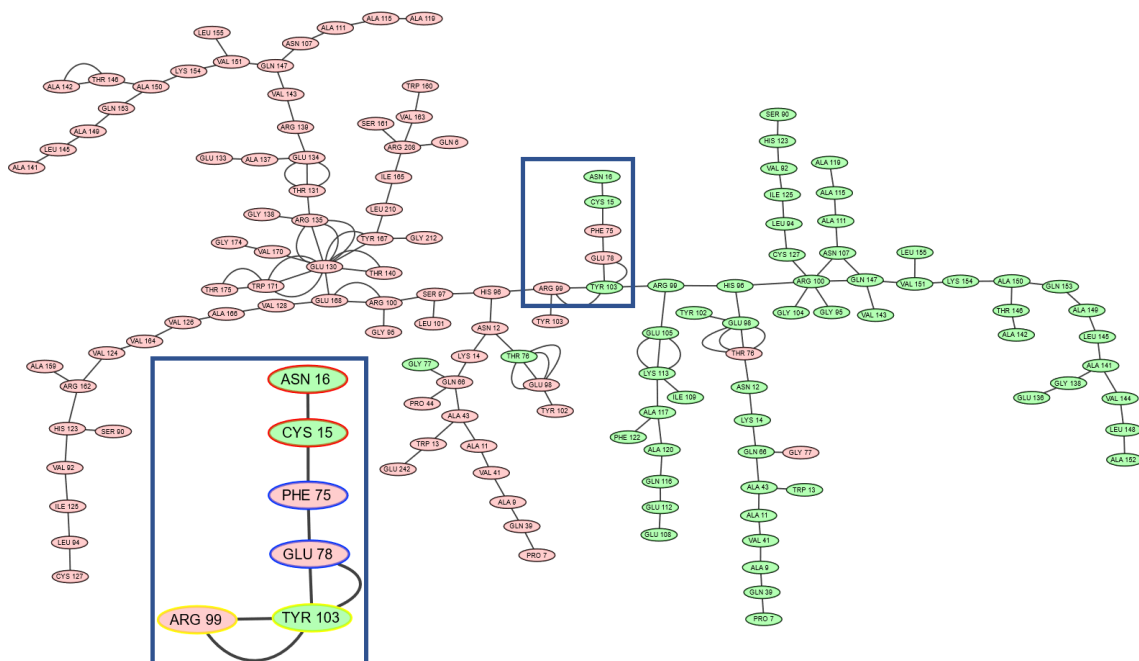


Figure S4.21 Hydrogen bond networks for TcTIM in the most populated cluster of the last 500 ns of the simulation. The residues in both monomers are connected through a single network of hydrogen bonds. Highlighted in blue are the residues at the interface of the hydrogen bond network (Fig. 4.11 in the main text). Amino acids in monomer A are shown in green and residues in monomer B in red.

4.8 Acknowledgments

The authors thank Dr. Mónica Rodríguez-Bolaños and Dr. Ruy Perez-Montfort for fruitful discussions. CCG thanks the Province of Ontario Trillium Scholarship Program. MK thanks the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Canada Research Chairs Program. Computing facilities were provided by SHARCNET (www.sharcnet.ca), Compute Canada (www.computecanada.ca).

4.9 References

- (1) Wierenga, R. K. The TIM-Barrel Fold: A Versatile Framework for Efficient Enzymes. *FEBS Lett.* **2001**, *492*, 193–198.

- (2) Wierenga, R. K.; Kapetaniou, E. G.; Venkatesan, R. Triosephosphate Isomerase: A Highly Evolved Biocatalyst. *Cell. Mol. Life Sci.* **2010**, *67* (23), 3961–3982. <https://doi.org/10.1007/s00018-010-0473-9>.
- (3) Roland, B. P.; Stuchul, K. A.; Larsen, S. B.; Amrich, C. G.; VanDemark, A. P.; Celotto, A. M.; Palladino, M. J. Evidence of a Triosephosphate Isomerase Non-Catalytic Function Crucial to Behavior and Longevity. *J. Cell Sci.* **2013**, *126* (14), 3151–3158. <https://doi.org/10.1242/jcs.124586>.
- (4) Brändén, C. I. The TIM Barrel-the Most Frequently Occurring Folding Motif in Proteins. *Curr. Opin. Struct. Biol.* **1991**, *1* (6), 978–983. [https://doi.org/10.1016/0959-440X\(91\)90094-A](https://doi.org/10.1016/0959-440X(91)90094-A).
- (5) Maes, D.; Zeelen, J. P.; Thanki, N.; Beaucamp, N.; Alvarez, M.; Minh Hoa Dao, T.; Backmann, J.; Martial, J. A.; Wyns, L.; Jaenicke, R.; Wierenga, R. K. The Crystal Structure of Triosephosphate Isomerase (TIM) from *Thermotoga Maritima*: A Comparative Thermostability Structural Analysis of Ten Different TIM Structures. *Proteins Struct. Funct. Genet.* **1999**, *37* (3), 441–453. [https://doi.org/10.1002/\(SICI\)1097-0134\(19991115\)37:3<441::AID-PROT11>3.0.CO;2-7](https://doi.org/10.1002/(SICI)1097-0134(19991115)37:3<441::AID-PROT11>3.0.CO;2-7).
- (6) Walden, H.; Bell, G. S.; Russell, R. J. M.; Siebers, B.; Hensel, R.; Taylor, G. L. Tiny TIM: A Small, Tetrameric, Hyperthermostable Triosephosphate Isomerase. *J. Mol. Biol.* **2001**, *306* (4), 745–757. <https://doi.org/10.1006/jmbi.2000.4433>.
- (7) Katebi, A. R.; Jernigan, R. L. The Critical Role of the Loops of Triosephosphate Isomerase for Its Oligomerization, Dynamics, and Functionality. *Protein Sci.* **2014**, *23* (2), 213–228. <https://doi.org/10.1002/pro.2407>.
- (8) Beaucamp, N.; Schurig, H.; Jaenicke, R. The PGK-TIM Fusion Protein from *Thermotoga Maritima* and Its Constituent Parts Are Intrinsically Stable and Fold Independently. *Biol. Chem.* **1997**, *378* (7), 679–685. <https://doi.org/10.1515/bchm.1997.378.7.679>.

- (9) Hernández-Alcántara, G.; Garza-Ramos, G.; Hernández, G. M.; Gómez-Puyou, A.; Pérez-Montfort, R. Catalysis and Stability of Triosephosphate Isomerase from *Trypanosoma Brucei* with Different Residues at Position 14 of the Dimer Interface. Characterization of a Catalytically Competent Monomeric Enzyme. *Biochemistry* **2002**, *41* (13), 4230–4238. <https://doi.org/10.1021/bi011950f>.
- (10) Kursula, I.; Partanen, S.; Lambeir, A. M.; Antonov, D. M.; Augustyns, K.; Wierenga, R. K. Structural Determinants for Ligand Binding and Catalysis of Triosephosphate Isomerase. *Eur. J. Biochem.* **2001**, *268* (19), 5189–5196. <https://doi.org/10.1046/j.0014-2956.2001.02452.x>.
- (11) Gómez-Puyou, A.; Saavedra-Lira, E.; Becker, I.; Zubillaga, R. A.; Rojo-Domínguez, A.; Pérez-Montfort, R. Using Evolutionary Changes to Achieve Species-Specific Inhibition of Enzyme Action - Studies with Triosephosphate Isomerase. *Chem. Biol.* **1995**, *2* (12), 847–855. [https://doi.org/10.1016/1074-5521\(95\)90091-8](https://doi.org/10.1016/1074-5521(95)90091-8).
- (12) Velanker, S. S.; Ray, S. S.; Gokhale, R. S.; Suma, S.; Balaram, H.; Balaram, P.; Murthy, M. R. N. Triosephosphate Isomerase from *Plasmodium Falciparum*: The Crystal Structure Provides Insights into Antimalarial Drug Design. *Structure* **1997**, *5* (6), 751–761. [https://doi.org/10.1016/S0969-2126\(97\)00230-X](https://doi.org/10.1016/S0969-2126(97)00230-X).
- (13) Téllez-Valencia, A.; Olivares-Illana, V.; Hernández-Santoyo, A.; Pérez-Montfort, R.; Costas, M.; Rodríguez-Romero, A.; López-Calahorra, F.; Tuena De Gómez-Puyou, M.; Gómez-Puyou, A. Inactivation of Triosephosphate Isomerase from *Trypanosoma Cruzi* by an Agent That Perturbs Its Dimer Interface. *J. Mol. Biol.* **2004**, *341* (5), 1355–1365. <https://doi.org/10.1016/j.jmb.2004.06.056>.
- (14) Somme, J.; Van Laer, B.; Roovers, M.; Steyaert, J.; Versées, W.; Droogmans, L. Characterization of Two Homologous 2'-O-Methyltransferases Showing Different Specificities for Their TRNA Substrates. *Rna* **2014**, *20* (8), 1257–1271. <https://doi.org/10.1261/rna.044503.114>.
- (15) Theßeling, A.; Rasmussen, T.; Burschel, S.; Wohlwend, D.; Kägi, J.; Müller, R.;

- Böttcher, B.; Friedrich, T. Homologous Bd Oxidases Share the Same Architecture but Differ in Mechanism. *Nat. Commun.* **2019**, *10* (1), 1–7. <https://doi.org/10.1038/s41467-019-13122-4>.
- (16) Garza-Ramos, G.; Cabrera, N.; Saavedra-Lira, E.; Tuena De Gómez-Puyou, M.; Ostoa-Saloma, P.; Pérez-Montfort, R.; Gómez-Puyou, A. Sulfhydryl Reagent Susceptibility in Proteins with High Sequence Similarity: Triosephosphate Isomerase from *Trypanosoma Brucei*, *Trypanosoma Cruzi* and *Leishmania Mexicana*. *Eur. J. Biochem.* **1998**, *253* (3), 684–691. <https://doi.org/10.1046/j.1432-1327.1998.2530684.x>.
- (17) Lindqvist, Y.; Branden, C. I.; Mathews, F. S.; Lederer, F. Spinach Glycolate Oxidase and Yeast Flavocytochrome B2 Are Structurally Homologous and Evolutionarily Related Enzymes with Distinctly Different Function and Flavin Mononucleotide Binding. *J. Biol. Chem.* **1991**, *266* (5), 3198–3207. [https://doi.org/10.1016/s0021-9258\(18\)49974-7](https://doi.org/10.1016/s0021-9258(18)49974-7).
- (18) García-Torres, I.; Cabrera, N.; Torres-Larios, A.; Rodríguez-Bolaños, M.; Díaz-Mazariegos, S.; Gómez-Puyou, A.; Perez-Montfort, R. Identification of Amino Acids That Account for Long-Range Interactions in Two Triosephosphate Isomerases from Pathogenic Trypanosomes. *PLoS One* **2011**, *6* (4). <https://doi.org/10.1371/journal.pone.0018791>.
- (19) Zomosa-Signoret, V.; Hernández-Alcántara, G.; Reyes-Vivas, H.; Martínez-Martínez, E.; Garza-Ramos, G.; Pérez-Montfort, R.; De Gómez-Puyou, M. T.; Gómez-Puyou, A. Control of the Reactivation Kinetics of Homodimeric Triosephosphate Isomerase from Unfolded Monomers. *Biochemistry* **2003**, *42* (11), 3311–3318. <https://doi.org/10.1021/bi0206560>.
- (20) Reyes-Vivas, H.; Martínez-Martínez, E.; Mendoza-Hernández, G.; López-Velázquez, G.; Pérez-Montfort, R.; Tuena De Gómez-Puyou, M.; Gómez-Puyou, A. Susceptibility to Proteolysis of Triosephosphate Isomerase from Two Pathogenic Parasites: Characterization of an Enzyme with an Intact and a Nicked Monomer.

- Proteins Struct. Funct. Genet.* **2002**, *48* (3), 580–590. <https://doi.org/10.1002/prot.10179>.
- (21) Rodríguez-Bolaños, M.; Cabrera, N.; Perez-Montfort, R. Identification of the Critical Residues Responsible for Differential Reactivation of the Triosephosphate Isomerases of Two Trypanosomes. *Open Biol.* **2016**, *6* (10). <https://doi.org/10.1098/rsob.160161>.
- (22) Vázquez-Raygoza, A.; Cano-González, L.; Velázquez-Martínez, I.; Trejo-Soto, P. J.; Castillo, R.; Hernández-Campos, A.; Hernández-Luis, F.; Oria-Hernández, J.; Castillo-Villanueva, A.; Avitia-Domínguez, C.; Sierra-Campos, E.; Valdez-Solana, M.; Téllez-Valencia, A. Species-Specific Inactivation of Triosephosphate Isomerase from *Trypanosoma Brucei*: Kinetic and Molecular Dynamics Studies. *Molecules* **2017**, *22* (12). <https://doi.org/10.3390/molecules22122055>.
- (23) Téllez-Valencia, A.; Ávila-Ríos, S.; Pérez-Montfort, R.; Rodríguez-Romero, A.; Tuena de Gómez-Puyou, M.; López-Calahorra, F.; Gómez-Puyou, A. Highly Specific Inactivation of Triosephosphate Isomerase from *Trypanosoma Cruzi*. *Biochem. Biophys. Res. Commun.* **2002**, *295* (4), 958–963. [https://doi.org/10.1016/S0006-291X\(02\)00796-9](https://doi.org/10.1016/S0006-291X(02)00796-9).
- (24) Cansu, S.; Doruker, P. Dimerization Affects Collective Dynamics of Triosephosphate Isomerase. *Biochemistry* **2008**, *47* (5), 1358–1368. <https://doi.org/10.1021/bi701916b>.
- (25) Díaz-Vergara, N.; Piñeiro, Á. Molecular Dynamics Study of Triosephosphate Isomerase from *Trypanosoma Cruzi* in Water/Decane Mixture. *J. Phys. Chem. B* **2008**, *112* (11), 3529–3539. <https://doi.org/10.1021/jp7102275>.
- (26) Quezada, A. G.; Díaz-Salazar, A. J.; Cabrera, N.; Pérez-Montfort, R.; Piñeiro, Á.; Costas, M. Interplay between Protein Thermal Flexibility and Kinetic Stability. *Structure* **2017**, *25* (1), 167–179. <https://doi.org/10.1016/j.str.2016.11.018>.
- (27) Dantu, S. C.; Groenhof, G. Conformational Dynamics of Active Site Loops 5, 6 and

- 7 of Enzyme Triosephosphate Isomerase: A Molecular Dynamics Study. *bioRxiv* **2018**. <https://doi.org/10.1101/459198>.
- (28) Liao, Q.; Kulkarni, Y.; Sengupta, U.; Petrović, D.; Mulholland, A. J.; Van Der Kamp, M. W.; Strodel, B.; Kamerlin, S. C. L. Loop Motion in Triosephosphate Isomerase Is Not a Simple Open and Shut Case. *J. Am. Chem. Soc.* **2018**, *140* (46), 15889–15903. <https://doi.org/10.1021/jacs.8b09378>.
- (29) Wierenga, R. K.; Borchert, T. V.; Noble, M. E. M. Crystallographic Binding Studies with Triosephosphate Isomerases: Conformational Changes Induced by Substrate and Substrate-Analogues. *FEBS Lett.* **1992**, *307* (1), 34–39. [https://doi.org/10.1016/0014-5793\(92\)80897-P](https://doi.org/10.1016/0014-5793(92)80897-P).
- (30) Borchert, T. V.; Kishan, K. R.; Zeelen, J. P.; Schliebs, W.; Thanki, N.; Abagyan, R.; Jaenicke, R.; Wierenga, R. K. Three New Crystal Structures of Point Mutation Variants of Mono TIM: Conformational Flexibility of Loop-1, Loop-4 and Loop-8. *Structure* **1995**, *3* (7), 669–679. [https://doi.org/10.1016/S0969-2126\(01\)00202-7](https://doi.org/10.1016/S0969-2126(01)00202-7).
- (31) Maldonado, E.; Soriano-García, M.; Moreno, A.; Cabrera, N.; Garza-Ramos, G.; Tuena de Gómez-Puyou, M.; Gómez-Puyou, A.; Perez-Montfort, R. Differences in the Intersubunit Contacts in Triosephosphate Isomerase from Two Closely Related Pathogenic Trypanosomes. *J. Mol. Biol.* **1998**, *283* (1), 193–203.
- (32) Wierengat, R. K.; Noble, M. E. M.; Vriend, G.; Nauche, S.; Hol, W. G. J. Refined 1.83 Å Structure of Trypanosomal Triosephosphate Isomerase Crystallized in the Presence of 2.4 M-Ammonium Sulphate: A Comparison with the Structure of the Trypanosomal Triosephosphate Isomerase-Glycerol-3-Phosphate Complex. *J. Mol. Biol.* **1991**, *220* (4), 995–1015.
- (33) Hekkelman, M. L.; te Beek, T. A. H.; Pettifer, S. R.; Thorne, D.; Attwood, T. K.; Vriend, G. WIWS: A Protein Structure Bioinformatics Web Service Collection. *Nucleic Acids Res.* **2010**, *38* (SUPPL. 2), 719–723. <https://doi.org/10.1093/nar/gkq453>.

- (34) Schrödinger, L. The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC. 2015.
- (35) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindah, E. GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, 1–2, 19–25. <https://doi.org/10.1016/j.softx.2015.06.001>.
- (36) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D. Comparison of Simple Potential Functions for Simulating Liquid Water Liquid Water. *J. Chem. Phys.* **1983**, 79, 926–935.
- (37) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; De Groot, B. L.; Grubmüller, H.; MacKerell, A. D. CHARMM36m: An Improved Force Field for Folded and Intrinsically Disordered Proteins. *Nat. Methods* **2016**, 14 (1), 71–73. <https://doi.org/10.1038/nmeth.4067>.
- (38) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An N·log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, 98 (12), 10089–10092. <https://doi.org/10.1063/1.464397>.
- (39) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* **1995**, 103 (19), 8577–8593. <https://doi.org/10.1063/1.470117>.
- (40) Bussi, G.; Donadio, D.; Parrinello, M. Canonical Sampling through Velocity Rescaling. *J. Chem. Phys.* **2007**, 126 (1), 1–7. <https://doi.org/10.1063/1.2408420>.
- (41) Parrinello, M.; Rahman, A. Polymorphic Transitions in Single Crystals : A New Molecular Dynamics Method. *J. Appl. Phys.* **1981**, 52, 7182–7190.
- (42) Hess, B. P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. *J. Chem. Theory Comput.* **2008**, 4 (1), 116–122. <https://doi.org/10.1021/ct700200b>.
- (43) Michaud-agrawal, N.; Denning, E. J.; Woolf, T. B.; Beckstein, O. Software News

- and Updates MDAnalysis: A Toolkit for the Analysis of Molecular Dynamics Simulations. *J. Comput. Chem.* **2011**, *32* (10), 2319–2327. <https://doi.org/10.1002/jcc>.
- (44) Gowers, R.; Linke, M.; Barnoud, J.; Reddy, T.; Melo, M.; Seyler, S.; Domański, J.; Dotson, D.; Buchoux, S.; Kenney, I.; Beckstein, O. MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. *Proc. 15th Python Sci. Conf.* **2016**, No. Scipy, 98–105. <https://doi.org/10.25080/majora-629e541a-00e>.
- (45) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graph.* **1996**, *14* (1), 33–38. [https://doi.org/10.1016/0263-7855\(96\)00018-5](https://doi.org/10.1016/0263-7855(96)00018-5).
- (46) Contreras-Riquelme, S.; Garate, J. A.; Perez-Acle, T.; Martin, A. J. M. RIP-MD: A Tool to Study Residue Interaction Networks in Protein Molecular Dynamics. *PeerJ* **2018**, *2018* (12), 1–18. <https://doi.org/10.7717/peerj.5998>.
- (47) Knowles, J. R. Enzyme Catalysis: Not Different, Just Better. *Nature* **1991**, *350* (6314), 121–124. <https://doi.org/10.1038/350121a0>.
- (48) Rozovsky, S.; McDermott, A. E. The Time Scale of the Catalytic Loop Motion in Triosephosphate Isomerase. *J. Mol. Biol.* **2001**, *310* (1), 259–270. <https://doi.org/10.1006/jmbi.2001.4672>.
- (49) Rozovsky, S.; Jogl, G.; Tong, L.; McDermott, A. E. Solution-State NMR Investigations of Triosephosphate Isomerase Active Site Loop Motion: Ligand Release in Relation to Active Site Loop Dynamics. *J. Mol. Biol.* **2001**, *310* (1), 271–280. <https://doi.org/10.1006/jmbi.2001.4673>.
- (50) Daura, X.; Mark, A. E.; Van Gunsteren, W. F. Parametrization of Aliphatic CH_n United Atoms of GROMOS96 Force Field. *J. Comput. Chem.* **1998**, *19* (5), 535–547. [https://doi.org/10.1002/\(SICI\)1096-987X\(19980415\)19:5<535::AID-JCC6>3.0.CO;2-N](https://doi.org/10.1002/(SICI)1096-987X(19980415)19:5<535::AID-JCC6>3.0.CO;2-N).
- (51) Berendsen, H. J. C.; Postma, J. P. M.; Van Gunsteren, W. F.; Hermans, J. Interaction Models for Water in Relation to Protein Hydration. *Jerusalem Symp. Quantum*

Chem. Biochem. **1981**, *14*. <https://doi.org/10.1007/978-94-015-7658-1>.

- (52) Seemayer, S.; Gruber, M.; Söding, J. CCMpred - Fast and Precise Prediction of Protein Residue-Residue Contacts from Correlated Mutations. *Bioinformatics* **2014**, *30* (21), 3128–3130. <https://doi.org/10.1093/bioinformatics/btu500>.
- (53) Tegge, A. N.; Wang, Z.; Eickholt, J.; Cheng, J. NNcon: Improved Protein Contact Map Prediction Using 2D-Recursive Neural Networks. *Nucleic Acids Res.* **2009**, *37* (SUPPL. 2), 515–518. <https://doi.org/10.1093/nar/gkp305>.
- (54) Monastyrskyy, B.; Fidelis, K.; Tramontano, A.; Kryshtafovych, A. Evaluation of Residue-Residue Contact Predictions in CASP9. *Proteins Struct. Funct. Bioinforma.* **2011**, *79* (SUPPL. 10), 119–125. <https://doi.org/10.1002/prot.23160>.
- (55) Li, Y.; Fang, Y.; Fang, J. Predicting Residue-Residue Contacts Using Random Forest Models. *Bioinformatics* **2011**, *27* (24), 3379–3384. <https://doi.org/10.1093/bioinformatics/btr579>.
- (56) Gallivan, J. P.; Dougherty, D. A. Cation- π Interactions in Structural Biology. *Proc. Natl. Acad. Sci. U. S. A.* **1999**, *96* (17), 9459–9464. <https://doi.org/10.1073/pnas.96.17.9459>.
- (57) Burley, A. S. K.; Petsko, G. a. Aromatic-Aromatic Interaction: A Mechanism of Protein Structure Stabilization Author(s): S. K. Burley and G. A. Petsko Source: *Science* (80-.). **1985**, *229* (4708), 23–28.
- (58) Donald, J. E.; Kulp, D. W.; Degrado, W. F. Salt Bridges: Geometrically Specific, Designable Interactions. *Proteins* **2011**, *79* (3), 898–915. <https://doi.org/10.1002/prot.22927.Salt>.
- (59) Harris, T. K.; Mildvan, A. S. High-Precision Measurement of Hydrogen Bond Lengths in Proteins by Nuclear Magnetic Resonance Methods. *Proteins Struct. Funct. Genet.* **1999**, *35* (3), 275–282. [https://doi.org/10.1002/\(SICI\)1097-0134\(19990515\)35:3<275::AID-PROT1>3.0.CO;2-V](https://doi.org/10.1002/(SICI)1097-0134(19990515)35:3<275::AID-PROT1>3.0.CO;2-V).

- (60) Ostoa-Saloma, P.; Garza-Ramos, G.; Ramírez, J.; Becker, I.; Berzunza, M.; Landa, A.; Gomez-Puyou, A.; Tuena De Gómez-Puyou, M.; Pérez-Montfort, R. Cloning, Expression, Purification and Characterization of Triosephosphate Isomerase from *Trypanosoma Cruzi*. *Eur. J. Biochem.* **1997**, *244* (3), 700–705. <https://doi.org/10.1111/j.1432-1033.1997.00700.x>.
- (61) Dougherty, D. A. The Cation- π Interaction. *Acc. Chem. Res.* **2013**, *46* (4), 885–893. <https://doi.org/10.1021/ar300265y>.The.
- (62) Kursula, I.; Partanen, S.; Lambeir, A. M.; Wierenga, R. K. The Importance of the Conserved Arg191-Asp227 Salt Bridge of Triosephosphate Isomerase for Folding, Stability, and Catalysis. *FEBS Lett.* **2002**, *518* (1–3), 39–42. [https://doi.org/10.1016/S0014-5793\(02\)02639-X](https://doi.org/10.1016/S0014-5793(02)02639-X).
- (63) Karshikoff, A.; Jelesarov, I. Salt Bridges and Conformational Flexibility: Effect on Protein Stability. *Biotechnol. Biotechnol. Equip.* **2008**, *22* (1), 606–611. <https://doi.org/10.1080/13102818.2008.10817520>.

5 Structure and dynamics of the Rett syndrome protein, MeCP2

Noriyuki Kodera^{†,1}, Anna A. Kalashnikova^{†,2}, Mary E. Porter-Goff², Catherine A. Musselman³, Cecilia Chávez-García^{4,5}, Mikko Karttunen^{4,5,6}, Borries Demeler⁷, Tatiana G. Kutateladze³, Toshio Ando^{*,1} and Jeffrey C. Hansen^{*}

¹Nano Life Science Institute (WPI-NanoLSI), Kanazawa University, Kakuma-machi, Kanazawa 920-1192, Japan; ²Department of Biochemistry and Molecular Biology, Colorado State University, Fort Collins, CO 80523, USA; ³Department of Pharmacology, University of Colorado School of Medicine, Aurora, CO 80045, USA; ⁴Department of Chemistry, The University of Western Ontario, London, Ontario, Canada N6A 3K7; ⁵The Center for Advanced Materials and Biomaterials Research, The University of Western Ontario, London, Ontario, Canada N6K 3K7; ⁶Department of Physics and Astronomy, The University of Western Ontario, London, Ontario, Canada N6A 5B7; ⁷Department of Chemistry and Biochemistry, The University of Lethbridge, Lethbridge, Alberta, Canada, T1K 3M4

† Equal contribution

*Corresponding authors: tando@staff.kanazawa-u.ac.jp, Jeffrey.C.Hansen@colostate.edu

Submission ID: NAR-01761-2020

5.1 Abstract

Methyl-CpG binding protein 2 (MeCP2) is a chromatin regulatory protein essential for brain development and activity in vertebrates. Specific missense and nonsense mutations in MeCP2 lead to the neurodevelopmental disorder, Rett syndrome (RTT). To understand the structure and dynamics of MeCP2 and gain insight into the molecular basis of RTT, we characterized MeCP2 properties using high speed atomic force microscopy and solution-state approaches. MeCP2 is an intrinsically disordered protein that displays highly dynamic behavior. MeCP2 transitions between a fully extended dumbbell-like structure with the methyl DNA binding domain (MBD) and C-terminal domain (CTD) at the extremities, and a compact structure where the MBD and CTD interact in *cis*. The MBD within the full length protein equilibrates between unfolded and well folded states. MBD–CTD interactions stabilize the MBD in its folded state and are essential for MeCP2 plasticity. The R106W, R133C, F155S and T158M RTT mutations all showed aberrant MBD dynamics compared to wild type. Our results indicate that MBD–CTD interactions in *cis* and the unfolding/refolding transition of the MBD are important features of MeCP2 structure that become dysregulated in RTT.

5.2 Introduction

MeCP2 is a 53 kD nuclear protein found in large amounts in the lung, spleen, and especially the brain of vertebrates. MeCP2 is named for its ability to selectively bind methylated DNA (1), although it can bind to unmethylated DNA (2, 3). MeCP2 has important roles in both neurodevelopment and adult brain function, with specific effects on transcription documented in the hypothalamus, cerebellum, and hippocampus (4). The importance of MeCP2 to normal brain function is further underscored by the finding that loss-of-function mutations in the X-linked MeCP2 gene cause Rett syndrome (RTT), a severe neurodevelopmental disorder. Girls with RTT develop normally until the age of 6–18 months, and then begin to lose language and fine motor skills. Further regression results in a host of serious neurological and cardiac symptoms, including intellectual disability, motor impairment, seizures, and characteristic hand wringing (4). Consequently, there has

been intense interest in deciphering the molecular mechanisms through which MeCP2 influences cellular function in the normal and disease states. Much attention has been focused on the role of MeCP2 as a methyl DNA binding transcriptional repressor (5). In support of this idea, MeCP2 directly interacts with the nuclear co-repressor NCoR/SMRT *in vitro* and *in vivo* and many RTT mutations disrupt this interaction (5). However, transcriptomics studies in mouse models have shown that loss of MeCP2 alters the expression of a large number of genes, roughly half of which are upregulated and half downregulated, and the magnitude of these changes are small (6). To explain these results, MeCP2 has been proposed to be a global regulator of transcription acting through effects on chromatin architecture (6).

Less is known about MeCP2 structure compared to its functions. Full-length MeCP2 contains 486 amino acids and is a monomer in solution (7). Early studies identified two functional domains, the methyl DNA binding domain (MBD, residues 78–163) and the transcription repression domain (TRD, residues 207–309) (8). This led to the proposed domain organization shown in Fig. 5.1A, where residues 1–77 were labeled the N-terminal domain (NTD), residues 164–206 the intervening domain (ID) and residues 310–486 the C-terminal domain (CTD). Structure-based evidence for the same domain architecture was obtained after limited protease digestion of the purified protein, which also revealed the CTD could be subdivided into the CTD- α (residues 310–355) and CTD- β (residues 356–486) (7). MeCP2 is an intrinsically disordered protein (IDP). When analyzed by circular dichroism (CD), over 60% of the protein sequence was estimated to be disordered (7, 9). Characterization of MeCP2 by hydrogen/deuterium exchange (H/DX) demonstrated that the entire polypeptide chain exhibited very fast exchange kinetics indicative of a disordered structure, with the exception of the MBD, which showed slower exchange kinetics and was more structured (10). CD studies of the isolated NTD, ID, TRD, and CTD confirmed that they lack stable secondary structure (9). The 3D structure of the isolated MBD has been determined by NMR (11), and of the MBD bound to methylated DNA by X-ray crystallography (12). Thus, while we have an atomic level insight into how MeCP2 recognizes methylated DNA, the extensive disorder present throughout the protein sequence has prevented a rigorous understanding of how full length MeCP2 functions as a structural unit. The importance of understanding the structure of full length MeCP2 is

further underscored by the presence of RTT missense mutations in all five domains of the protein (13).

Here we have characterized the structure and dynamics of full length MeCP2 using high speed atomic force microscopy (HS-AFM) (14–16). HS-AFM previously has been employed to visualize myosin V walking on actin filaments (17), the structural changes of F₁-ATPase (18), and the conformational dynamics of the ClpB chaperone (19), and holds great promise for determining the structure and motions of intrinsically disordered proteins (20–22). Our HS-AFM analyses indicate that MeCP2 rapidly interconverts between many different structures, one of which has the shape of a dumbbell. The two more ordered parts of the dumbbell correspond to the MBD and CTD and are connected by the long flexible intrinsically disordered ID/TRD. The rapid conformational sampling of MeCP2 visualized by HS-AFM was not random, but rather was driven by transient intramolecular MBD–CTD interactions. Weak interaction of the isolated MBD and CTD in solution was documented by analytical ultracentrifugation and NMR. The MBD within the full length protein was unstable, undergoing an unfolding/refolding transition that could be observed and quantified by HS-AFM. The unfolding/refolding transition of the MBD was influenced by MBD–CTD interactions and was differentially altered by four missense mutations that cause RTT. Taken together our results provide novel insight into the structure and dynamics of MeCP2 and their possible misregulation in RTT.

5.3 Materials and methods

Protein expression and purification

Full-length human MeCP2 isoform e2, MeCP2-GFP, MeCP2 R294X, MeCP2 lacking amino acids 311–328 or 370–415 in the CTD, isolated MBD₇₄₋₁₇₁, isolated CTD₃₀₀₋₄₈₆, CTD₃₆₃₋₄₀₂ and full-length MeCP2 bearing RTT point mutations in the MBD (R106W, R133C, F155S, T158M) were expressed in *Escherichia coli* and purified using the Intein Mediated Purification with an Affinity Chitin-binding Tag (IMPACT) system followed by Heparin column (New England Biolabs) using a modification of the protocol described

previously (7). The MBD construct contained an added sequence, EFLEGSSC, on its C-terminal ends as a result of previously described cloning methods. *Escherichia coli* BL21RP cells were transformed with the ptyb1 plasmid vectors containing MeCP2 constructs using heat shock. The clones with the best inducing expression were selected for each construct and stored at -80°C . Bacteria were grown in lysogeny broth (LB) at 37°C to an optical density of 0.5 absorbance unit, induced with 0.4 mM isopropyl 1-thio-D-galactopyranoside and incubated at 30°C for 2–3 h prior to harvest. Expression hosts were pelleted in an Avanti J-26 XPI preparative centrifuge (Beckman Coulter) in a JLA-8.100 rotor at 5,000 g for 10 min. Pellets were resuspended in wash buffer (25 mM Tris, pH 7.5, 100 mM NaCl) and repelleted under the same conditions. Clean pellets were resuspended in column buffer (25 mM Tris-HCl, pH 7.5, 500 mM NaCl) supplemented with 0.5% Triton X-100, 0.2 mM PMSF, and Protease Inhibitor Mixture Set II and Set III (Calbiochem), followed by two rounds of sonication, 90 s each, using a Branson Sonifier 450 with a large tip at 50% duty cycle and a power output of 7. The lysate was transferred to Oakridge tubes and spun at 21,000 g for 25 min at 4°C in the preparative centrifuge in a JA-17 rotor (Beckman Coulter). The supernatant was mixed with 7 ml chitin beads (New England Biolabs) previously equilibrated in column buffer and incubated overnight at 4°C on a rotator on a low revolution rate. Chitin beads with supernatant mixture was applied on an empty chromatography column (Life Sciences), and supernatant was allowed to flow through. The column was washed with five column volumes of column buffer followed by five column volumes of column buffer containing 900 mM NaCl to remove bacterial DNA non-specifically bound to MeCP2. The chitin beads were washed with an additional 5 column volumes of 500 mM NaCl column buffer and incubated with column buffer containing freshly added 50 mM DTT for 72 hours to complete cleavage. Protein was eluted from the chitin column with column buffer, diluted from 500 mM to 300 mM NaCl, and loaded onto a HiTrap Heparin HP column (GE Healthcare) using HPLC. Proteins were eluted from the heparin column via step gradient from 300 mM NaCl to 1 M NaCl buffer using 100 mM NaCl steps in 25 mM Tris (pH 7.5), 10% glycerol background buffer. Peak fractions were pooled and dialyzed into 10 mM Tris (pH 7.5), 10 mM NaCl, 2% glycerol and 0.25 mM EDTA. The concentration of labeled protein was determined using a Pierce BCA test (Thermo Scientific).

For the samples subjected to analytical ultracentrifugation analysis, we used site-directed mutagenesis to replace with alanines (MonoC) all native Cysteines but Cys412 in the CTD (300–486) domain. Next, we expressed and purified the construct as described above and labeled the protein with Alexa 488-maleimide (Molecular Probes). The excess fluorophore was removed using HiTrap Desalting column (GE Healthcare). The concentration of labeled protein was determined using a BCA test, ensuring the fluorophore alone had no signal at 488 nm. The proteins were 90% clean as observed by SDS-PAGE, imaged with Typhoon 9500.

HS-AFM imaging

The HS-AFM imaging of protein molecules was performed as described (23). A glass sample stage (diameter, ~2 mm; height, ~2 mm) with a thin mica disc (1 mm in diameter and 0.05 mm thick) glued onto the top by epoxy was attached onto the top of the Z-scanner using a drop of nail polish. A freshly cleaved mica surface was prepared by removing the top layers of mica using Scotch tape. Then, a drop (~2 μ l) of each diluted sample (2–5 nM) in Buffer A (2 mM MgCl₂, 10 mM Tris-HCl, pH 7.5) was deposited onto the mica surface. After incubation for ~3 minutes, the mica surface was rinsed with 20 μ l of Buffer A to remove unattached protein molecules. The sample stage was then immersed in a liquid cell containing ~60 μ l of Buffer A. The HS-AFM observation was performed in the tapping mode using a laboratory built apparatus (23). The short cantilevers (BL-AC7DS-KU4) were custom-made by Olympus (Tokyo, Japan); resonant frequency ~1 MHz in water, quality factor ~2 in water, and spring constant 0.1–0.15 N/m. The cantilever's free oscillation amplitude A_0 was set at 1–2 nm and set point amplitude A_s was set at $\sim 0.9 \times A_0$, so that the loss of cantilever's oscillation energy *per* tap was adjusted at 1–3 $k_B T$ on average. The images were captured at a rate of 14.9 frames *per* sec (fps) or 10 fps for a scan area of $125 \times 125 \text{ nm}^2$ with a pixel size of 80×80 .

Analysis of AFM images

To measure topographical parameters of MeCP2 constructs from AFM images, a pixel-search software program was used (24). AFM images were first edited with a low-pass filter to remove spike noises and next with a flatten filter to make the overall xy-plane flat. The (X, Y, Z) coordinate of globular domains (MBD and CTD) in their folded states were measured semi-automatically using the following procedures. First, we selected manually the most probable highest point on each domain and several molecule-free positions in close proximity to each domain. Second, each highest point (X, Y, Z) was automatically determined by searching a 5×5 pixel area ($\sim 8 \times 8 \text{ nm}^2$) around the manually selected point. The (X, Y, Z) coordinate of the end region of IDR1 was also measured semi-automatically using the following procedure. First, we determined manually an end region of IDR1 and chose several molecule-free positions in close proximity to the end region. Next, the pixel search program automatically found a pixel position (X, Y) having the largest height value Z , as described above. The heights of globular domains (H_{MBD} and H_{CTD}) and the end of IDR1 (H_e) were obtained by subtracting respective average heights of the substrate surface from the corresponding Z values. To measure the end-to-end distance of IDR1 (R_{IDR1}), the direct distance D between the N-terminal end and the highest point within the MBD was measured. The value of R_{IDR1} was estimated as $R_{\text{IDR1}} = D - H_{\text{MBD}}/2 - H_e/2$. Similarly, the end-to-end distance of IDR2 (R_{IDR2}) was estimated as $R_{\text{IDR2}} = D - H_{\text{MBD}}/2 - H_{\text{CTD}}/2$, where D represents the direct distance between the highest points of MBD and CTD. The height of IDR (H_{IDR}) was obtained by subtracting the average height of the substrate surface from the Z values at positions along the ridgeline of the IDR. Note that the entire IDR1 is fully disordered judging from its height (0.4–0.5 nm). However, the CTD appeared to show partial order-disorder transitions with a small height change, so that IDR2 contains a region that is not fully disordered. Therefore, D_{IDR2} does not represent the length of a fully disordered IDR.

Transition rate determination

The mutants, R294X, F155S, and MeCP2-GFP, exhibited folding/unfolding transitions in their MBD. The autocorrelations $G(\tau)$ s of their time-series data of H_{MBD} were best fitted to single-exponential functions. However, the decay constants λ of these $G(\tau)$ s is the sum of respective rate constants of the low-to-high transition ($k_{\text{L} \rightarrow \text{H}} = 1/\tau_{\text{L}}$) and high-to-low transition ($k_{\text{H} \rightarrow \text{L}} = 1/\tau_{\text{H}}$); τ_{L} and τ_{H} are the lifetimes of partially unfolded and well-folded states, respectively. To determine the values of $k_{\text{L} \rightarrow \text{H}}$ and $k_{\text{H} \rightarrow \text{L}}$, we used the two Gaussian components of each H_{MBD} frequency distributions that overlap in a medium height region. The ratio $k_{\text{H} \rightarrow \text{L}}/k_{\text{L} \rightarrow \text{H}}$ (or $\tau_{\text{L}}/\tau_{\text{H}}$) was estimated from the area ratio ($A_{\text{L}}/A_{\text{H}}$) of the corresponding two Gaussian components, i.e., $k_{\text{H} \rightarrow \text{L}}/k_{\text{L} \rightarrow \text{H}} (\equiv \tau_{\text{L}}/\tau_{\text{H}}) = A_{\text{L}}/A_{\text{H}}$. For the cases of WT MeCP2 and d311–328 and d370–415 deletion mutants, their $G(\tau)$ s were best fitted to double-exponential functions. The values of rate constants of the stable (S)-to-unstable (U) ($k_{\text{S} \rightarrow \text{U}}$), U-to-S ($k_{\text{U} \rightarrow \text{S}}$), high-to-low ($k_{\text{H} \rightarrow \text{L}}$) and low-to-high ($k_{\text{L} \rightarrow \text{H}}$) state transitions were estimated as described in Supplementary Text S1.

Molecular dynamics simulations

Three systems were set up for molecular dynamics simulations: (i) the MBD domain alone, (ii) the MBD with the half the NTD domain (36 amino acids) and (iii) the MBD with the full NTD domain. The sequences used for these three cases are as follows:

MBD (PDBid: 1QK9):

ASASPKQRRSIIRDRGPMYDDPTLPEGWTRKLRKQKSGRSAGKYDVYLINPQGGK
AFRSKVELIAYFEKVGDTSLDPNDFDFTVTGRGSGSGC

MBD with half NTD:

GKHEPVQPSAHHSAEPAEAGKAETSEGSGSAPAVPEASASPKQRRSIIRDRGPMY
DDPTLPEGWTRKLRKQKSGRSAGKYDVYLINPQGGKAFRSKVELIAYFEKVGDTSL
LDPNDFDFTVTGRGSGSGC

MBD with NTD:

MVAGMLGLREEKSEDQDLQGLKDKPLKFKKVKKDKKKEEKEGKHEPVQPSAHH
SAEPAEAGKAETSEGSGSAPAVPEASAPKQRRSIIRDRGPMYDDPTLPEGWTRK
LKQRKSGRSAGKYDVYLINPQGKAFRSKVELIAYFEKVGDTSLDPNDFDFTVTG
RGS GSGC

The initial structure for the MBD was taken from the Protein Data Bank (PDBid: 1QK9) and the residues for the NTD domain were generated using the CNS-SOLVE software (25). The peptide was placed in a dodecahedral box in which the distance from the edges of the box to every atom in the protein was at least 1 nm. The box was solvated with explicit water and an excess ion concentration of 150 mM was added to reproduce physiological conditions. The simulations were performed using GROMACS 2016.3 software (26) with the TIP3P water model (27) and the Amber99SB*-ILDNP force field (28). The system was energy minimized and equilibrated in the NVT (constant particle number, volume and temperature) ensemble. Equilibration was followed by a production run in the NPT (constant particle number, pressure and temperature) ensemble with a time step of 2 fs. The particle-mesh Ewald method (29) was used with a cutoff of 1.2 nm. The temperature was set to 310 K with the V-rescale algorithm (30) and pressure was kept at 1 atmospheric pressure using the Parrinello-Rahman barostat (31). The MBD domain was simulated for 3 μ s and the MBD with NTD systems for 1 μ s. At least three separate sets of simulations starting from different initial conditions were run to ensure that the results were independent of the starting conformations.

A second set of systems was built for the case where the mica surface exists, using configurations from each of the above proteins systems after the first 100 ns of simulation. Each protein was placed in a cubic box with a minimum distance of 1 nm between its atoms and the edges of the box. A surface of size of each of the boxes was generated and a charge was added to the surface atoms to reproduce the experimental surface charge of -0.48 e/nm². The surface structure was modeled using graphene structure to ensure commensurability with periodic boundary conditions. The box was solvated with explicit water and an excess ion concentration of 150 mM. The system was energy minimized and

equilibrated with NVT dynamics. Equilibration was followed by a production run in the NVT ensemble with a time step of 2 fs. The temperature was set to 310 K with the V-rescale algorithm. The systems were run for 1 μ s and repeated as above.

Analytical ultracentrifugation

Sedimentation and diffusion transport in the ultracentrifugation cell are described by the Lamm equation, which can be solved using adaptive finite element methods (32). Whole boundary data obtained in SV experiments are fitted by linear combinations of such solutions using advanced optimization routines (33–35) that are typically implemented on a supercomputer (36). Sedimentation velocity experiments were performed using a Beckman XL-A analytical ultracentrifuge equipped with the Aviv fluorescence detector (37) at the Center for Analytical Ultracentrifugation of Macromolecular Assemblies at the University of Texas Health Science Center at San Antonio, using an An60Ti 4-hole rotor and standard 2-channel epon centerpieces with 1.2 cm pathlength (Beckman-Coulter). The CTD was fluorescently labeled with Alexa 488. The MBD was not labeled. Samples were prepared in 20 mM Tris-HCl, containing 100 mM NaCl, 0.5 mM EDTA, and 0.1 mM PMSF. A titration was performed where a constant amount of labeled CTD (200 nM) was mixed with 100, 200 and 400 μ M of unlabeled MBD. The experiment was performed in a 4-hole An60Ti rotor, at 50,000 rpm, 20°C, and fluorescence detection, collecting 964 scans for each sample before reaching equilibrium (~22 hours). All data were analyzed with UltraScan-III ver. 4.0, release 2655 (38). Hydrodynamic corrections for buffer density and viscosity were estimated by UltraScan to be 1.0017 g/ml and 1.0046 cP. The partial specific volumes of CTD (0.727 ml/g) and MBD (0.7252 ml/g) were estimated by UltraScan-III based on their amino acid sequence analogous to methods outlined in Laue et al. (39).

SV data were analyzed according to the workflow described in (40). Optimization was performed by 2-dimensional spectrum analysis (2DSA) (33) with simultaneous removal of time- and radially-invariant noise contributions (41). After inspection of the 2DSA solutions, a global genetic algorithm-Monte Carlo analysis was performed to quantify the relative concentrations of free and complexed CTD (35), and to obtain a model that

described all four titrations equally well. The calculations are computationally intensive and are carried out on high-performance computing platforms (36). All calculations were performed on the Lonestar cluster at the Texas Advanced Computing Center (TACC) at the University of Texas at Austin or on XSEDE clusters at TACC (Jetstream, Stampede 2) or the San Diego Supercomputing Center (Comet). Integral distributions of sedimentation coefficients were evaluated with the enhanced van Holde–Weischet method (42) to determine if shifts of sedimentation distributions occurred as a function of mass action.

NMR spectroscopy

The sub-domain of CTD, containing residues 363–402 (CTD2), was expressed in minimal media in the presence of ammonium N15-chloride, purified as described above and concentrated using Amicon Ultra 15 mL centrifugal filter (Millipore). ^1H , ^{15}N heteronuclear single quantum coherence (HSQC) spectra were collected on ^{15}N -labeled MeCP2 CTD (aa 363–402) at 100 μM , free and in the presence of MBD. Spectra were collected at 25°C on a 600 MHz Varian INOVA spectrometer equipped with a cryogenic probe. Data were processed using NMRPipe.

5.4 Results

Visualization and characterization of full length MeCP2 and its dynamics by HS-AFM

To determine the structural features of MeCP2, we first imaged full-length wild type (WT) MeCP2 using HS-AFM (16). The HS-AFM images document the highly dynamic behavior of MeCP2; we see rapid interconversion between ensembles of different structures, including a typical dumbbell-like structure (Fig. 5.1B). In the dumbbell structure, two intrinsically disordered regions (IDRs) mostly with a height of 0.4–0.5 nm (Fig. 5.1C) were observed: a short IDR1 at one end and a longer IDR2 connecting the two more ordered ends of the dumbbell. The two-dimensional end-to-end-distances of IDR1 (R_{IDR1}) and

IDR2 (R_{IDR2}) were $\langle R_{\text{IDR1}} \rangle = 12.5 \pm 3.6$ and $\langle R_{\text{IDR2}} \rangle = 18.5 \pm 8.0$ nm (mean \pm s.d.) (Fig. 5.1D). The broad distribution of R_{IDR2} values reflects the conformational distortions of the IDR2 resulting from the high degree of flexibility of this long segment. The globule between the IDR1 and IDR2 had a peak height of 1.4 nm in its height distribution (Fig. 5.1E), indicating it is well folded. Interestingly, this well folded globule itself is dynamic and undergoes unfolding/refolding transitions over time (Figs. 5.1E and 5.2A). The ordered region at the other end of MeCP2 had a height of 0.8 nm, indicating it is partially or “loosely” folded (Fig. 5.1F). We next assigned the N- and C-terminal ends of the MeCP2 molecule. HS-AFM images of a construct with GFP fused at the C-terminal end of MeCP2 indicated that the GFP tag was at the opposite end from the well folded globule (Fig. 5.2B, and Figs. S5.1 and S5.2A). In contrast, the TRD–CTD construct lacked the well folded globule (Fig. 5.2C and Fig. S5.3). From these results we conclude that IDR1 is the NTD, the well folded globule that undergoes folding/unfolding transitions is the MBD, IDR2 is composed of the ID, TRD and possibly a part of the CTD- α , and the loosely folded region is the CTD or CTD- β (see Fig. 5.2D). A summary of the HS-AFM results is shown in Fig. 5.2D.

The unfolding/folding transitions of the MBD within WT MeCP2 were further characterized by calculating the autocorrelation function of time-series of MBD height data (ACF_{MBD}). Notably, it showed a two-exponential decay (Fig. 5.2C and Supplementary Text S5.1), suggesting that MBD structural transitions are more complex than a simple equilibrium between two states, as detailed below in the next section. To determine if absorption onto mica influenced the observed folding/unfolding transition of the MBD, atomistic molecular dynamics simulations (MD) both in solution and in the presence of a surface (to model the HS-AFM experiments) were performed. Results indicated that the solution structure of the MBD was not perturbed by the mica surface over the time scale of the simulation (Fig. 5.2F and Supplementary Movie S5.1). This is consistent with previous H/DX studies indicating that the MBD samples folded and unfolded states in solution (8).

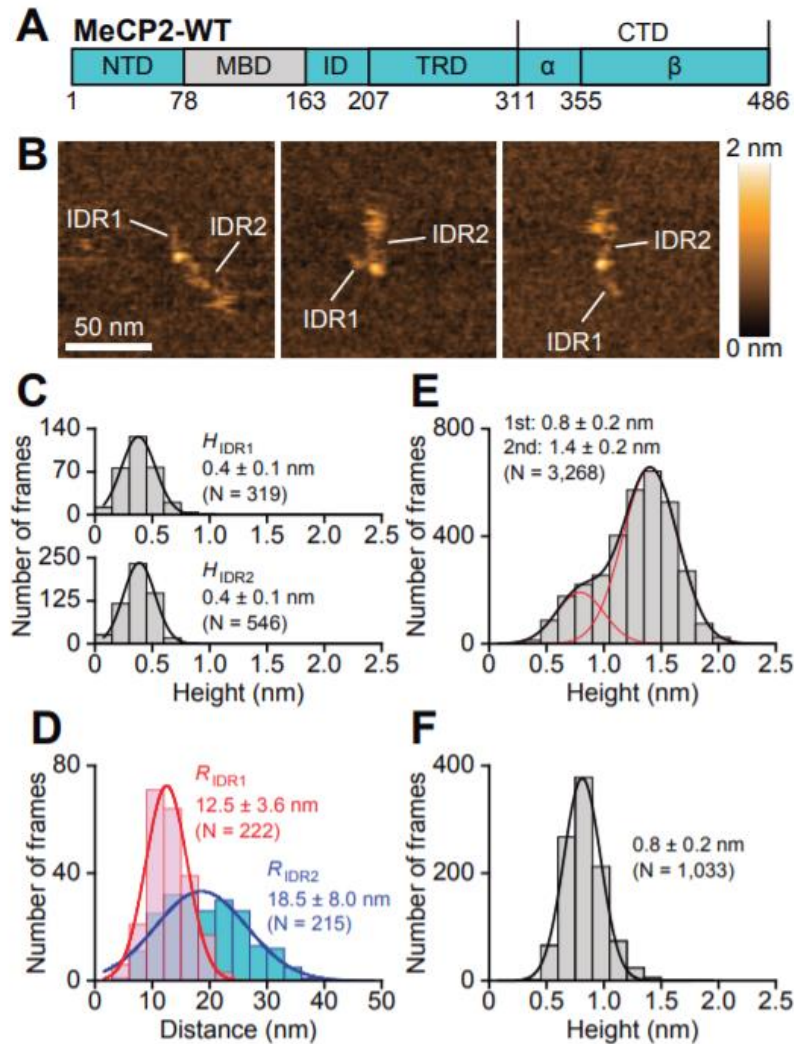


Figure 5.1 Structural features of WT MeCP2. (A) Domain diagram. (B) Typical HS-AFM images captured at 10 fps. (C) Height histograms for IDR1 (top) and IDR2 (bottom). (D) End-to-end distance histograms for IDR1 and IDR2. These distances were also measured for images in which the MBD and CTD appeared clearly as globules. (E) Height histogram of the globule locating between IDR1 and IDR2. This globule was identified as the MBD, as shown in Fig. 5.2D. (F) Height histogram for the loosely folded globule. This globule was identified as the CTD or a part of the CTD, as shown in Fig. 5.2D.

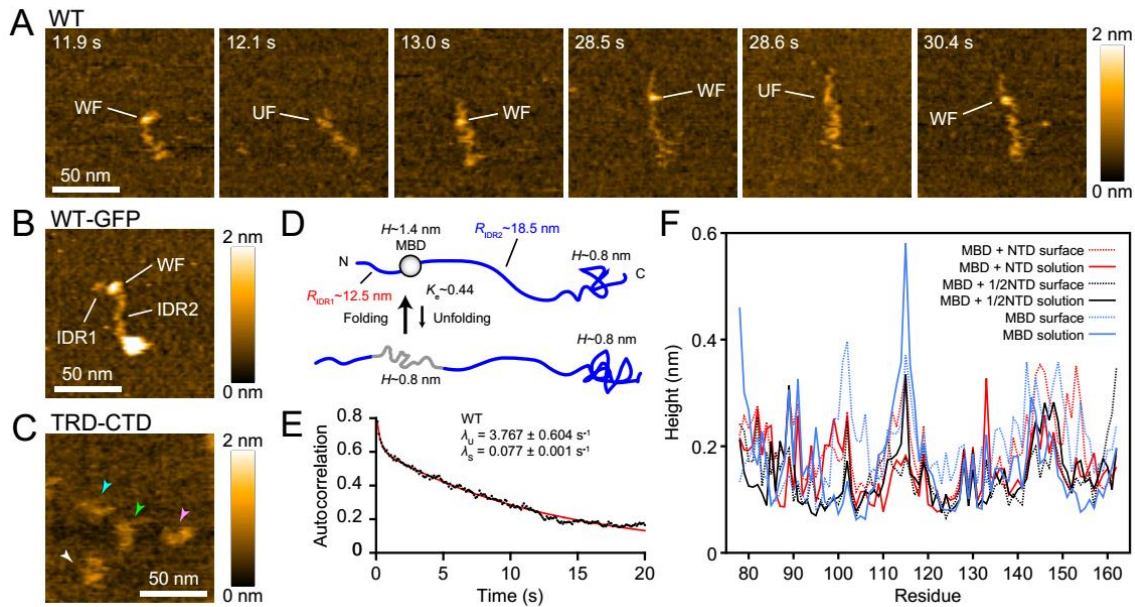


Figure 5.2 Domain identification and dynamic conformational changes of MBD in WT MeCP2. (A) HS-AFM images captured at 10 fps showing transitions of MBD between well folded (WF) and loosely folded (LF) conformations. (B) AFM image of MeCP2–GFP fusion. (C) AFM image of TRD–CTD. Arrowheads point individual molecules. (D) Schematic of MeCP2 dynamics. (E) ACF_{MBD} in WT MeCP2 (dots) and the best result of its fitting to a sum of two exponential functions (solid line) (see Supplementary Text S1). (F) Root mean square fluctuations (RMSF) of MBD height of three MeCP2-derived peptides in solution and in the presence of a surface, as observed by molecular dynamics simulation. The systems include MBD, MBD plus C-terminal half of NTD, and MBD plus NTD.

Intramolecular MBD–CTD interactions influence MeCP2 structural dynamics

Previous studies have shown that the intrinsic fluorescence intensity of MeCP2 due to W104 in the MBD is affected by removal of the CTD, suggestive of MBD–CTD interactions (43). To examine this question directly, we performed HS-AFM imaging of the RTT nonsense mutant, R294X, lacking the CTD. Remarkably, the MBD of R294X existed predominantly in the unfolded state (Fig. 5.3A, B), despite no mutations in the

MBD. This behavior was reflected in its ACF_{MBD} , which had a single-exponential decay (Fig. 5.3C). These results suggest that MBD–CTD interactions occur within the full-length MeCP2 in *cis*, shifting the MBD towards its well-folded conformation for a longer time. Further support for this conclusion came from three HS-AFM observations: (i) some of the HS-AFM images of full-length MeCP2 showed the MBD and CTD in contact to form a compact structure (Fig. 5.3D), (ii) removal of residues 311–328 in the CTD- α largely shifted the MBD equilibrium toward the unfolded state, and the equilibrium change was moderate when residues 370–415 in the CTD- β were removed (Supplementary Fig. S5.4), and (iii) the MeCP2-GFP fusion showed a MBD height distribution expected for the mostly unfolded state (Supplementary Fig. S5.2B) and the ACF_{MBD} with a single-exponential decay (Supplementary Fig. S5.2C), indicating that the GFP-dependent immobilization of MeCP2 on mica prevented MBD–CTD interactions.

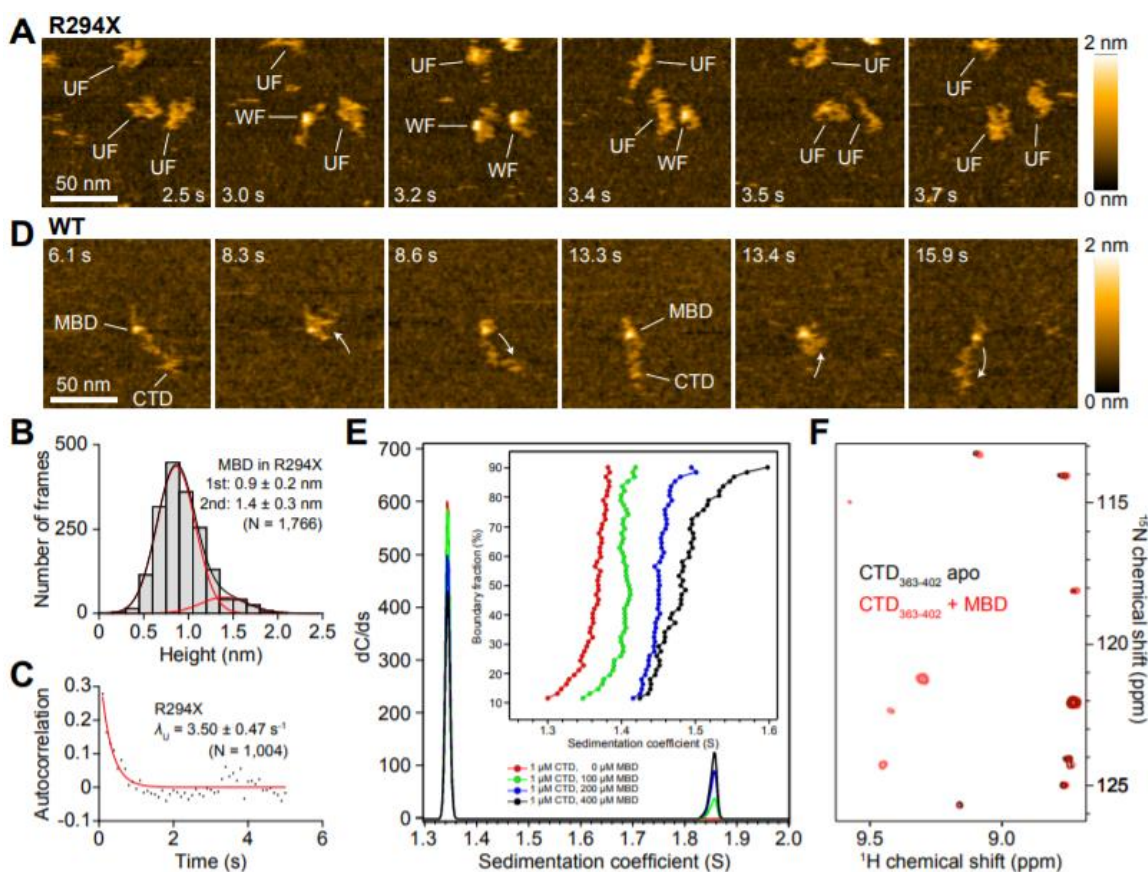


Figure 5.3 MBD–CTD interactions in *cis* and *trans*. (A) HS-AFM images of R294X RTT mutant captured at 10 fps. (B) Height histogram for MBD in R294X RTT mutant. The most probable fitting curve is shown with the solid black line. The red lines represent the Gaussian components in double-Gaussian fitting. (C) ACF_{MBD} in R294X (dots) and the best result of its fitting to a single exponential function (solid line). (D) HS-AFM images of WT MeCP2 showing transient MBD–CTD association (at 8.3 s and 13.4 s). (E) Sedimentation coefficient distributions of CTD (red) and three titration points with different concentrations of MBD (green, 100 μ M; blue, 200 μ M; black, 400 μ M). Inset: Integral distribution of sedimentation coefficient plot for the same experiment. (F) Superimposed ^1H , ^{15}N HSQC spectra of the uniformly ^{15}N -labeled CTD₃₆₃₋₄₀₂ region in the absence (black) and presence of MBD (red).

To determine if the MBD and CTD interact in solution as free domains, we purified the isolated CTD and MBD and characterized mixtures of the two using analytical ultracentrifugation (AUC) and NMR. The sedimentation coefficient distribution of the fluorescently labeled CTD (~20 kD) increased progressively when the CTD was titrated with increasing amount of unlabeled MBD (~10 kD), as would be expected for the formation of a reversibly associating complex (Fig. 5.3E, inset) (44). Global genetic algorithm analysis (33,38) of the same data fits all experimental concentrations simultaneously and indicated that the free CTD (1.35S) was progressively converted to a MBD–CTD complex (1.85S) with increasing amount of MBD (Fig. 5.3E and Supplementary Fig. S5.5A). From the titration curves, the K_D for the MBD–CTD interaction was estimated to be 1.29 ± 0.66 mM (Supplementary Fig. S5.5B), although the data were limited. To further test for MBD–CTD interactions in *trans*, we collected heteronuclear single quantum coherence (HSQC) spectra of uniformly ^1H , ^{15}N -labeled CTD₃₆₃₋₄₀₂ in the absence and presence of the MBD (Fig. 5.3F). The large chemical shift changes of CTD₃₆₃₋₄₀₂ in the presence of the MBD provided further evidence that the two domains directly interact in solution. Furthermore, appearance of a number of resonances downfield of 9 ppm in the ^1H dimension revealed that the MBD–CTD interaction induces partial CTD folding. Altogether, we conclude from the HS-AFM and solution-state studies that weak MBD–CTD interactions occur in both *cis* and *trans*. The *cis* interaction is facilitated by the highly flexible ID/TRD. The effective CTD concentration in the close vicinity of the MBD was estimated to be ~50 μM from the end-to-end distance distribution of the IDR2 (Fig. 5.1D and Supplementary Text S5.2). From this value and the estimated K_D of ~1.3 mM, the MBD in ~4% of MeCP2 molecules is bound to the CTD in the steady state.

The ACF_{MBD} of WT MeCP2 decayed with two rate constants, $\lambda_S = 0.077$ s⁻¹ and $\lambda_U = 3.77$ s⁻¹ (Fig. 5.2E), whereas the ACF_{MBD} of the R294X RTT mutant showed a single-exponential decay with a rate constant of 3.50 s⁻¹ (Fig. 5.3C), nearly identical to the value of $\lambda_U = 3.77$ s⁻¹. These results suggest that when not interacting with the CTD, the MBD is in an unstable (U) state undergoing fast transitions between folded (high) and unfolded (low) conformations ($\text{L} \leftrightarrow \text{H}$). The transient MBD–CTD interaction converts the MBD from the U state to a stable (S) state ($\text{U} \rightarrow \text{S}$). In the S state, the folded (high) conformation

of the MBD is sustained even after the dissociation of CTD (i.e., structural plasticity), and decays slowly to the U state ($S \rightarrow U$). According to this model, the rate constants for $L \rightarrow H$, $H \rightarrow L$, $U \rightarrow S$, and $S \rightarrow U$ transitions in the WT were determined as $k_{L \rightarrow H} = 0.45 \text{ s}^{-1}$, $k_{H \rightarrow L} = 3.32 \text{ s}^{-1}$, $k_{U \rightarrow S} = 0.059 \text{ s}^{-1}$, and $k_{S \rightarrow U} = 0.018 \text{ s}^{-1}$ (Supplementary Text S1 and Table S1).

Aberrant dynamics of RTT mutant MBDs

We next determined whether the unfolding/refolding transition of the MBD was influenced by RTT mutations in the MBD (Fig. 5.4). Unfolding/refolding of the WT MBD occurred at the sub-second to a second time scale in the U state, while in the S state stabilized by transient MBD–CTD interaction the well-folded conformation is sustained for the tens of seconds time scale as described above. On average, in the WT the MBD occupied the well folded (1.4 nm height) conformation ~80% of the time and the unfolded (0.8 nm height) conformation about ~20% of the time (Fig. 5.1E). HS-AFM imaging of full-length MeCP2 bearing RTT point mutations in the MBD (R106W, R133C, F155S, and T158M) (Fig. 5.4) revealed that abnormal MBD dynamics is a common feature of all RTT mutants analyzed, although the nature of the defect was mutant-specific. The R133C and T158M mutants behaved most similarly, showing predominantly lower MBD height distributions compared to the WT MBD (Fig. 5.4C, D). This indicates that the MBDs of these two mutants spend most of their time in the unfolded state, i.e., the mutations destabilize the well folded MBD structure. The F155S mutant existed in two populations, an unfolded state and a misfolded state with a peak height of only 1.0 nm (Fig. 5.4E, F and Supplementary Table S5.1). By contrast, the R106W mutation stabilized the MBD in a well folded conformation as indicated by the predominance of a MBD species with ~1.4 nm height (Fig. 5.4B). These results argue that the inherent transitioning of the WT MBD between its well folded and unfolded states (and stable and unstable states) is required for proper MeCP2 function, and when compromised may contribute to RTT.

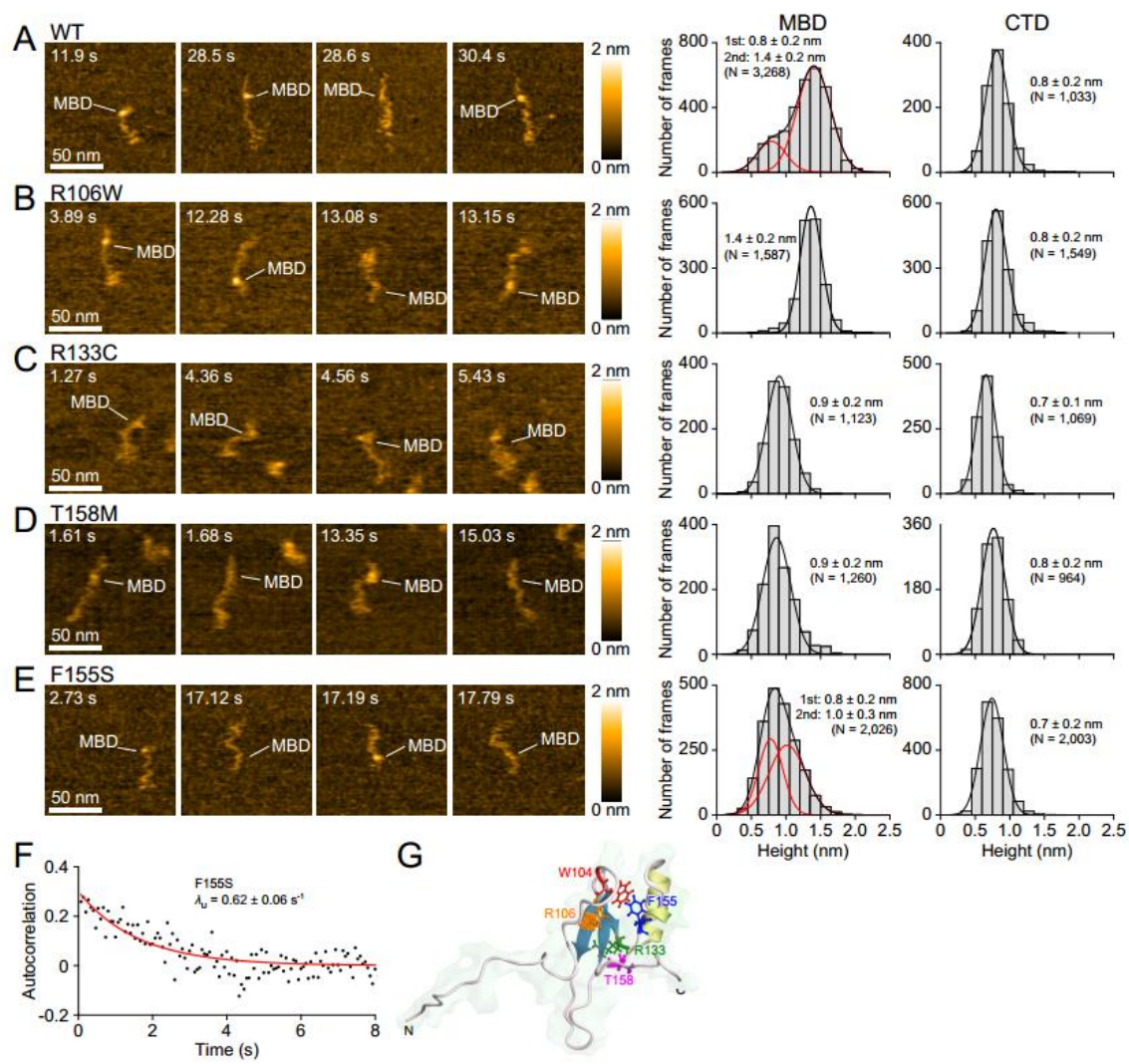


Figure 5.4 Structural features of MBD and CTD in MeCP2 bearing RTT point mutations in MBD, and their comparison to the WT. Left panels, typical HS-AFM images captured at 10 fps (A) or 15 fps (B–E). Middle panels, MBD height histograms. The black lines are most probable fitting curves, while the red lines (A, E) show Gaussian components in double-Gaussian fitting. The black lines are most probable fitting curves. Right panels, height histograms for C-terminal globule. (F) ACF_{MBD} in F155S (dots) and the best result of its fitting to a single component exponential function (solid line). (G) NMR structure of MBD.

5.5 Discussion

As expected from its intrinsically disordered nature, the conformation of MeCP2 is highly dynamic. In its most elongated state MeCP2 resembles a dumbbell. The MBD and CTD form the ends of the dumbbell and the long intervening ID/TRD is maximally extended (Fig. 5.1B). In its most compact state, the MBD and CTD are in contact, and the ID/TRD appears to be folded upon itself or interacting with other parts of the protein (Fig. 5.3D). Collectively, our data indicate that the conformational motions of MeCP2 (i) result from equilibration between the extended and compact structures, (ii) are driven by MBD–CTD interactions, and (iii) are facilitated by the flexible, intrinsically disordered ID/TRD segment. The MBD–CTD interaction is direct and not an artifact of MeCP2 being adsorbed to the mica surface. Indeed, the isolated MBD and CTD interact when free in solution as judged by analytical ultracentrifugation and NMR (Fig. 5.3E, F). Moreover, Ghosh et al. (43) showed that removal of the CTD reduced the fluorescence emission maxima of W104 in the MBD, indicating changes in the local tryptophan environment upon CTD deletion and implying that MBD–CTD interactions occur within MeCP2 under solution conditions (43).

The MBD is the only MeCP2 domain with classical tertiary structure. The NMR and X-ray structures of the MBD reported a three-stranded antiparallel β -sheet packed against an 11-residue α -helix (Fig. 5.4G) (11, 12). Extending from the C-terminal end of the α -helix is a short 3-10 helix followed by an Asx-ST motif. The first and second β -strands are connected by an elongated nine residue loop that fits in the major groove of DNA (12). Our HS-AFM experiments have revealed that the MBD transitions between its well folded conformation and an unfolded state, indicating that the MBD within full length MeCP2 is only marginally stable. These results are in close agreement with previous H/DX analyses of full length MeCP2 in solution. Whereas a stably folded protein exhibits slow exchange kinetics, a moderate level of exchange occurred throughout the MBD (10), demonstrating that the MBD samples unfolded and folded states. The R106W, R133C, F155S, and T158M RTT mutants each affected the MBD folding/unfolding equilibrium, although in different ways. The R133C and T158M mutants greatly destabilized the folded state of the MBD (Fig. 5.4C, D). The chief characteristic of the F155S mutant was that the MBD was

misfolded and more unstable (Fig. 5.4E). All three of these mutations are located in positions that would be expected to disrupt proper folding of the MBD (Fig. 5.4G). Strikingly, the stability of the MBD was enhanced by the R106W mutation, such that the unfolded state of the MBD could only rarely be detected in this mutant (Fig. 5.4B). When analyzed by H/DX, the isolated MBD bearing the R106W mutation showed very similar exchange kinetics as wild type except for increased protection of the residues in the β 1 strand surrounding the mutation (10). These observations suggest that the β -sheet found in the MBD is stabilized by the R106W mutation (Fig. 5.4G), perhaps through gain of pi-pi interaction with W104, resulting in a decreased propensity of the MBD to unfold. Taken together, our results imply that misregulation of the MBD unfolding/folding transition—in either direction—may contribute to RTT.

The unfolding/refolding transition of the MBD appears to be integral to the mechanism of MeCP2 binding to unmethylated and methylated DNA. In the H/DX experiments performed with full length WT MeCP2, the protection observed throughout the MBD was enhanced by DNA binding and enhanced further by methylated DNA binding (10). This would be expected if the MBD was transitioning between unfolded and folded states and binding to DNA and methylated DNA sequentially stabilized the folded conformation. This observation is consistent with the results of Ghosh et al. (43), who found that binding to unmethylated and methylated DNA successively increased the T_m of MeCP2 in thermal melting experiments. Despite having a stable well folded structure (Fig. 5.4B) that is almost identical to wild type (10), the R106W mutant does not bind normally to either unmethylated or methylated DNA *in vitro* (43, 45), is not retained in chromatin *in vivo* (46), and has a severe RTT phenotype (43) consistent with disrupted MBD function. These results are difficult to reconcile with a classical one-step mechanism of protein-DNA recognition mediated by a stable DNA binding domain. At the same time, the R133C, F155S, and T158M mutations all destabilize the well folded MBD state (Figs. 5.4C–E), and all show impaired binding to methylated DNA (43, 45), indicating the importance of the MBD fold for methylated DNA recognition. One possible explanation is that the MBD initially engages with DNA when it is unfolded and subsequently assumes the MBD fold to create a stable complex with methylated DNA. We note that the properties of the RTT mutants identified in our studies have implications for disease treatment. Based on the HS-

AFM behavior of the RTT mutants, small molecules that stabilize the MBD fold may prove useful for treating patients with the R133C, F155S, and T158M mutations, while patients with the R106W mutation may benefit from small molecules that destabilize the MBD.

The inter-domain interactions of MeCP2 are likely to be functionally important. MeCP2 condenses chromatin fibers into unique higher order structures characterized by edge-to-edge clustering of neighboring nucleosomes (47), although how it accomplishes this is unknown. Both the MBD and CTD bind to DNA and nucleosomes (48). We therefore speculate that the MBD and CTD bind to different nucleosomes when in the extended MeCP2 conformation, and subsequent MBD–CTD interactions help bring the nucleosomes together to condense the fiber. The plasticity observed in our experiments (Supplementary Text S5.1) may help stabilize such structures once formed. MeCP2 recruits the transcription factors to methylated DNA *in vivo* (49). The interaction sites for some of these factors, such as transcription co-repression complex NCoR, are located near the TRD–CTD boundary (50). In these cases, the transcription factors will be brought into physical proximity with methylated DNA upon MBD–CTD contact. If the MBD and CTD in the extended MeCP2 conformation bind to two linearly distant genomic loci that are in close proximity in three dimensions, MBD–CTD interactions and recruitment of the CCCTC-binding transcription factor, CTCF (51), may facilitate formation of chromatin loops. Taken together, the function of MeCP2 in transcription factor recruitment may involve more than a simple tethering process. In a broader sense, we speculate that the conformational dynamics of MeCP2 provides the structural basis for its multifunctionality. We note that the situation in which two structured domains are separated by a long intrinsically disordered region is predicted to be common among IDPs (51). Consequently, it seems likely that the intramolecular domain-domain interactions observed in our studies may be shared by many disordered proteins. We postulate that conformational malleability driven by domain-domain interactions *in cis* and structural flexibility of the intervening polypeptide chain is a common feature of many IDPs and is essential for mediating their functionality in three dimensions.

5.6 Supplemental information

The pixel-search program for AFM images can be accessed at the following URL:

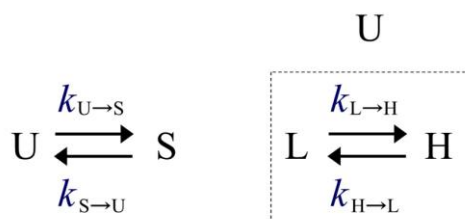
<https://elifesciences.org/content/4/e04806/article-data#fig-data-supplementary-material>

Movies from the molecular dynamics simulations can be accessed at the following URL:

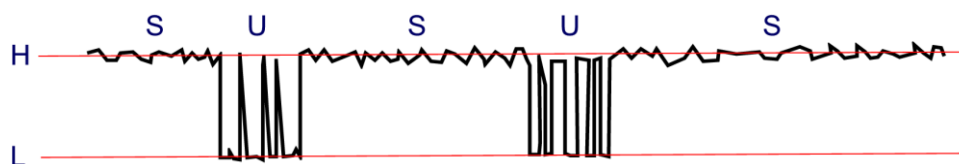
<https://doi.org/10.5281/zenodo.5774094>

S5.1. Analysis of folding/unfolding dynamics of MBD

The state transitions of MBD contained in the WT MeCP2 are considered to take place according to the model shown below:



The transition from the unstable (U) state to the stable (S) state is triggered by transient interactions between the MBD and the CTD, whereas the transition from S to U state takes place autonomically. The U state is in dynamic equilibrium between the unfolded state (low state L) and the well-folded state (high state H). In the S state, the MBD is well folded and its height is identical to that in the H state (1.4 nm). The time course of MBD height variations and state transitions are schematized below.



Let's define N_U as an average number of transitions ($L \leftrightarrow H$) occurring during the single U state. The following relationship holds:

$$1/k_{U \rightarrow S} \text{ (lifetime of U state)} = N_U \times (1/k_{L \rightarrow H} + 1/k_{H \rightarrow L}). \quad (S1)$$

Eq.(S1) can be rewritten as

$$N_U = k_{L \rightarrow H} \times k_{H \rightarrow L} / [k_{U \rightarrow S} \times (k_{L \rightarrow H} + k_{H \rightarrow L})]. \quad (S2)$$

The total time during which the MBD assumes the unfolded (low) conformation in the single U state (T_L) is identical to $N_U/k_{L \rightarrow H}$ on average, while the total time during which the MBD assumes the well-folded (H) conformation in the single U and S states (T_H) is $N_U/k_{H \rightarrow L} + 1/k_{S \rightarrow U}$ on average.

The area ratio ($\alpha_{WT} \equiv A_L/A_H = 0.24$) of two Gaussian components of MBD height distribution in WT MeCP2 (**Fig. 5.1E** in the main text) is identical to T_L/T_H . Therefore, we obtain

$$\alpha_{WT} = (N_U/k_{L \rightarrow H}) / (N_U/k_{H \rightarrow L} + 1/k_{S \rightarrow U}) = \\ k_{S \rightarrow U} \times (k_{H \rightarrow L}/k_{L \rightarrow H}) / [k_{U \rightarrow S} + k_{S \rightarrow U} + k_{U \rightarrow S} \times (k_{H \rightarrow L}/k_{L \rightarrow H})]. \quad (S3)$$

For the case of WT MeCP2, the autocorrelation function of time-series of MBD height data (**Fig. 5.2D** in the main text) was best fitted to the sum of two exponential functions, $A \times \text{Exp}(-\lambda_S t) + B \times \text{Exp}(-\lambda_U t)$, from which we obtained the values of $\lambda_S = 0.077 \text{ s}^{-1}$ and $\lambda_U = 3.77 \text{ s}^{-1}$. Since the two different transitions, $L \leftrightarrow H$ in the U state and $U \leftrightarrow S$, take place independently, the autocorrelation function of this kinetic system is expressed as a sum of two exponential functions, consistent with the above result. The decay rates in the autocorrelation function, λ_S and λ_U , are therefore, expressed as

$$\lambda_S = k_{U \rightarrow S} + k_{S \rightarrow U} \quad (S4)$$

and

$$\lambda_U = k_{L \rightarrow H} + k_{H \rightarrow L}. \quad (S5)$$

The conformational transitions $L \leftrightarrow H$ in the U state of the WT MeCP2 are considered to be identical to those occurring in the MBD contained in R294X, because R294X lacks the CTD and thus its MBD always stays in the U state. In fact, its autocorrelation function of time-series of MBD height data showed a single-exponential decay (**Fig. 5.3C** in the main text). Moreover, its decay rate ($\lambda_U = 3.50 \text{ s}^{-1}$) was close to the value, $\lambda_U = 3.77 \text{ s}^{-1}$, estimated for the WT MeCP2. Therefore, the lifetime ratio [$\tau_L/\tau_H = k_{H \rightarrow L}/k_{L \rightarrow H}$] in the U state of MBD in WT MeCP2 can be approximately obtained from the area ratio ($\alpha_{294X} \equiv A_L/A_H = 7.43$) of two Gaussian components of the MBD height distribution in R294X (**Fig. 5.3B** in the main text), i.e.,

$$k_{H \rightarrow L}/k_{L \rightarrow H} = \alpha_{294X}. \quad (S6)$$

Thus, $k_{L \rightarrow H}$ and $k_{H \rightarrow L}$ are expressed as

$$k_{L \rightarrow H} = \frac{\lambda_U}{1 + \alpha_{294X}}, \quad (S7)$$

and

$$k_{H \rightarrow L} = \frac{\alpha_{294X} \cdot \lambda_U}{1 + \alpha_{294X}}. \quad (S8)$$

From *Eqs.* (S3), (S4), and (S6), we finally obtained the following relationships for $k_{U \rightarrow S}$ and $k_{S \rightarrow U}$ in the WT MeCP2:

$$k_{U \rightarrow S} = \lambda_S \cdot \frac{(\alpha_{294X} - \alpha_{WT})}{\alpha_{294X}(1 + \alpha_{WT})}, \quad (S9)$$

and

$$k_{S \rightarrow U} = \lambda_S \cdot \frac{\alpha_{WT}(1 + \alpha_{294X})}{\alpha_{294X}(1 + \alpha_{WT})}. \quad (S10)$$

Using these equations and the values of $\lambda_S = 0.077 \text{ s}^{-1}$, $\lambda_U = 3.77 \text{ s}^{-1}$, $\alpha_{WT} = 0.25$, and $\alpha_{R294X} = 7.43$, we obtained $k_{U \rightarrow S} = 0.059 \text{ s}^{-1}$ ($\tau_U = 16.9 \text{ s}$), $k_{S \rightarrow U} = 0.018 \text{ s}^{-1}$ ($\tau_S = 55.6 \text{ s}$), $k_{L \rightarrow H} = 0.45 \text{ s}^{-1}$ ($\tau_L = 2.22 \text{ s}$), and $k_{H \rightarrow L} = 3.32 \text{ s}^{-1}$ ($\tau_H = 0.30 \text{ s}$) for the MBD in WT MeCP2. For R294X, we obtained $k_{L \rightarrow H} = 0.42 \text{ s}^{-1}$ ($\tau_L = 2.38 \text{ s}$), and $k_{H \rightarrow L} = 3.09 \text{ s}^{-1}$ ($\tau_H = 0.32 \text{ s}$).

In the MeCP2–GFP fusion, the GFP moiety is firmly attached to the mica surface, which largely suppresses Brownian motion of CTD on mica. Therefore, MBD–CTD interactions are hampered. In fact, the MDB height distribution in this construct (**Fig. S5.2B**) was similar to that of R294X (**Fig. 5.3B** in the main text), and the area ratio ($\alpha_{GFP} \equiv A_L/A_H$) was 10.24. The autocorrelation function of time-series of MBD height variations in the MeCP2–GFP fusion was best fitted to a single exponential function with a decay rate of $\lambda_U = 3.08 \text{ s}^{-1}$ (**Fig. S5.2C**). From these results and Eqs. (S7) and (S8), we obtained $k_{L \rightarrow H} = 0.27 \text{ s}^{-1}$ and $k_{H \rightarrow L} = 2.81 \text{ s}^{-1}$, which are roughly identical to the corresponding values estimated for R294X. Similarly, we obtained the values of rate constants for d311–328, d370–415, and F155S. These results are summarized in Supplementary **Table S5.1**.

S5.2. Effective concentration of CTD around MBD

The MBD and CTD are linked with the highly flexible IDR2 chain. Therefore, the effective concentration (C_{eff}) of the CTD around the MBD must be high. We here estimate the value of C_{eff} , as shown below.

1. The first approximation

We assume that IDR2 encompasses residues 164–309 (ID/TRD), although it may also contain a part of the CTD- α . From the number of residues contained in the ID/TRD, $N_{\text{aa}} = 146$, the stretched IDR2 length (contour length) is estimated to be $L = 52.6 \text{ nm}$, using the relationship of $L = (N_{\text{aa}} - 1) \times d_{\text{aa}}$, where $d_{\text{aa}} (= 0.36 \text{ nm})$ is an average distance between adjacent residues. In this first approximation, we further assume that the IDR2 is extremely flexible, so that the probability of finding the CTD is nearly uniform over the spherical

space of radius L , centered at the MBD. The effective concentration C_{eff} can be simply calculated as

$$C_{\text{eff}} = 10^{-3} / \left(\frac{4\pi}{3} L^3 \cdot N_A \right), \quad (S11)$$

where N_A is the Avogadro constant. Here we neglected the dimensions of MBD and CTD, as they are much smaller than L . Eq. (S11) gives $C_{\text{eff}} \approx 2.7 \mu\text{M}$ for $L = 52.6 \text{ nm}$.

2. The second approximation

In reality, the CTD is not uniformly distributed around the MBD. In fact, the two-dimensional distance R_{IDR2} between the two globular domains (MBD and CTD) showed a Gaussian distribution with a peak at 18.5 nm (i.e., mean R_{IDR2} , $\langle R_{\text{IDR2}} \rangle$), ~ 3 -times shorter than the full-stretched length L . Supposing that IDR2 is neither extended nor compacted by contact with the mica surface, the mean end-to-end distance of IDR2 in solution (i.e., not on mica) is given by $\langle R \rangle = \langle R_{\text{IDR2}} \rangle / \sqrt{2} = 13.08 \text{ nm}$. However, our measurements of two-dimensional end-to-end distances $\langle R_{2D} \rangle$ for various fully disordered IDRs have indicated that mica-IDR interactions extend $\langle R_{2D} \rangle$ by a factor 1.24. This is due to frictional forces locally exerted from mica against fast Brownian motion of the IDR chain, which would increase the IDR chain's undulation wavelength and/or decrease the undulation amplitude, resulting in swelling of its two-dimensional dimensions [Kirk, J. and Ilg, P. (2017) Chain dynamics in polymer melts at flat surfaces. *Macromolecules*, **50**, 3703–3718]. Therefore, $\langle R \rangle$ is estimated to be $13.08/1.24 = 10.55 \text{ nm}$. An fully disordered IDR is considered to behave as an ideal Gaussian chain. In this case, the distribution function of R is given by

$$P(R)dR = 4\pi R^2 \left[\frac{3}{2\pi \langle R^2 \rangle} \right]^{3/2} \exp \left[-\frac{3}{2} \frac{R^2}{\langle R^2 \rangle} \right] dR, \quad (S12)$$

although an region in IDR2 close to (or contained in) the CTD- α may not be fully disordered. The two globular domains associate with each other when they approach within a certain range of distance, ($r \leq r_0$). However, the distance r between the two domains is a complex function of R , when we consider the actual dimensions of the two globules. To avoid this complexity, we assume that MBD-CTD association occurs when R becomes R_0

or shorter (i.e., $R \leq R_0$). In this approximation, the probability of finding the CTD in the volume ($R \leq R_0$) is given by

$$p = \int_0^{R_0} P(R) dR. \quad (S13)$$

Therefore, the effective concentration of the CTD in the volume in close proximity to the MBD is given as

$$C_{\text{eff}} = p \times 10^{-3} / \left(\frac{4\pi}{3} R_0^3 \cdot N_A \right). \quad (S14)$$

For $R_0 = 1$ nm, *Eq.(S13)* provides $p = 0.32 \times 10^{-3}$, and *Eq.(S14)* provides $C_{\text{eff}} \approx 120$ μM . Note that $C_{\text{eff}}(R_0)$ slowly decays with increasing R_0 . Therefore, the approximate value of ~ 120 μM holds for the range of $0 < R_0 < 2.0$ nm. When we consider the actual dimensions of the two globules, the value of C_{eff} is reduced. Since the CTD binding site on the MBD is localized at its certain surface area, the CTD available for MBD binding is reduced approximately to a half. Therefore, the likely value of C_{eff} is approximately ~ 50 μM .

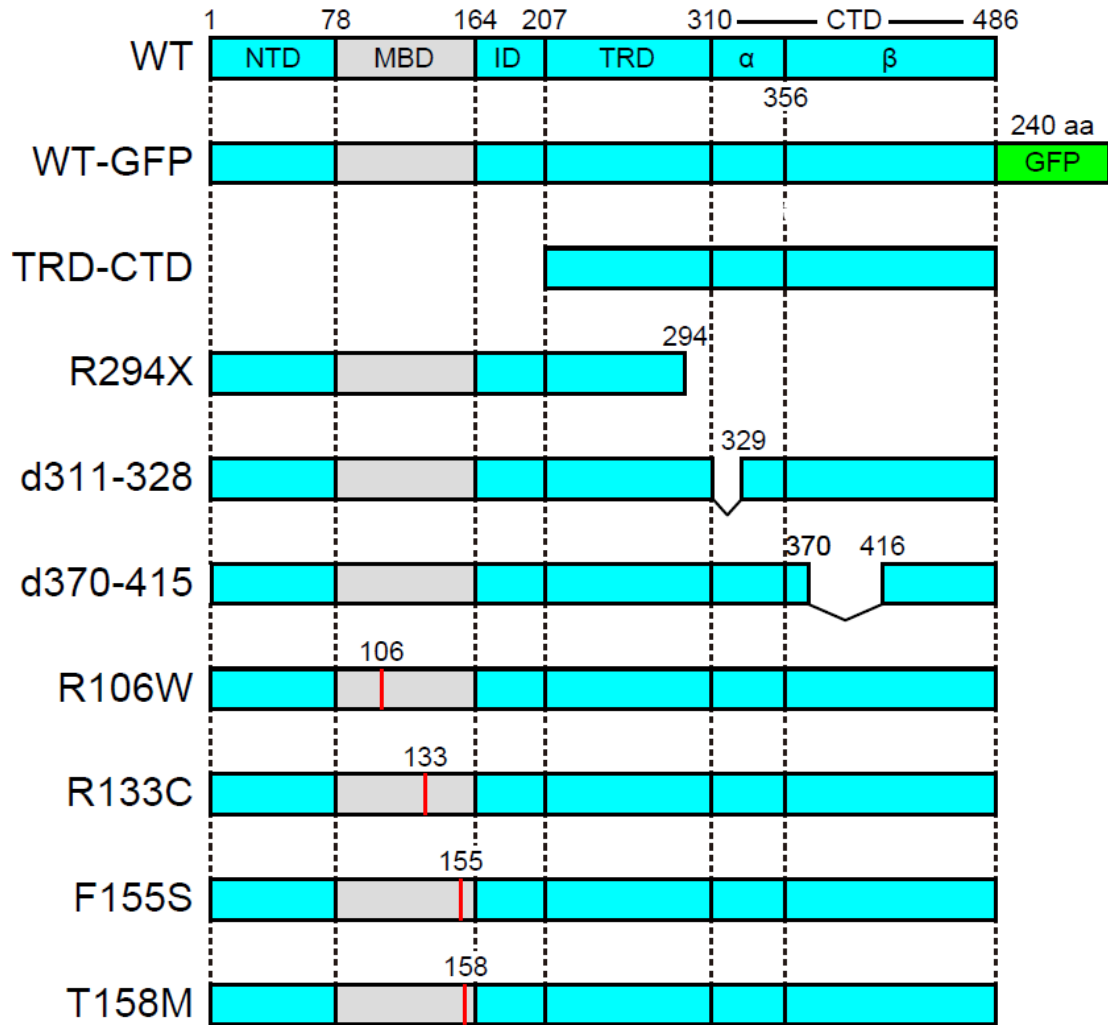


Figure S5.1 Domain diagrams of wild type MeCP2 and its mutants used in this study.

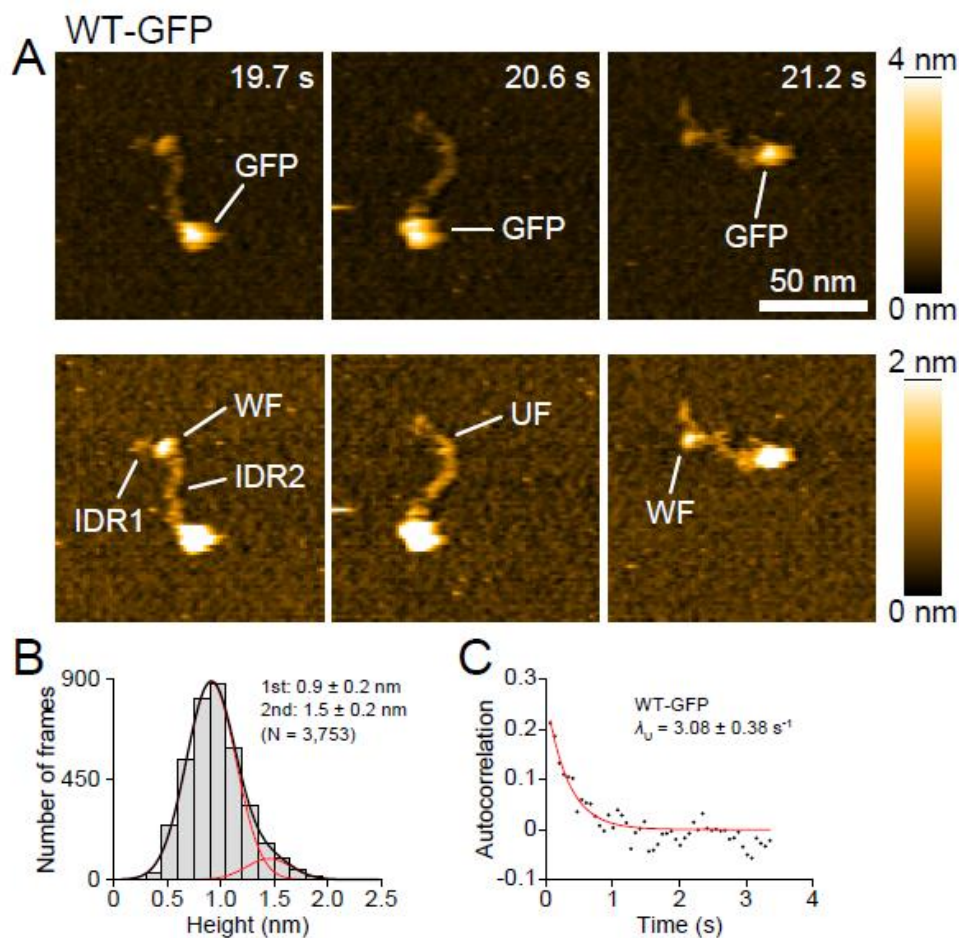


Figure S5.2 Structural features of MeCP2 fused to GFP at the C-terminus of the former. (A) Typical HS-AFM images captured at 15 fps (top). The images (bottom) were obtained by increasing the brightness contrast on the corresponding images (top). (B) MBD height histogram. Note that the MBD is mostly in the lower (unfolded) state, in contrast to the case of WT MeCP2. (C), Autocorrelation function of MBD height variation over time. The best result of its fitting to a single exponential function is shown in the solid line (also see Supplementary Text S1).

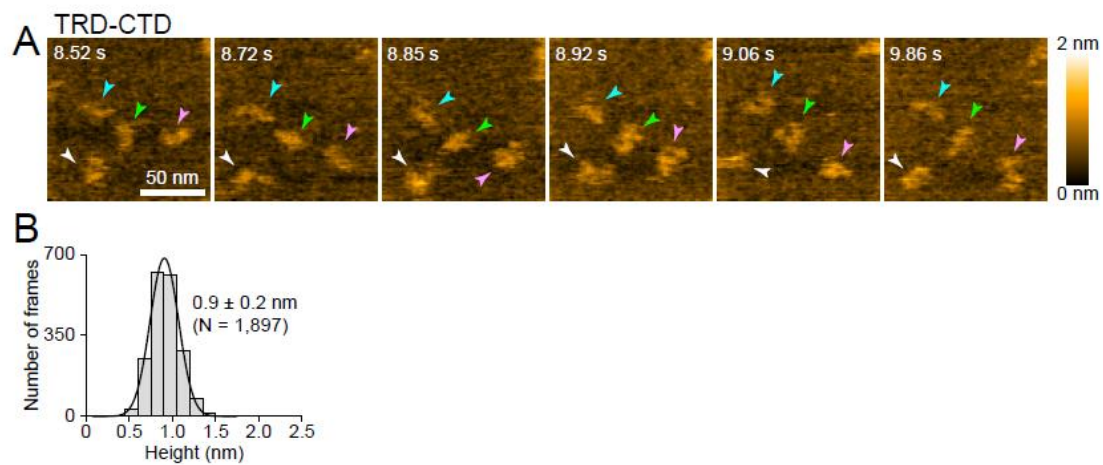


Figure S5.3 Structural features of TRD–CTD. (A) Typical HS-AFM images captured at 15 fps. Four individual molecules are marked with arrow heads with different colors. (B) Height histogram for highest pixel positions in the molecules.

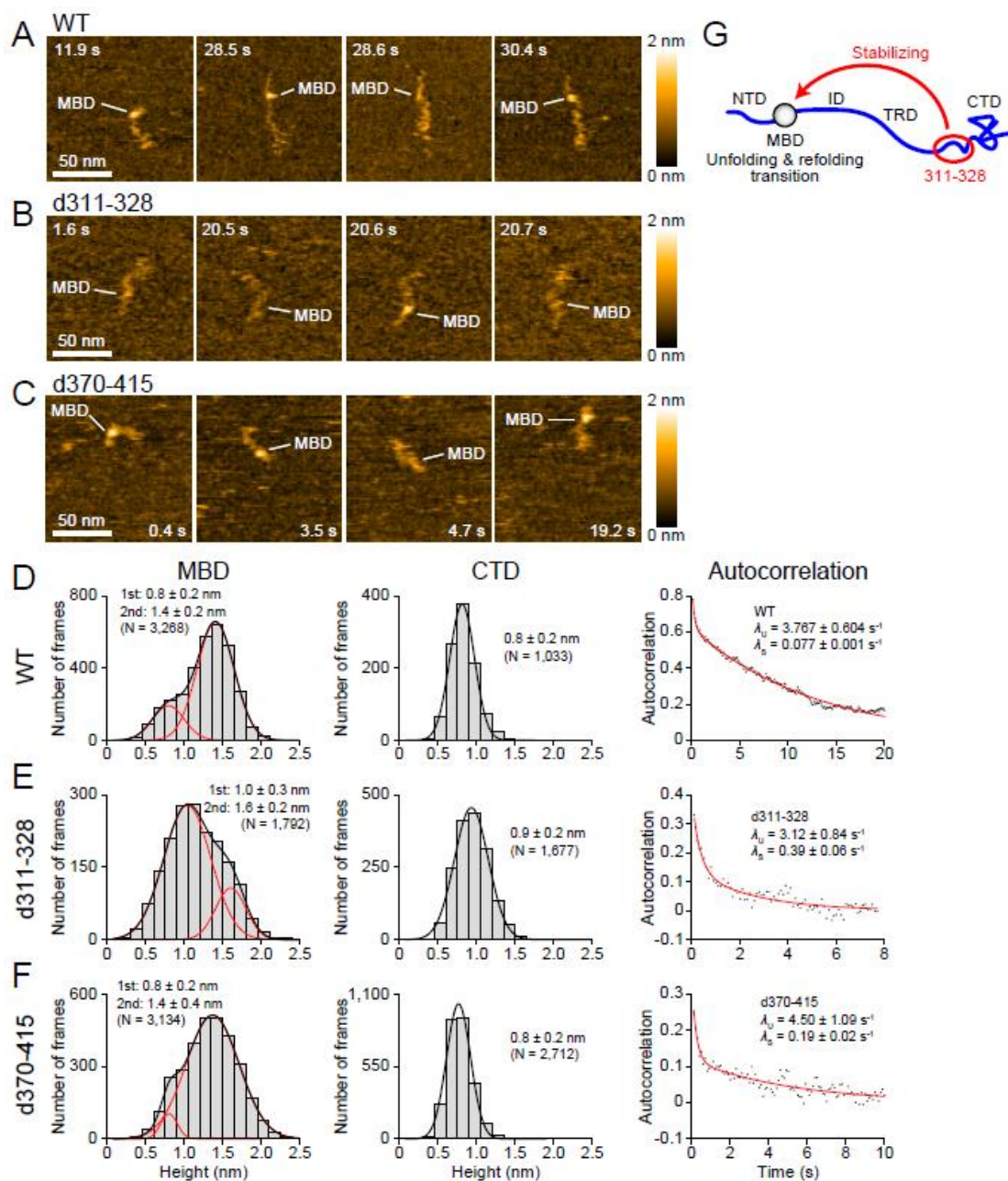


Figure S5.4 Structural features of d311–328 and d370–415 mutants and their comparison to the wild type. (A–C), Typical HS-AFM images captured at 10 fps. (D–F), Height histograms for MBD and CTD, and Autocorrelation functions of MBD height variations over time. (G), Schematic showing MBD structure stabilization by interaction between MBD and residues 311–328.

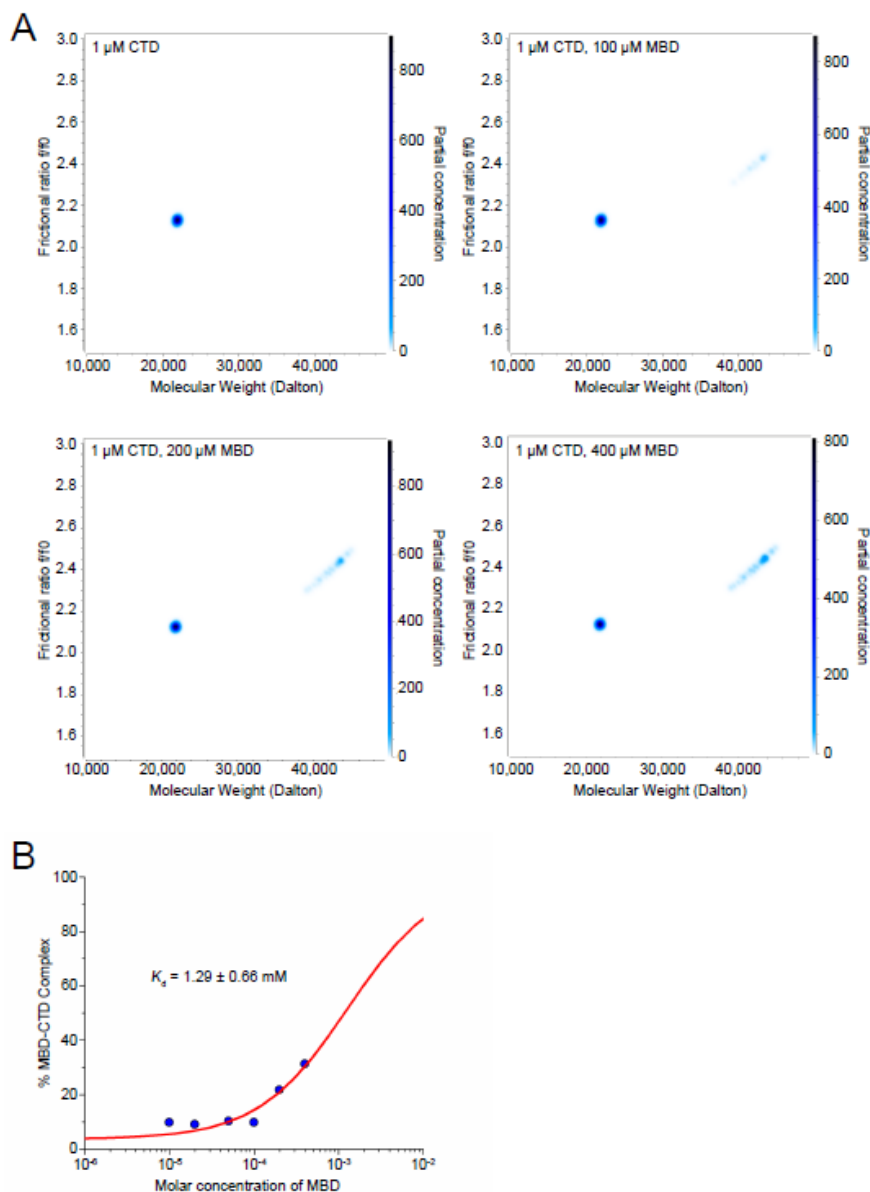


Figure S5.5 Analysis of ultracentrifugation data. (A) Two-dimensional representation of the global genetic algorithm Monte Carlo analysis. Solutes are represented by colored spots, whose color intensity is proportional to their concentration. The position of the spots indicates their approximate buoyant molar masses and their anisotropies. An isotropic particle has an anisotropy of one, higher values indicate increasingly non-globular structure. An anisotropy of 2 or higher indicates significantly elongated shape in solution. Molar masses are approximate since partial specific volumes used to transform sedimentation and diffusion coefficients for each

species are only estimated. Additional uncertainty is present for the low concentration complex because only a small amount of signal is available. (B) Estimation of K_d value for the CTD–MBD complex. A limited number of data points were available for the fit of the binding isotherm, producing a relatively large error in the estimate.

Table S5.1 Decay rates of auto-correlation functions calculated from time-series of MBD height data and rate constants for structural transition dynamics of MBD.

Constructs	Decay rates		Area ratios	Rate constants			
	λ_U (s^{-1})	λ_S (s^{-1})	A_L/A_H	$k_{U \rightarrow S}$ (s^{-1})	$k_{S \rightarrow U}$ (s^{-1})	$k_{L \rightarrow H}$ (s^{-1})	$k_{H \rightarrow L}$ (s^{-1})
WT	3.77 ± 0.60	0.077 ± 0.001	0.25 ± 0.03	0.059 ± 0.002	0.018 ± 0.002	0.45 ± 0.17	3.32 ± 1.02
WT-GFP	3.08 ± 0.38	–	10.24 ± 4.95	–	–	0.27 ± 0.13	2.81 ± 1.12
R294X	3.50 ± 0.47	–	7.43 ± 2.98	–	–	0.42 ± 0.16	3.09 ± 0.94
d311-328	3.12 ± 0.84	0.39 ± 0.06	4.18 ± 0.98	0.033 ± 0.024	0.357 ± 0.077	0.60 ± 0.20	2.52 ± 0.38
d370-415	4.50 ± 1.09	0.19 ± 0.02	0.06 ± 0.01	0.178 ± 0.019	0.012 ± 0.002	4.25 ± 1.03	0.25 ± 0.06
F155S	0.62 ± 0.06	–	0.89 ± 0.67	–	–	0.33 ± 0.12	0.29 ± 0.02

All values are of mean \pm s.e.

Other supplementary materials for this manuscript include the followings:

Movie S5.1. Four movies from molecular dynamics simulations of MBD and half NTD domain (top) and MBD and the full NTD domain (bottom). Systems in the presence of a surface are shown on the left-hand side and systems in solution are on the right. The simulations were performed in explicit water but water has been removed from the movies for clarity. Importantly, the structures in the systems remain the same both in the presence and absence of the surface. The movies are from the end of independent 1 μ s simulations.

5.7 Acknowledgments

Province of Ontario Trillium Scholarship Program to CCG; the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Canada Research Chairs Program to MK; International Rett Syndrome Foundation and American Heart Association to AAK; the NIH [grant number GM 066834 to JCH, GM120600 to BD]; the NSF [grant number MCB-1814012 to JCH, ACI-1339649 to BD]; NSF/XSEDE [grant number TG-MCB070039N to BD]; the Japan Society for the Promotion of Science (JSPS) [KAKENHI grant number 21113002, 26119003, 17H06121 to TA]; the Japan Science and Technology Agency (JST) [CREST program, grant number JPMJCR13M1 to TA; JPMJCR1762 to NK]; University of Texas [grant number TG457201 to BD] and the San Antonio Cancer Institute [grant number P30 to BD].

Computing facilities for MK have been provided by SHARCNET and Compute Canada. Analytical ultracentrifugation-related supercomputer calculations were performed on Comet at the San Diego Supercomputing Center and on Lonestar-5 at the Texas Advanced Computing Center.

5.8 References

1. Lewis, J.D., Meehan, R.R., Henzel, W.J., Maurer-Fogy, I., Jeppesen, P., Klein, F. and Bird, A. (1992) Purification, sequence, and cellular localization of a novel chromosomal protein that binds to methylated DNA. *Cell*, **69**, 905–914.
2. Fraga, M.F., Ballestar, E., Montoya, G., Taysavang, P., Wade, P.A. and Esteller, M. (2003) The affinity of different MBD proteins for a specific methylated locus depends on their intrinsic binding properties. *Nucleic Acids Res.*, **31**, 1765–1774.
3. Nikitina, T., Shi, X., Ghosh, R.P., Horowitz-Scherer, R.A., Hansen, J.C. and Woodcock, C.L. (2007) Multiple modes of interaction between the methylated DNA binding protein MeCP2 and chromatin. *Mol. Cell. Biol.*, **27**, 864–877.

4. Sanfeliu,A., Hokamp,K., Gill,M. and Tropea,D. (2019) Transcriptomic analysis of Mecp2 mutant mice reveals differentially expressed genes and altered mechanisms in both blood and brain. *Front. Psychiatry*, **10**, article number 278.
5. Tillotson,R. and Bird,A. (2019) The molecular basis of MeCP2 function in the brain. *J. Mol. Biol.*, **432**, 1602–1623.
6. Connolly,D.R. and Zhou,Z. (2019) Genomic insights into MeCP2 function: A role for the maintenance of chromatin architecture. *Curr. Opin. Neurobiol.*, **59**, 174–179.
7. Adams,V.H., McBryant,S.J., Wade,P.A., Woodcock,C.L. and Hansen,J.C. (2007) Intrinsic disorder and autonomous domain function in the multifunctional nuclear protein, MeCP2. *J. Biol. Chem.*, **282**, 15057–15064.
8. Guy,J., Cheval,H., Selfridge,J. and Bird,A. (2011) The Role of MeCP2 in the Brain. *Annu. Rev. Cell Dev. Biol.*, **27**, 631–652.
9. Hite,K.C., Kalashnikova,A.A. and Hansen,J.C. (2012) Coil-to-helix transitions in intrinsically disordered methyl CpG binding protein 2 and its isolated domains. *Protein Sci.*, **21**, 531–538.
10. Hansen,J.C., Wexler,B.B., Rogers,D.J., Hite,K.C., Panchenko,T., Ajith,S. and Black,B.E. (2011) DNA binding restricts the intrinsic conformational flexibility of methyl CpG binding protein 2 (MeCP2). *J. Biol. Chem.*, **286**, 18938–18948.
11. Wakefield,R.I., Smith,B.O., Nan,X., Free,A., Soteriou,A., Uhrin,D., Bird,A.P. and Barlow,P.N. (1999) The solution structure of the domain from MeCP2 that binds to methylated DNA. *J. Mol. Biol.*, **291**, 1055–1065.
12. Ho,K.L., Mcnae,I., Schmiedeberg,L., Klose,R.J., Bird,A.P. and Walkinsha,M. (2008) MeCP2 binding to DNA depends upon hydration at methyl-CpG. *Mol. Cell*, **29**, 525–531.

13. Krishnaraj,R., Ho,G. and Christodoulou,J. (2017) RettBASE: Rett syndrome database update. *Hum. Mutat.*, **38**, 922–931.
14. Ando,T., Uchihashi,T. and Scheuring,S. (2014) Filming biomolecular processes by high-speed atomic force microscopy. *Chem. Rev.*, **114**, 3120–3188.
15. Ando,T. (2019) High-speed atomic force microscopy. *Curr Opin Chem Biol.*, **51**, 105–112.
16. Ando,T., Uchihashi,T. and Fukuma,T. (2008) High-speed atomic force microscopy for nano-visualization of dynamic biomolecular processes. *Prog. Surf. Sci.*, **83**, 337–437.
17. Kodera,N., Yamamoto,D., Ishikawa,R. and Ando,T. (2010) Video imaging of walking myosin V by high-speed atomic force microscopy. *Nature*, **468**, 72–76.
18. Uchihashi,T., Iino,R., Ando,T. and Noji,H. (2011) High-speed atomic force microscopy reveals rotary catalysis of rotorless F1-ATPase. *Science*, **333**, 755–758.
19. Uchihashi,T., Watanabe,Y., Nakazaki,Y., Yamasaki,T., Watanabe,H., Maruno,T., Ishii,K., Uchiyama,S., Song,C., Murata,K., Iino,R. and Ando,T. (2018) Dynamic structural states of ClpB involved in its disaggregation function. *Nat. Commun.*, **9**, article number 2147.
20. Yamamoto,H., Fujioka,Y., Suzuki,S.W., Noshiro,D., Suzuki,H., Kondo-Kakuta,C., Kimura,Y., Hirano,H., Ando,T., Noda,N.N. and Ohsumi,Y. (2016)The intrinsically disordered protein Atg13 mediates supramolecular assembly of autophagy initiation complexes. *Dev. Cell*, **38**, 86–99.
21. Watanabe-Nakayama,T., Ono,K., Itami,M., Takahashi,R., Teplow,D.B. and Yamada,M. (2016) High-speed atomic force microscopy reveals structural dynamics of amyloid β 1-42 aggregates. *Proc. Natl. Acad. Sci. U. S. A.*, **113**, 5835–5840.

22. Zhang, Y., Hashemi, M., Lv, Z., Williams, B., Popov, K.I., Dokholyan, N.V. and Lyubchenko, Y.L. (2018) High-speed atomic force microscopy reveals structural dynamics of α -synuclein monomers and dimers. *J. Chem. Phys.*, **148**, article number 123322.
23. Uchihashi, T., Kodera, N. and Ando, T. (2012) Guide to video recording of structure dynamics and dynamic processes of proteins by high-speed atomic force microscopy. *Nat. Protoc.*, **7**, 1193–1206.
24. Ngo, K.X., Kodera, N., Katayama, E., Ando, T. and Uyeda, T.Q.P. (2015) Cofilin-induced unidirectional cooperative conformational changes in actin filaments revealed by high-speed atomic force microscopy. *Elife*, **4**, article number e04806.
25. Brünger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N.S., Read, R.J., Rice, L.M., Simonson, T. and Warren, G.L. (1998) A New Software Suite for Macromolecular Structure Determination. *Acta Crystallogr. D. Biol. Crystallogr.*, **54**, 905–921.
26. Abraham, M.J., Murtola, T., Schulz, R., Páll, S., Smith, J.C., Hess, B. and Lindahl, E. (2015) GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, **1–2**, 19–25.
27. Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W. and Klein, M.L. (1983) Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, **79**, 926–935.
28. Aliev, A.E., Kulke, M., Khaneja, H.S., Chudasama, V., Sheppard, T.D. and Lanigan, R.M. (2014) Motional timescale predictions by molecular dynamics simulations: Case study using proline and hydroxyproline sidechain dynamics. *Proteins Struct. Funct. Bioinforma.*, **82**, 195–215.
29. Darden, T., York, D. and Pedersen, L. (1993) Particle mesh Ewald: An $N \log(N)$ method for Ewald sums in large systems. *J. Chem. Phys.*, **98**, 10089–10092.

30. Bussi,G., Donadio,D. and Parrinello,M. (2007) Canonical sampling through velocity rescaling. *J. Chem. Phys.*, 126, article number 014101.
31. Parrinello,M. and Rahman,M. (2007) Polymorphic transitions in single crystals : A new molecular dynamics method. *J. Appl. Phys.*, 52, 7182–7190.
32. Cao,W. and Demeler,B. (2008) Modeling analytical ultracentrifugation experiments with an adaptive space-time finite element solution for multicomponent reacting systems. *Biophys. J.*, 95, 54–65.
33. Brookes,E.H. and Demeler,B. (2007) Parsimonious regularization using genetic algorithms applied to the analysis of analytical ultracentrifugation experiments. 9th Annu. Conf. Genet. Evol. Comput. GECCO '07, 361–368.
34. Brookes,E., Cao,W. and Demeler, B. (2010) A two-dimensional spectrum analysis for sedimentation velocity experiments of mixtures with heterogeneity in molecular weight and shape. *Eur. Biophys. J.*, 95, 405–414.
35. Demeler,B. and Brookes,E. (2008) Monte Carlo analysis of sedimentation experiments. *Colloid Polym. Sci.*, 286, 129–137.
36. Brookes,E. and Demeler,B. (2008) Parallel computational techniques for the analysis of sedimentation velocity experiments in UltraScan. *Colloid Polym. Sci.*, 286, 139–148.
37. MacGregor,I.K., Anderson,A.L. and Laue,T.M. (2004) Fluorescence detection for the XLI analytical ultracentrifuge. *Biophys. Chem.*, 108, 165–185.
38. Demeler,B. and Gorbet, G.E. (2016) Analytical ultracentrifugation data analysis with UltraScan-III. In Uchiyama,S., Arisaka,F., Stafford,W.F. and Laue,T. (ed.), *Analytical ultracentrifugation: Instrumentation, software, and applications*. Springer, Japan, pp. 119-143.
39. Laue,T., Shah,B., Ridgeway,T. and Pelletier,S. (1992) Computer-aided interpretation of analytical sedimentation data for proteins. In *Analytical*

Ultracentrifugation in Biochemistry and Polymer Science. Harding, S.E., Rowe, A.J. and Horton, J.C. (eds). Cambridge: Royal Society of Chemistry, pp. 90–125.

40. Demeler, B. (2010) Methods for the design and analysis of sedimentation velocity and sedimentation equilibrium experiments with proteins. *Curr. Protoc. Protein Sci.*, **60**, 1–24.
41. Schuck, P. and Demeler, B. (1999) Direct sedimentation analysis of interference optical data in analytical ultracentrifugation. *Biophys. J.*, **76**, 2288–2296.
42. Demeler, B. and Van Holde, K.E. (2004) Sedimentation velocity analysis of highly heterogeneous systems. *Anal. Biochem.*, **335**, 279–288.
43. Ghosh, R.P., Horowitz-Scherer, R.A., Nikitina, T., Gierasch, L.M. and Woodcock, C.L. (2008) Rett syndrome-causing mutations in human MeCP2 result in diverse structural changes that impact folding and DNA interactions. *J. Biol. Chem.*, **283**, 20523–20534.
44. Demeler, B., Saber, H. and Hansen, J.C. (1997) Identification and interpretation of complexity in sedimentation velocity boundaries. *Biophys. J.*, **72**, 397–407.
45. Ballestar, E. and Yusufzai, T.M. (2000) Wolffe, A.P. Effects of rett syndrome mutations of the Methyl-CpG binding domain of the transcriptional repressor MeCP2 on selectivity for association with methylated DNA. *Biochemistry*, **39**, 7100–7106.
46. Kumar, A., Kamboj, S., Malone, B.M., Kudo, S., Twiss, J.L., Czymmek, K.J., LaSalle, J.M. and Schanen, N.C. (2008) Analysis of protein domains and Rett syndrome mutations indicate that multiple regions influence chromatin-binding dynamics of the chromatin-associated protein MECP2 in vivo. *J. Cell Sci.* **121**, 1128–1137.
47. Georgel, P.T., Horowitz-Scherer, R.A., Adkins, N., Woodcock, C.L., Wade, P.A. and Hansen, J.C. (2003) Chromatin compaction by human MeCP2. Assembly of novel

secondary chromatin structures in the absence of DNA methylation. *J. Biol. Chem.*, **278**, 32181–32188.

48. Ghosh,R.P., Horowitz-Scherer,R.A., Nikitina,T., Shlyakhtenko,L.S. and Woodcock,C.L. (2010) MeCP2 binds cooperatively to its substrate and competes with histone H1 for chromatin binding sites. *Mol. Cell. Biol.*, **30**, 4656–70.
49. Della Ragione,F., Vacca,M., Fioriniello,S., Pepe,G. and D’Esposito,M. (2016) MECP2, a multi-talented modulator of chromatin architecture. *Brief. Funct. Genomics*, **15**, 420–431.
50. Lyst,M.J., Ekiert,R., Ebert,D.H., Merusi,C., Nowak,J., Selfridge,J., Guy,J., Kastan,N.R., Robinson,N.D., Alves,Fde L., Rappsilber,J., Greenberg,M.E. and Bird,A. (2013) Rett syndrome mutations abolish the interaction of MeCP2 with the NCoR/SMRT co-repressor. *Nat. Neurosci.*, **16**, 898–902.
51. Van der Lee,R., Buljan,M., Lang,B., Weatheritt,R.J., Daughdril,G.W., Dunker,A.K., Fuxreiter,M., Gough,J., Gsponer,J., Jones,D.T., Kim,P.M., Kriwacki,R.W., Oldfield,C.J., Pappu,R.V., Tompa,P., Uversky,V.N., Wright,P.E. and Babu,M.M. (2014) Classification of intrinsically disordered regions and proteins. *Chem. Rev.*, **114**, 6589–6631.

6 A multiscale computational study of the conformation of the full-length intrinsically disordered protein MeCP2

Cecilia Chávez-García^{1,2}, Jérôme Hénin³ and Mikko Karttunen^{*1,2,4}

¹Department of Chemistry, The University of Western Ontario, 1151 Richmond Street, London, Ontario, Canada, N6A 5B7, ²The Center for Advanced Materials and Biomaterials Research, The University of Western Ontario, London, Ontario, Canada, N6A 3K7, ³Laboratoire de Biochimie Théorique UPR 9080, CNRS and Université de Paris, Paris, France, ⁴Department of Physics and Astronomy, The University of Western Ontario, London, Ontario, Canada, N6A 3K7

*Corresponding author: mikko.karttunen@uwo.ca

Link: <https://www.biorxiv.org/content/10.1101/2021.11.08.467619v1>

Submission ID: ci-2021-01354p

6.1 Abstract

The malfunction of the Methyl CpG binding protein 2 (MeCP2) is associated to the Rett syndrome, one of the most common causes of cognitive impairment in females. MeCP2 is an intrinsically disordered protein (IDP), making its experimental characterization a challenge. There is currently no structure available for the full-length MeCP2 in any of the databases, and only the structure of its MBD domain has been solved. We used this structure to build a full-length model of MeCP2 by completing the rest of the protein via *ab initio* modelling. Using a combination of all-atom and coarse-grained simulations, we characterized its structure and dynamics as well as the conformational space sampled by the ID and TRD domains in the absence of the rest of the protein. The present work is the first computational study of the full-length protein. Two main conformations were sampled in the coarse-grained simulations: a globular structure similar to the one observed in the all-atom force field and a two-globule conformation. Our all-atom model is in good agreement with the available experimental data, predicting amino acid W104 to be buried, amino acids R111 and R133 to be solvent accessible, and having 4.1% of α -helix content, compared to the 4% found experimentally. Finally, we compared the model predicted by AlphaFold to our Modeller model. The model was not stable in water and underwent further folding. Together, these simulations provide a detailed (if perhaps incomplete) conformational ensemble of the full-length MeCP2, which is compatible with experimental data and can be the basis of further studies, e.g., on mutants of the protein or its interactions with its biological partners.

6.2 Introduction

Methyl CpG binding protein 2 (MeCP2) is a transcriptional regulator essential for growth and synaptic activity of neurons¹. The malfunction of this protein is associated to the Rett syndrome, one of the most common causes of cognitive impairment in females^{2,3}. The MeCP2 gene is X-linked in mammals. Mutations that affect the protein function were initially thought to be lethal in males⁴, but these are now frequently identified in cognitively impaired male patients⁵.

MeCP2 is an intrinsically disordered protein (IDP), and little is known about its molecular architecture during normal cellular processes and in disease⁶. IDPs are characterized by a low proportion of bulky hydrophobic amino acids and high proportions of charged and hydrophilic amino acids. Consequently, they cannot bury sufficient hydrophobic core to fold spontaneously into stable, highly organized three-dimensional structures; instead, they fluctuate through an ensemble of conformations⁷. The physical characteristics of IDPs makes their structural characterization a challenge as these proteins are more sensitive to degradation.

MeCP2 contains 486 amino acids, is a monomer in solution and is composed of six different domains⁸. Residues 78-162 specifically bind to methylated CpG dinucleotides and have been termed the methyl-CpG binding domain (MBD)⁹. Another functionally annotated region corresponds to the transcriptional repression domain (TRD) whose main function is to repress the transcription of genes¹⁰. Biophysical and protease digestion experiments identified three other domains: the N-terminal domain (NTD), the intervening domain (ID) and the C-terminal domain (CTD), which can be subdivided into CTD- α and CTD- β ⁸ (Fig. 6.1).

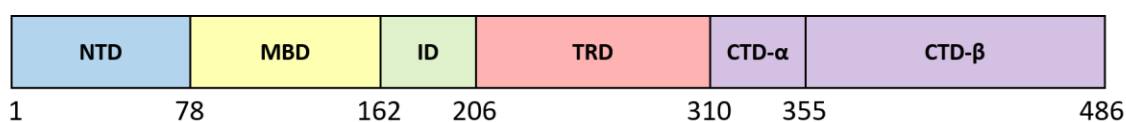


Figure 6.1 MeCP2 is composed of six domains: The N-terminal domain (NTD), the methyl-CpG binding domain (MBD), the intervening domain (ID), the transcription repression domain (TRD) and the C-terminal domain (CTD) which can be subdivided into CTD- α and CTD- β . The only available structure¹ contains solely the MBD domain, which is the only ordered region in the protein.

There is currently no structure available for the full-length MeCP2 in any of the protein databases. MBD is the only domain for which the secondary structure is known, and it only

accounts for ~17% of the amino acids¹; MBD is also the only ordered domain. Circular dichroism (CD) of recombinant human MeCP2 has shown that the protein consists of ~35% β -strand/turn, 5% α -helix and almost 60% is unstructured². Characterization of MeCP2 by hydrogen/deuterium exchange has indicated disorder in the entire polypeptide chain with the exception of the MBD domain³. Further CD studies of isolated NTD, ID, TRD and CTD domains confirmed their lack of stable secondary structure⁴. It has been experimentally demonstrated that the NTD, CTD and TRD domains can undergo a coil to helix transition, with the TRD showing the greatest tendency for helix formation⁴.

To date, two computational studies of MeCP2 have been reported, and both focus on the ordered MBD domain only. Kucukkal and Alexov reported comparative MD simulations of the R133C mutant and wild-type MBD⁵, and Yang et al. studied the effects of Rett syndrome-causing mutations on the binding affinity of MBD to CpG dinucleotides⁶. The scarcity of computational studies is due to the lack of a three-dimensional structure of the full-length protein. Nevertheless, computer simulations have been able to predict the structures of IDPs. For example, using a coarse-grained model, *ab initio* simulations of pKID successfully modeled its coupled folding and binding to KIX⁷, and a combination of homology and *ab initio* modelling provided valuable insight into the three-dimensional structures of intrinsically disordered e7 proteins⁸. In this work, we used the known structure for the MBD domain as a starting point with the rest of the protein built by *ab initio* modeling. Using a combination of all-atom and coarse-grained simulations, the folding of the full-length MeCP2 and the conformational ensemble it could sample were studied.

6.3 Materials and methods

Ab initio modelling

Modeller⁹ version 9.19 was used to build a model for the full-length MeCP2 protein. Using the BLAST algorithm¹⁰, we searched the UniProt database¹¹ for homologues of MeCP2 with a 3D structure. Unfortunately, the only known structures belong to homologues of the MBD domain, which accounts for only ~17% of MeCP2 amino acids and whose structure

has already been determined. Thus, we used the Protein Data Bank 1QK9,¹ which contains the MBD domain structure, as a template. Twenty different models were generated with Modeller⁹. There was little variation between the different models and thus the first model was chosen as the starting structure for the simulations (Fig. S6.1). With the aim of having a different starting structure for our coarse-grained simulations, a second model was built by refining the loops of the first model using the loopmodel class in Modeller. There is no structural information on this protein besides its known disorder and the structure of the MBD domain, and thus no quality assessment predictors were used to evaluate the generated models. The evaluation will come from the data obtained during the simulations.

The AlphaFold¹² model for human MeCP2 (UniProt code: P51608) was downloaded from the database hosted by the European Bioinformatics Institute (<https://alphafold.ebi.ac.uk>). Three simulations were performed with this model as the initial structure, using the procedure described in the next section.

All-atom simulations

The following procedure was used in all of the all-atom MD simulations: The initial structure was placed in a dodecahedral box in which the distance from the edges of the box to every atom in the protein was at least 1 nm. The box was solvated with water and 150 mM of NaCl was added to reproduce physiological conditions. Counterions were added to maintain the overall charge neutrality of the system. Simulations were performed using GROMACS 2016.3¹³ with the TIP3P water model¹⁴ and the Amber99SB*-ILDNP force field¹⁵. The only exception is the set of five replicas for the ID and TRD domains that were run with the CHARMM36IDPSFF force field that is parameterized specifically for intrinsically disordered proteins¹⁶. This IDPs-specific force field has been shown to produce good results when compared to other force fields in a recent study of amyloid- β ¹⁷, an extensively studied IDP. Table S6.1 contains the details of the all-atom simulations: three simulations of Modeller models, three simulations of the AlphaFold model and 12 simulations of sections of the ID and TRD domains of different lengths.

Each system was first energy minimized using the method of steepest descents and pre-equilibrated in the canonical ensemble, i.e., at constant particle number, temperature and volume, for 100 ps. Pre-equilibration was followed by a production run with a time step of 2 fs. The Lennard-Jones potential was truncated using a shift function between 1.0 and 1.2 nm. Electrostatic interactions were calculated using the particle-mesh Ewald method (PME)^{18,19} with a real space cut-off of 1.2 nm. The temperature was set to 310 K with the V-rescale algorithm²⁰ and pressure was kept at 1 atm using the Parrinello-Rahman barostat²¹. Bonds involving hydrogens were constrained using the Parallel Linear Constraint Solver (P-LINCS) algorithm²².

Some systems (marked as “resized” in Table S6.1) were moved into a smaller simulation box after an initial run in which the protein became more compact. The final configuration of the initial simulation was placed into a new simulation box in which the distance from the edges of the box to every atom in the protein was again at least 1 nm. The new box was solvated with water, 150mM of NaCl and counterions. Each new system was energy minimized and pre-equilibrated in the canonical ensemble before moving to the production run. All parameters mentioned above were kept the same. Trajectory analysis was performed using Gromacs built-in tools¹³ and MDAnalysis^{23,24}.

Coarse-grained simulations

The intermediate-resolution implicit solvent coarse-grained protein model PLUM²⁵ by Bereau and Deserno was used to further explore the conformational landscape of MeCP2. This model represents the backbone with near-atomistic resolution, with beads for the amide group N, central carbon C α and carbonyl group C'. The side chains are represented by single beads located at the first carbon C β of the all-atom model. The N and C' beads can hydrogen bond through a directional potential which depends on the implicit positions of hydrogen and oxygen atoms within them. The PLUM model has been successfully used to study a variety of scenarios such as the aggregation of polyglutamine²⁶, β -barrel formation at the interface between virus capsid proteins²⁷, folding of transmembrane

peptides²⁸, and it has been shown to be able to reproduce the secondary structure of small IDPs involved in biomineralization²⁹.

Simulations using this model were carried out in GROMACS 4.5.5¹³ specifically modified to support the PLUM model. All interaction parameters were taken from the original work of Berau and Deserno²⁵. The simulations were run in the canonical ensemble (NVT) with a Langevin thermostat with friction constant $\Gamma = \tau^{-1}$ and an integration timestep of $\delta t = 0.01\tau$, where τ is the natural time unit in the simulation. The reduction in degrees of freedom removes friction and speeds up the motion through phase space and thus this time unit is not equivalent to the time step in an all-atom simulation²⁵. Table S2 contains the simulation details.

6.4 Results

The all-atom protein is largely unstructured

A full-length MeCP2 all-atom protein model, henceforth referred to as MeCP2_1, was simulated for 1,550 ns. The protein started with an extended conformation (Fig. 6.2) in order to minimize bias towards any particular fold. After an initial simulation of 150 ns, the protein had become more compact and it was moved to a smaller box to increase efficiency. Figure 6.2 shows a drastic decrease in the radius of gyration (R_g) during the first 20 ns of the simulation, when it went from 8.35 nm to 4.56 nm. Although R_g continued to fluctuate, it never surpassed 5 nm. Moving the protein to a smaller box allowed a reduction in the number of water molecules from ~793,000 to ~120,000 (Table S6.1).

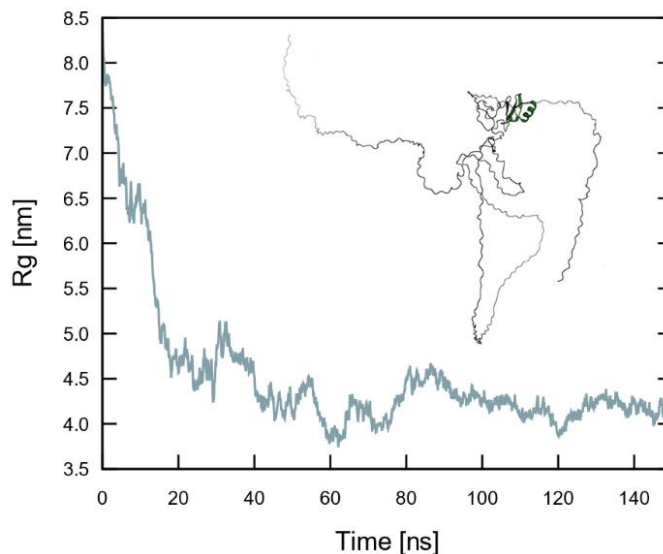


Figure 6.2 Radius of gyration (R_g) of the full-length MeCP2 in the initial simulation box. The protein becomes more compact during the first 20 ns. The inset shows the initial structure. MBD, the only ordered domain, is clearly visible.

The protein was then run in the new box for an additional 1,400 ns. The protein remained highly flexible; its root-mean-square deviation (RMSD) from the initial structure continued to show small fluctuations throughout the trajectory, as expected for an IDP (Fig. 6.3A). The most populated cluster in the last 400 ns of the simulation is largely unstructured with only small motifs of secondary structure (Fig. 6.3B). Shown in Fig. 6.3B red are the residues with a root mean square fluctuation (RMSF) larger than 0.6 nm. Figure 6.3C shows the RMSF of each amino acid throughout the last 400 ns of the trajectory. The residues with the highest RMSF are located in the NTD and CTD- β domains, at the opposite ends of the protein, and in two solvent-exposed loops.

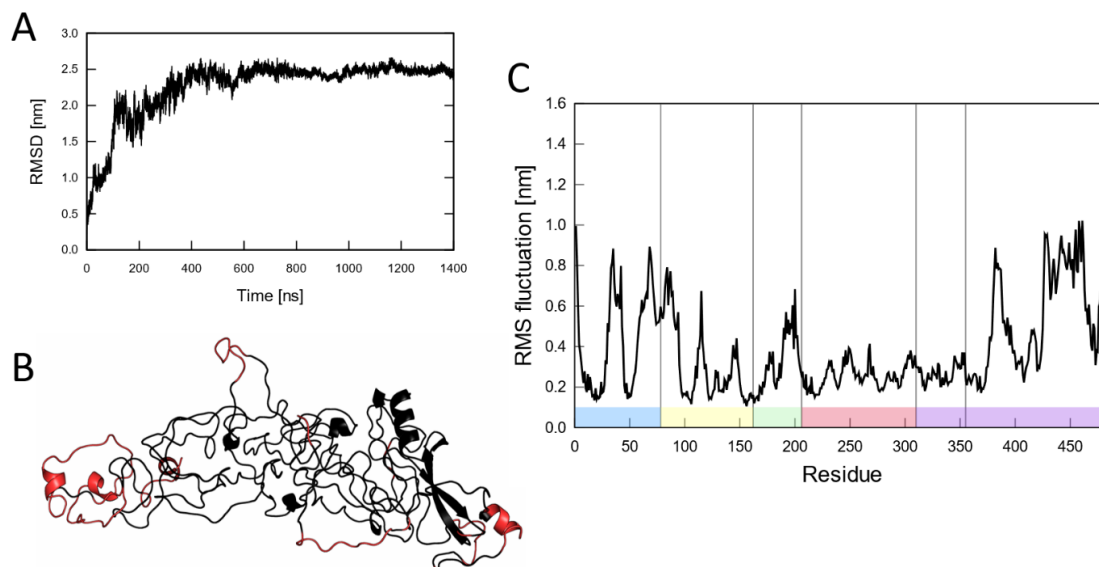


Figure 6.3 All-atom MD simulation of the full-length MeCP2. A) RMSD of the protein in the smaller simulation box. B) Most sampled cluster throughout the last 400 ns of the simulation. Red: residues with an RMSF higher than 0.6 nm. C) RMSF of the protein throughout the last 400 ns of the simulation. The different domains are marked following the color code in Fig 1.

The last 400 ns of the 1,400 ns trajectory were clustered using the method of Daura *et al.*³⁰ with a 0.5 nm cut-off. The secondary structure was computed for the most representative structure in each of the 11 clusters obtained (Table S6.3). The weighted averages show that 20 residues had an α -helix conformation (4.1%), 113 residues were in β -strands or turns (23.2%) and 338 residues were in random coil (69.7%). This is very similar to experimental data of Adams *et al.*, in particular the amount of α -helix compared to the experimentally (by CD) determined 4%³¹.

We also computed the secondary structure of the protein throughout the last 400 ns of the simulation using DSSP^{32,33}. The secondary structure elements in the MBD domain are very stable, appearing in at least 80% of all frames (Fig. 6.4). Adams *et al.*³¹ reported the secondary structure for the MBD domain on its own to be 10% α -helix, 51% β -strands or turns and 38% unstructured, and the NMR structure (PDBid: 1QK9¹) contains 12% α -helix,

20% β -strands or turns and is 69% unstructured. The MBD domain in our simulation had 15% α -helix, 28% β -strands or turns, and is 57% unstructured. Overall, our simulation is in good agreement with the experimental data. The most disordered domains are the ID domain (Fig. S6.4) and the CTD α domain (Fig. S6.6). The NTD domain has two short β -strands and two helices, with one of them present in 60% of the simulation frames (Fig. S6.3). The TRD domain formed a helix in residues 241 to 244 in 70% of the simulation and two β -strands were observed in 6% of the frames (Fig. S6.5). The residues in the α -helix correspond to 4% of the TRD residues and the unstructured residues to 87% of the TRD amino acids. This is in good agreement with the 3% of α -helix and 85% of unstructured residues measured by Adams *et al*³¹. Five helices are observed in the CTD β domain, with two of them present 80% of the simulation (Fig. S6.7).

Even though some of the secondary structure elements appeared in only a small fraction of the frames, these could become stable upon interaction with another protein, DNA or a small molecule. In fact, most domains in MeCP2 can bind to DNA; the MBD domain binds to symmetrically methylated 5'CpG3' pairs with a preference for A/T-rich motifs^{34,35}, an autonomous DNA binding domain has been identified in the ID domain³⁶, the TRD domain possesses a non-specific DNA binding site^{31,36} and there is a distinct non-specific binding site for unmethylated DNA in the CTD α domain³⁶.

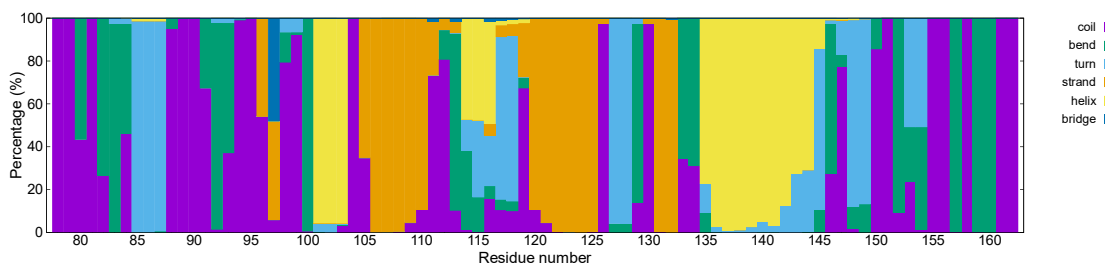


Figure 6.4 Percentage of frames in the last 400 ns of the MBD domain with every type of secondary structure. The secondary structure of the MBD domain observed in the MeCP2_1 simulation, 15% α -helix, 28% β -strands or turns, and 57% unstructured, is in good agreement with experimental data³¹.

Principal component analysis (PCA) of the trajectory underlines the structural rearrangements that the protein undergoes during the first 600 ns of the simulation (Fig.

S6.8A). After this time, the protein explores a much smaller portion of the conformational space. In contrast, the TRD domain begins to sample more conformational space in the second half of the simulation (Fig. S6.8B).

The experiments by Ghosh *et al.*³⁷ showed that the single tryptophan of MeCP2, which is located at position 104 in the MBD domain, is strongly protected from the aqueous environment. Using the STRIDE web server³⁸, we computed the relative solvent accessible surface area (rSA) of residue W104 in the four most populated clusters (Table S6.4). The first four clusters contain 98% of all frames in the last 400 ns simulation. Although there is no consensus on where to set the threshold to determine if an amino acid is buried, it is typically set between 10% and 20%^{39,40}. The weighted average for the four clusters gave a rSA of 8.1% and thus it can be considered to be buried inside the protein, in agreement with the experimental data.³⁷

R133C is one of the most common disease-causing mutations in the MBD domain³⁷. The x-ray structure of an MBD-DNA complex has revealed that Arg 133 is involved in the DNA interaction surface⁴¹, and the study by Lei *et al.*⁴² found that this residue, together with Arg 111, forms hydrogen bonds with DNA. In order to see if these two residues are solvent accessible in our simulation, we computed their rSA (Tables S6.5 and S6.6). Residue R111 had a rSA of 12.7% in the most populated cluster, which can be considered to be buried. However, this amino acid had a high rSA value in the second most populated cluster, giving a weighted average of 20.4%. Therefore, this residue is actually solvent accessible. Residue R133 had a weighted average rSA of 53.7% and thus is also solvent accessible. Kucukkal and Alexov⁵ reported an average number of hydrogen bonds with water of 1.68 for residue R133 and 0.47 for residue R111 in their MBD-only simulations. We obtained an average of 2.96 for R133 and 1.59 for R111 in the last 800 ns of the simulation. It is thus evident, that these residues are more solvent accessible when the full-length protein is considered. Kucukkal and Alexov⁵ did not report the total number of salt bridges observed in their simulations, however, they reported the loss of two salt bridges (R133-E137 and K119-D121) upon mutation of residue R133. We computed all salt bridges in the same manner as them, using the Salt Bridges plugin for VMD⁴³. A total of 499 salt bridges were identified but most of them appeared in only a small fraction of the

frames and only 35 were stable during the last 400 ns of the simulation (Table S6.7). Most of these salt bridges occur between the NTD and the MBD domains. Salt bridge K119-D121 is only present in very few frames (Fig. S6.9A). Lys119 formed hydrogen bonds with neighbouring residues 115-117 and Asp121 with Lys109 and Arg111. The salt bridge R133-E137 can be observed at the beginning of the simulation but is lost in the last 400 ns of the simulation. This is consistent with the study by Kucukkal and Alexov⁵ who observed this salt bridge in their 220 ns simulation (Fig. S6.9B), but it underlines the need for sufficiently long sampling times.

Coarse-grained simulations sample two different conformations

In order to investigate other possible folds of the protein, we ran four coarse-grained simulations using the PLUM model²⁵. Three different configurations were used as starting points: A) the structure of the all-atom simulation after 800 ns, B) the initial structure built with Modeller, and C) model “B” with its loops refined (Fig. 6.5).

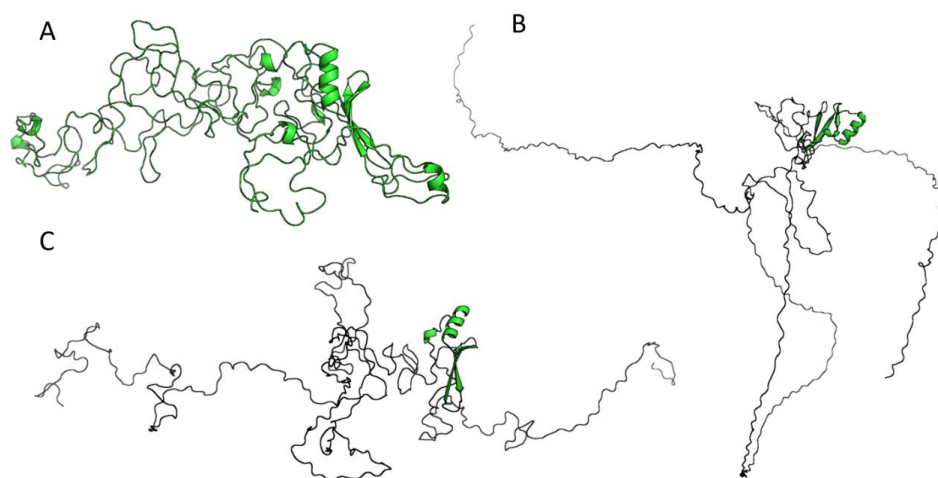


Figure 6.5 Coarse-grained simulations of MeCP2 using the PLUM model²⁵. Simulations started from three different conformations: structure of the all-atom simulation after 800 ns (A), the initial structure built with Modeller⁹ (B) and structure “B” with refined loops (C).

The configuration at 800 ns in the MeCP2_1 simulation (Fig. 6.5A) was used as the starting point for a coarse-grained simulation, henceforth referred to as CG1. Similar to the RMSD in the all-atom simulation, the RMSD of the protein converges to 3.5 nm but continues to fluctuate. The large RMSD value indicates that the overall topology of the structure changed. A cluster analysis was used help to identify the differences.

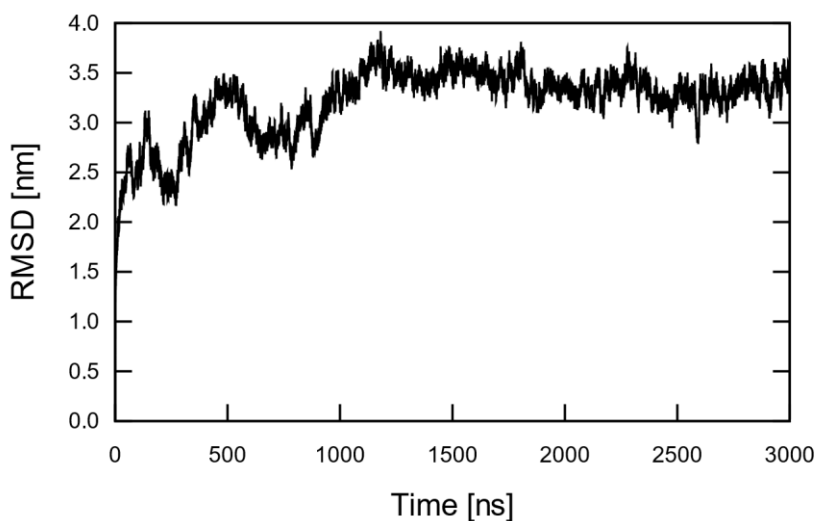


Figure 6.6 RMSD from the initial structure of the protein (the MeCP2_1 model after 800 ns of simulation) in the CG1 PLUM simulation.

We clustered the conformations sampled in the entire trajectory using the method of Daura *et al.*³⁰ with a 2.0 nm cut-off. Two main conformations were revealed: 1) a single globule and 2) two globules connected by a loop (Fig. 6.7). The first two clusters had a single globule configuration but the third had two distinct globules connected by a loop. In this structure, the connecting loop starts at residue 228 and ends in residue 242. A similar conformation can be observed in the eight cluster. The minimum distance between the amino acids of the two globules throughout the simulation shows that the two-globule conformation was sampled at the beginning of the simulation, around 700 ns and at 2,600 ns of simulation. The first two times this conformation was sampled, the linker between

the two globules was long enough to stabilize it for ~100 ns. In contrast, the two-globule conformation sampled at 2,600 ns had a shorter linker and it coalesced into a single globule after 20 ns.

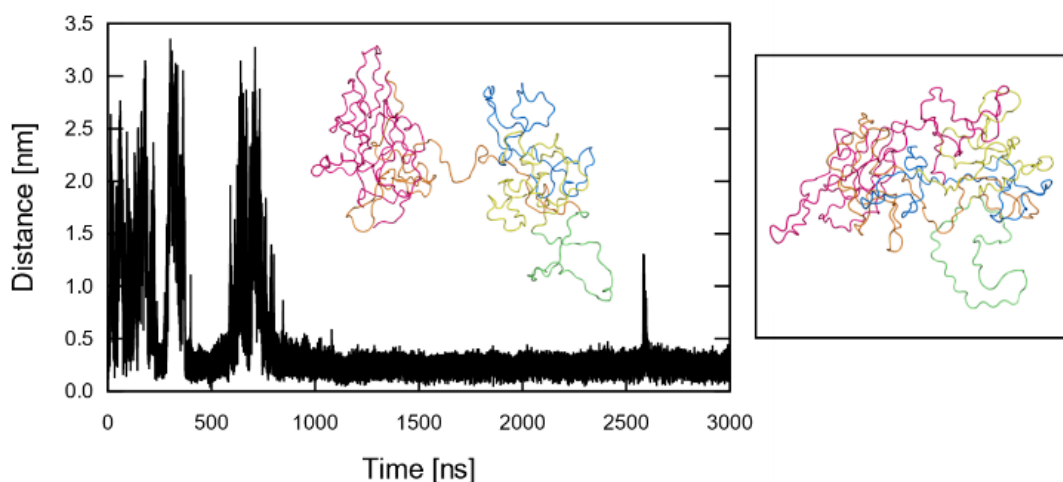


Figure 6.7 Minimum distance between the two globules in the CG1 PLUM simulation. A two-globule conformation is sampled at the beginning of the simulation at 700 ns and once again at 2,600 ns. The single globule and two-globule conformations have their different domains marked following the color code in Fig. 1.

Two different replicas (simulations CG2 and CG3) were run for the model built with Modeller (Fig. 6.5B). Their RMSD converged in the first 200 ns but the simulations were extended to 500 ns (Fig. 6.8A). Since the reference structure for the RMSD calculation is the initial frame i.e., the unfolded structure, a high RMSD value is to be expected. Simulation CG2 sampled conformations similar to those observed in the all-atom MeCP2_1 simulation, albeit more compact (Fig. 6.8B). Simulation CG3 collapsed into a globule which appears to be an energetic minimum since the system could not sample any other conformations (Fig. 6.8A). Interestingly, the loop that remained solvent-exposed for this entire simulation, spans residues 168 to 201 and corresponds to the ID domain (Fig. 6.1).

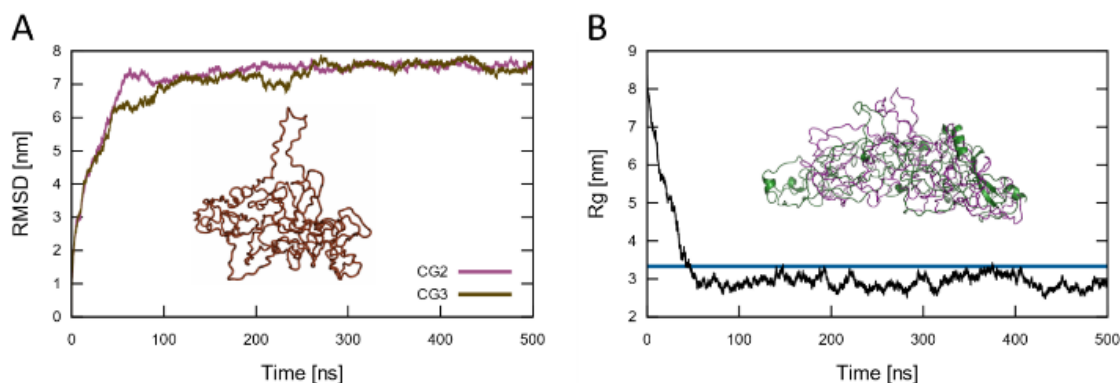


Figure 6.8 (A) RMSD from the initial structure (Modeller model) of the CG2 and CG3 PLUM simulations. Brown: Most populated cluster in the entire CG3 PLUM simulation. (B) Radius of gyration of the CG2 PLUM simulation. Blue line: The average R_g of the MeCP2_1 system. Green: The average structure of the most populated cluster in the MeCP2_1 simulation. Magenta: The average structure of the second most populated cluster in the last 400 ns of the CG2 PLUM simulation.

A fourth coarse-grained simulation (CG4) started from the Modeller model with its loops refined (Fig. 6.5C). Since the starting structure had not been energy minimized, a high RMSD value is to be expected. This simulation sampled two-globule conformations similar to those observed in the CG1 simulation albeit with the connecting loop located between residues 161 and 205. Interestingly, the location of this loop matches the ID domain (Fig. 6.1). The protein underwent two main transitions during the simulation. It became more compact during the first 40 ns, it sampled two-globule conformations from 40 to 315 ns, and it sampled a single globule for the rest of the simulation (Fig. 6.9).

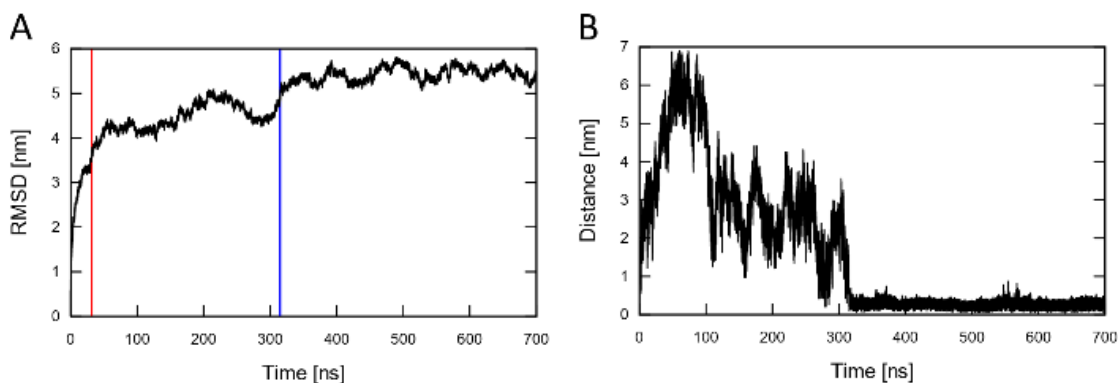


Figure 6.9 (A) RMSD from the initial structure (Modeller model with refined loops) and (B) minimum distance between the two globules of simulation CG4. The protein becomes more compact and at 40 ns (red line) it starts to sample two-globule conformations. After 270 ns (blue line) the two globules merge together and a single globule is sampled.

A cluster analysis with the method of Daura *et al.*³⁰ and a 2.5 nm cutoff of all coarse-grained trajectories concatenated found 23 different clusters (Table S6.8). The first eight clusters contain 95.3% of all structures sampled. Four of these clusters are single globules and four are two-globule conformations. Only the single globule conformations had overlap between trajectories.

To further understand the conformational space sampled by all coarse-grained trajectories, we performed a single Principal Components Analysis (PCA) on all simulations. Even though each simulation sampled different conformations, these get closer to one another over time, when projected onto the first two eigenvectors (Fig. 6.10). This implies that the protein tends toward a similar, limited conformational ensemble in all four simulations.

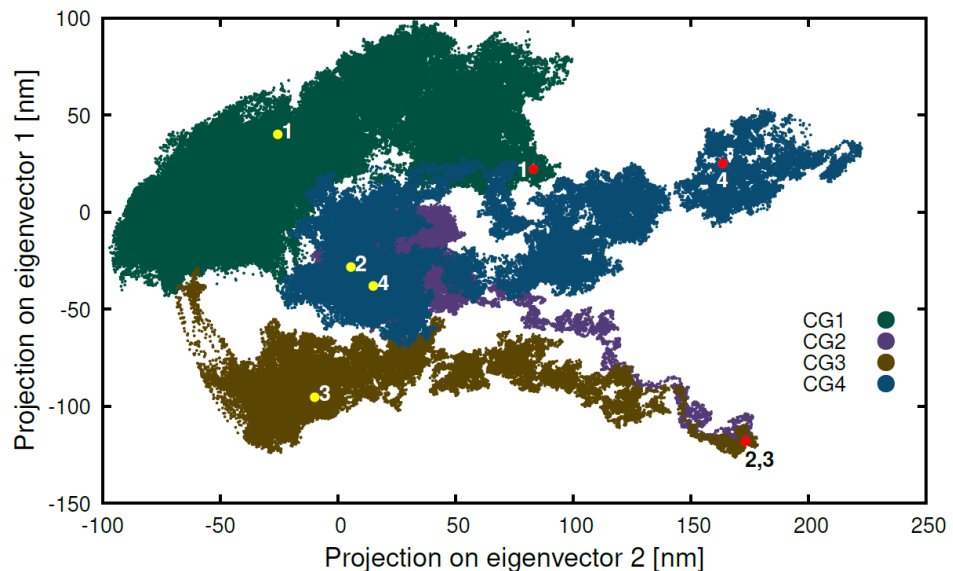


Figure 6.10 Principal Component Analysis. Projection of all coarse-grained trajectories on the first two eigenvectors, each trajectory is depicted with a different colour. Simulations CG2 and CG3 started from the same conformation. The starting points of all simulations are marked in red and the end points in yellow.

All-atom two-globule conformations transition into a single globule

Given that the two-globule conformation had only been sampled by the coarse-grained force field, new all-atom simulations were run starting from this conformation. Modeller⁹ was used to generate the initial structures via homology modeling. Three templates were used to generate the models: One for the first globule (NTD and MBD domains), one for residues in the connecting loop (ID and TRD domains) and one for the second globule (CTD domain).

Model MeCP2_2 was built using the two globules from the most populated cluster with a two-globule conformation in the first 500 ns of simulation CG1. The first template contained residues 1-235, the second template had an extended peptide with residues 230-249, and the third one contained residues 311-486 from the second globule (Table S6.9). The peptide used in the second template was generated using Pymol⁴⁴. Using a longer

peptide for the second template produced single-globule models, with the two globules merged into one and a long loop forming a hoop.

In order to study whether the secondary structure in the MBD domain would have any impact on the stability of the two-globule conformation, we generated another model using an all-atom configuration as a template for the first globule. We used the final structure after 400 ns of simulation as the first template for model MeCP2_3. The loop in Model MeCP2_2 was used as the second template and the second globule (residues 311-486) from simulation CG1 as the third template (Table S6.9).

Model MeCP2_2 collapsed into a single globule after only 20 ns of simulation and did not sample any other conformations, therefore, we did not continue the production run beyond 60 ns. Model MeCP2_3 did not dwell into the same local minimum and its production run was extended to 600 ns. It sampled the two-globule conformation for a longer time but eventually the two globules melted into one, albeit with a more extended structure than the previous model and retaining the secondary structure (Fig. 6.11). It is possible that this conformation was not stable enough because the connecting loop was not in a water-soluble conformation. Simulations of the connecting loop could help us shed some light on this matter.

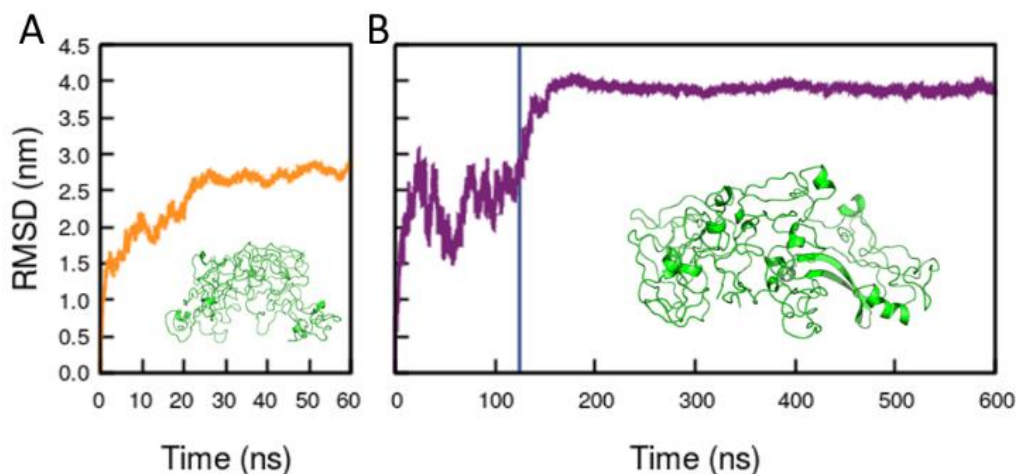


Figure 6.11 RMSD to the initial structure of models (A) MeCP2_2 and (B) MeCP2_3. Green: The most populated cluster in each trajectory. Model MeCP2_2 samples two-

globule conformations during the first 125 ns (blue line), it then undergoes a transition to a single globule.

Comparing all simulations

Using the PLUMED plugin⁴⁵ for GROMACS¹³, we computed the α -helical content of the all-atom simulations, as well as acylndricity and asphericity of all simulations.

The α -helical content was computed by generating a set of all possible six residue sections in the protein and calculating the RMSD distance between each residue configuration and an idealized α -helical structure. This is done by calculating the following sum of functions of the RMSD distances,

$$s = \sum_i \frac{1 - \left(\frac{r_i - d_0}{r_0}\right)^n}{1 - \left(\frac{r_i - d_0}{r_0}\right)^m},$$

where the sum runs over all possible segments of an α -helix. This collective variable was first defined by Pietrucci and Laio⁴⁶ and all parameters were set equal to those used in their original paper: $d_0 = 0.0$, $r_0 = 0.08$ nm, $n = 8$ and $m = 12$.

Model MeCP2_3 sampled conformations with a wider array of values for both the α -helical content and the R_g than the MeCP2_1 simulation (Fig. 6.12). The trajectory analyzed for this all-atom simulation does not include the 150 ns from the bigger simulation box in which the protein underwent initial folding.

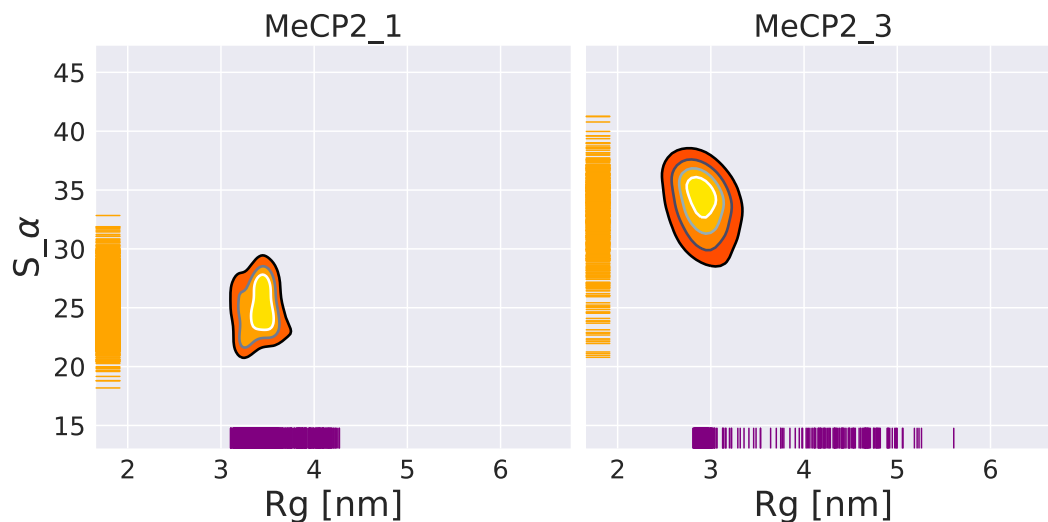


Figure 6.12 α -helical content of the protein structure vs radius of gyration in the all-atom and two-globule all-atom simulations. Comparison between the all-atom simulation (MeCP2_1, left) that started from an extended structure and the one that started from a two-globule conformation (MeCP2_3, right). Orange: Individual measurements of α -helical content. Purple: Individual measurements of R_g .

In 1971, Šolc showed that the shape of polymers can be quantified using the eigenvalues (L_1 , L_2 and L_3) of the tensor of gyration⁴⁷. The symmetry of a polymer, or in this case, of a peptide, can be described by asphericity,

$$b = L_1 - \frac{1}{2}(L_2 + L_3)$$

and acylndricity,

$$c = L_2 - L_1.$$

Figure 6.13 shows the results. All simulations sampled similar values; however, the coarse-grained simulations sampled a wider array of values. From the four coarse-grained simulations, simulation CG1 is the most akin to the all-atom simulations.

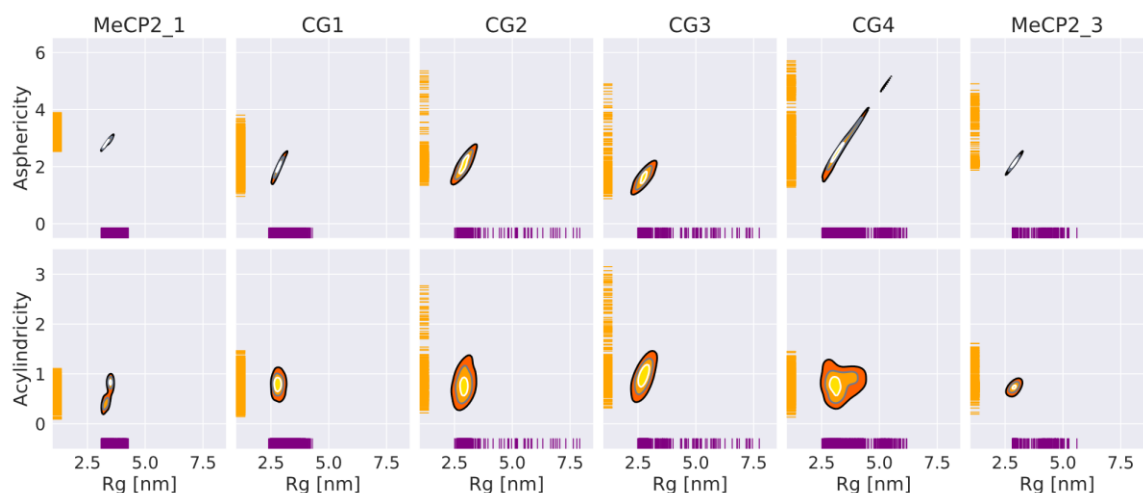


Figure 6.13 Acylindricity and asphericity vs radius of gyration in the all-atom, coarse-grained and two-globule all-atom simulations. The coarse-grained simulations sampled structures with lower asphericity, higher acylindricity and higher radius of gyration than the all-atom simulations. Orange: Individual measurements of asphericity and acylindricity. Purple: individual measurements of radius of gyration.

The ID and TRD domains are highly flexible

In order to thoroughly explore the conformations that the flexible ID and TRD domains that form the connective loop can sample, all-atom simulations were run on the ID and TRD domains (residues 164-310). Five replicas were run with two different force fields: Amber99SB*-ILDNP¹⁵ (simulations A1-A5 in Table S6.1) and CHARMM36IDPSFF¹⁶ (simulations C1-C5 in Table S6.1), using the loop in model MeCP2_3 as the initial structure.

Figure 6.14 shows R_g and end-to-end distance of all structures sampled by the ten simulations. One of these simulations sampled very compact structures but the other nine sampled an array of structures with end-to-end distances between from 3 nm to 23 nm, and radius of gyration from 2.5 nm to 6.5 nm. Table S6.10 shows the most sampled conformations in all ten simulations. Overall, the Amber force field sampled more compact structures than the Charmm force field, in agreement with previous studies⁴⁸⁻⁵⁰.

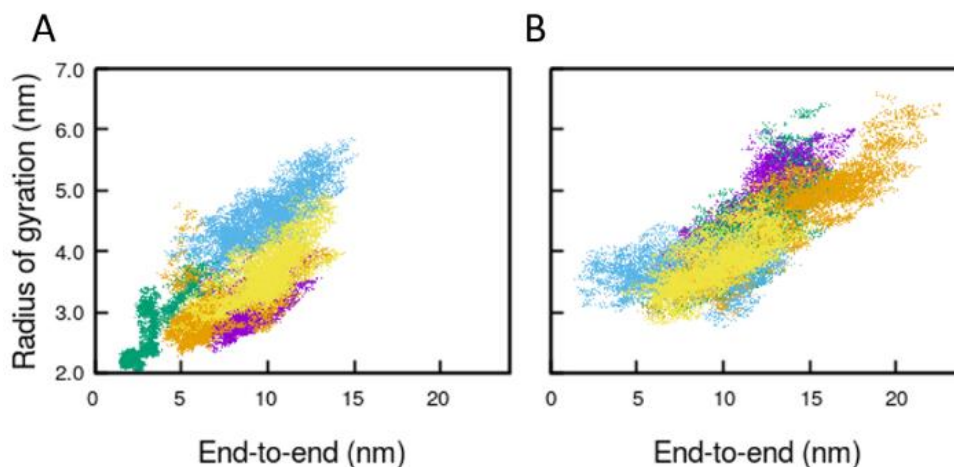


Figure 6.14 Radius of gyration vs end-to-end distance of five all-atoms simulations of the ID and TRD domains run with Amber99SB*-ILDNP (A) and CHARMM36IDPSFF (B). The peptide is unstructured and can sample a large number of conformations, from compact (low radius of gyration and end-to-end distances) to extended structures (large end-to-end distances). Each simulation is shown in a different color.

In order to understand the role of length in the connecting loop between globules, we simulated the two connecting loops found in the coarse-grained simulations. Two all-atom simulations were performed, one in which the loop spanned residues 228 to 242 (observed in simulation CG1, see Table S6.2), and another with the loop containing residues 161 to 205 (observed in CG4, see Table S6.2). The initial structures were taken from the most representative structure of the two-globule conformation in the corresponding coarse-grained trajectory. Modeller⁹ was used to add the missing side-chains and to obtain all-atom structures. The shorter loop (residues 228-242) sampled conformations with the radius of gyration lower than 1.5 nm, whereas the longer loop (residues 161-205) had conformations with the radius of gyration of up to 2.5 nm (Fig. 6.15).

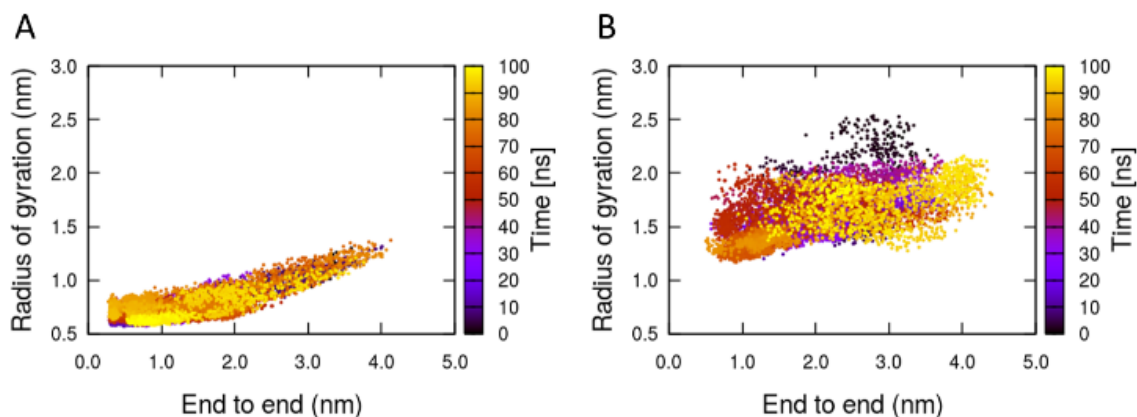


Figure 6.15 Radius of gyration vs end-to-end distance in all-atom simulations of the two connecting loops found in the coarse-grained simulations. **A)** Loop with residues 228-242. **B)** Loop with residues 161-205. The shorter loop sampled more compact structures.

Comparing these two simulations with those of the entire ID and TRD domains (Fig. 6.14) further underlines the relationship between the length of the loop and its compactness for these particular sequences and range of lengths. The shorter loops observed between the two globules may result in insufficient spacing to stabilize the two-globule conformations sampled in the coarse-grained trajectories, which would explain why they eventually merged into a single globule. Moreover, the two-globule all-atom simulations may be unstable due to the poor initial conditions of the loop generated with Modeller. We hypothesize that a stable two-globule conformation would feature a longer separating loop than observed in our simulations.

Comparing the simulations with AlphaFold prediction

Last year, the field of bioinformatics had a major breakthrough when the deep learning model AlphaFold was able to successfully predict the three-dimensional structure of proteins from their sequence¹². Since then, the model has been used to predict 98.5% of the proteins in the human proteome⁵¹; all the structures are available to the community in a

database hosted by the European Bioinformatics Institute (<https://alphafold.ebi.ac.uk>). Nevertheless, predicting the structure of IDPs remains a challenge, as the vast number of low and very low confidence regions from the structures predicted by AlphaFold overlap with regions predicted to be disordered⁵².

The model we built with Modeller⁹ (MeCP2_1) has the N- and C-terminal ends extended into the solvent, and its radius of gyration is large (8.4 nm). In contrast, the model predicted by AlphaFold¹² is much more compact ($R_g = 4.9$ nm) and with an overall spherical shape (Fig. 6.16). The per-residue confidence score (pLDDT) of almost all residues is either low or very low; only the MBD domain was predicted with confidence (pLDDT > 70). Since the model had such low confidence, we used it as the initial structure for three all-atom MD simulations (Table 6.1).

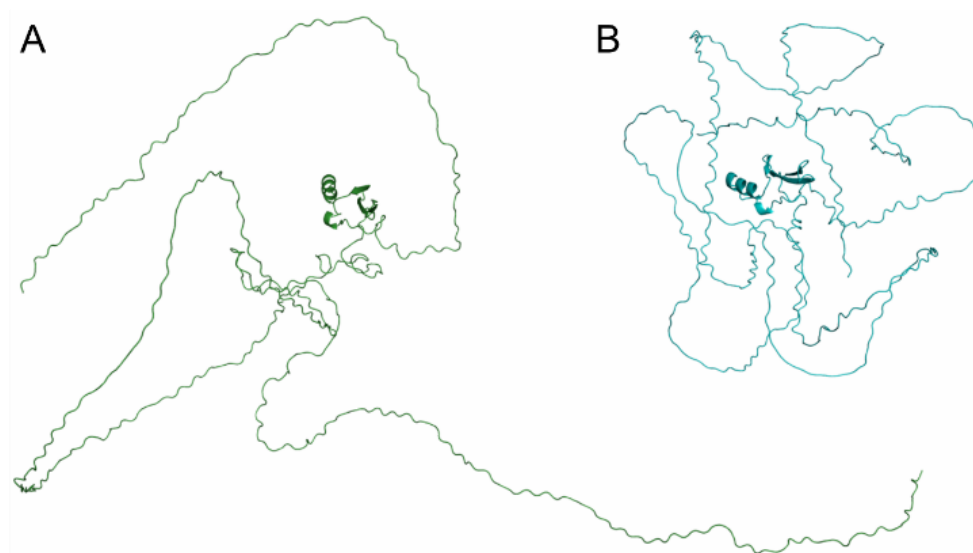


Figure 6.16 Comparison of the models generated by A) Modeller⁹ and B) AlphaFold¹², with their respective MBD domains aligned with each other. The model generated by AlphaFold is much more compact than the one generated by Modeller.

Three replicas of the AlphaFold¹² model were run for 400 ns each. Each simulation sampled a different folding path and converged to a different conformation (Fig. S6.10). Using the PLUMED plugin⁴⁵ for GROMACS¹³, their α -helical content, acylindricity and

asphericity were determined (Fig. S6.11). The conformations sampled by all three AlphaFold¹² simulations have similar asphericity and acylindricity values. They sampled conformations that are more spherical and less cylindrical than the conformations sampled by the MeCP2_1 Modeller⁹ simulation. Since the starting structure has a low radius of gyration ($R_g = 4.9$ nm) we argue that it introduced a bias in the folding path towards more compact structures. Each AlphaFold¹² replica had a different α -helical content, and only one of them sampled values to similar those observed in the Modeller⁹ simulation.

The AlphaFold prediction was not stable in water and, the only exception being the MBD domain, underwent further folding of all of its domains. Although AlphaFold does a remarkable job predicting the presence of disorder, it cannot solve IDP structures⁵³. These simulations should serve as a cautionary tale on the use of predicted models for IDPs; as explained by Strodel in her review⁵⁴, extensive simulations are recommended to equilibrate the protein and sample its conformational space.

6.5 Conclusions

In this work we have presented a multiscale study of MeCP2, comprising six all-atom and four coarse-grained simulations of the full-length protein, as well as twelve all-atom simulations of the ID and TRD domains. Together, they represent the first computational attempt to study the full-length MeCP2 protein.

The initial model was built starting from the NMR structure of the MBD domain¹ and building the rest of the protein by *ab initio* modeling. Two main different conformations were sampled in the coarse-grained simulations: a globular structure similar to the one observed in the all-atom force field and a two-globule conformation. This second conformation was not stable in the all-atom force field, probably because the length of the connecting loop was not long enough to be water-soluble. The conformational ensemble sampled by the 1,550 ns all-atom simulation is in good agreement with the available experimental data^{1,31}. Our model had 4.1% of α -helix content compared to 4% found experimentally³¹. In addition, our model predicted amino acid W104 to be buried, and

amino acids R111 and R133 to be solvent accessible, in accordance with experiments^{37,41,42}. Finally, we used the model predicted by AlphaFold¹² to run three all-atom simulations. The model was not stable in water and underwent further folding. This model is more compact than the one predicted with Modeller⁹, and consequently, it sampled conformations more compact and spherical than those sampled in our Modeller simulations. We recommend caution when using structures of intrinsically disordered proteins predicted by AlphaFold.

With a total of 3 μ s of atomistic simulations and 4.7 μ s of coarse-grained trajectories of full-length MeCP2 models, extensive conformational space of this protein was sampled. Our longest atomistic simulation (MeCP2_1) converged after 800 ns to a very stable structure. When compared to CG, it is reasonable to assume that the all-atom models are more accurate, so the drift of the CG models towards more compact structures is likely to be an artifact. The results show that no single method (atomistic or CG simulations, or AlphaFold modelling) is sufficient on its own for predicting the conformational ensemble of a large IDP such as MeCP2. Our simulations add structural and dynamical detail to the low-resolution information previously available from experiments and could help study disease-associated mutations in their structural context.

We finish by speculating on how the one- and two-globule conformations that were observed in CG simulations and also transiently in MD simulations, could be investigated experimentally – as discussed above, IDPs pose formidable challenges to both experiments and simulations. One possible way might be high-speed atomic force microscopy (HS-AFM) that has very recently been demonstrated to be able to characterize the structure and dynamics of IDPs (polyglutamine tract binding protein-1 and four of its variants as well as two other IDPs) by Kodera *et al*⁵⁵. In particular, for some of their systems they reported temporarily appearing two-globule conformations and order-disorder transition with an associated change in the (relatively short) linking intrinsically disorder region between the globules. Given that MeCP2 has a long and very flexible disordered region spanning the ID and TRD domains, it is tempting to speculate that fluctuations between the one- and two-globule conformation might be directly detectable or/and inducible in HS-AFM. This seems feasible since force spectroscopy⁵⁶ and MD simulations⁵⁷ have shown that for

intrinsically disordered regions forces in the range of a few tens of pN may cause significant stretching and that the free energy barriers are very low. In HS-AFM, the forces are higher up to about 100 pN and there is frictional interaction, albeit very small, with the substrate^{55,58}. Thus, the two-globule state that was only marginally stable in current simulations might also be observable in HS-AFM. Such experiments would potentially also allow investigation of the properties of the linker and the globules.

6.6 Supplemental information

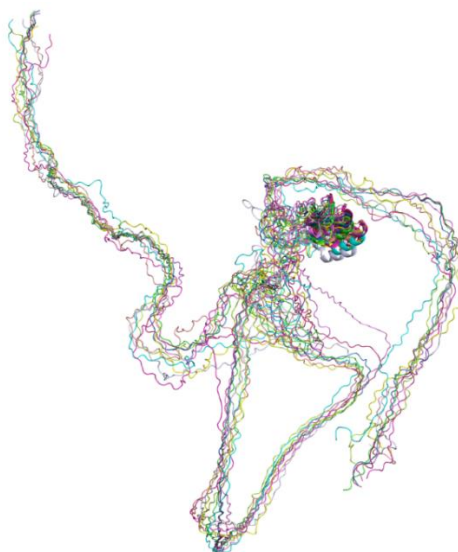


Figure S6.1 Alignment of ten of the rejected models of the full-length MeCP2 protein built with Modeller¹⁸.

Table S6.1 Details of all-atom simulations.

<i>System</i>	<i>Water molecules</i>	<i>Na/Cl each</i>	<i>Counterions (Cl)</i>	<i>Duration (ns)</i>	<i>Starting structure</i>
<i>MeCP2_1</i>	793,027	2,205	37	150	Built with Modeller
<i>MeCP2_1 resized</i>	119,575	337	37	1000	MeCP2_1
<i>MeCP2_2</i>	113,980	324	37	60	Built with Modeller
<i>MeCP2_3</i>	458,344	1,277	37	60	Built with Modeller
<i>MeCP2_3 resized</i>	119,989	339	37	600	MeCP2_3
<i>Loop Amber</i>	324,046	0	31	10	loop in MeCP2_3
<i>Loop A1 resized</i>	71,863	0	31	100	Loop Amber
<i>Loop A2 resized</i>	44,741	0	31	100	Loop Amber
<i>Loop A3 resized</i>	224,502	0	31	100	Loop Amber
<i>Loop A4 resized</i>	129,396	0	31	100	Loop Amber
<i>Loop A5 resized</i>	131,425	0	31	100	Loop Amber
<i>Loop Charmm</i>	324,046	0	31	10	loop in MeCP2_3
<i>Loop C1 resized</i>	186,560	0	31	100	Loop Charmm

<i>Loop C2 resized</i>	281,911	0	31	100	Loop Charmm
<i>Loop C3 resized</i>	54,506	0	31	100	Loop Charmm
<i>Loop C4 resized</i>	293,861	0	31	100	Loop Charmm
<i>Loop C5 resized</i>	102,059	0	31	100	Loop Charmm
<i>Loop 228-242</i>	4,665	0	0	100	Loop in two- globule, CG1
<i>Loop 161-205</i>	13,658	0	12	100	Loop in two- globule, CG4
<i>AlphaFold R1</i>	144,569	411	37	400	AlphaFold model
<i>AlphaFold R2</i>	144,569	411	37	400	AlphaFold model
<i>AlphaFold R3</i>	144,569	411	37	400	AlphaFold model

Table S6.2 Details of coarse-grained simulations. These were run using the PLUM model with implicit water.

<i>System</i>	<i>Duration (ns)</i>
<i>CG1</i>	3,000

<i>CG2</i>	500
<i>CG3</i>	500
<i>CG4</i>	700

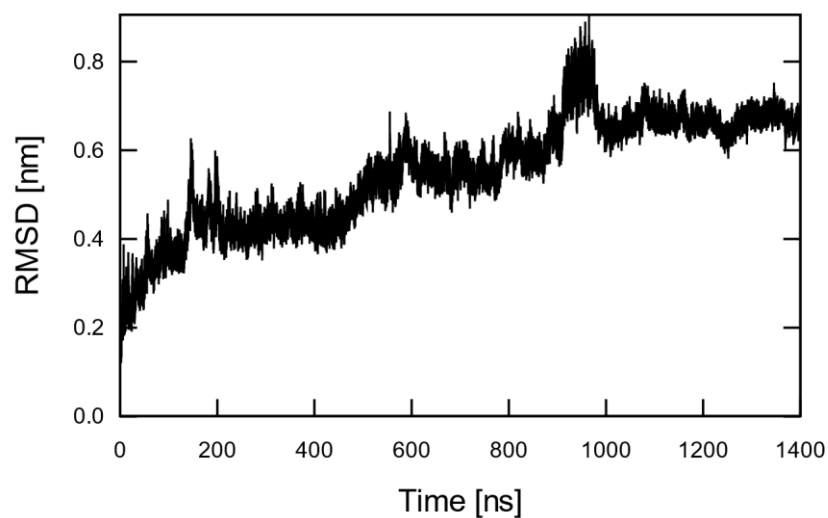


Figure S6.2 RMSD of the TRD domain in the all-atom full-length protein simulation.

Table S6.3 Secondary structure content in the last 400 ns of the all-atom MeCP2_1 simulation clustered with a 0.5 nm cutoff and the Daura *et al.* method³⁹.

<i>Cluster #</i>	<i># frames</i>	<i>α-helix</i>	<i>β-strand/turn</i>	<i>coil/bend</i>
1	23,559	22	121	333
2	11,556	16	99	353
3	2,920	26	110	338
4	1,189	15	112	342

5	499	31	109	327
6	137	8	120	340
7	58	14	109	343
8	58	21	109	338
9	19	23	101	345
10	3	15	115	343
11	3	22	91	350
<i>Total/Average</i>	40,001	20	113	339

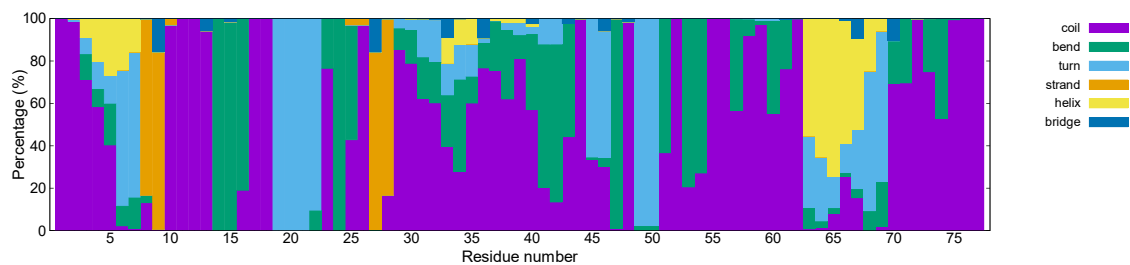


Figure S6.3 Percentage of frames in the last 400 ns of the NTD domain with every type of secondary structure.

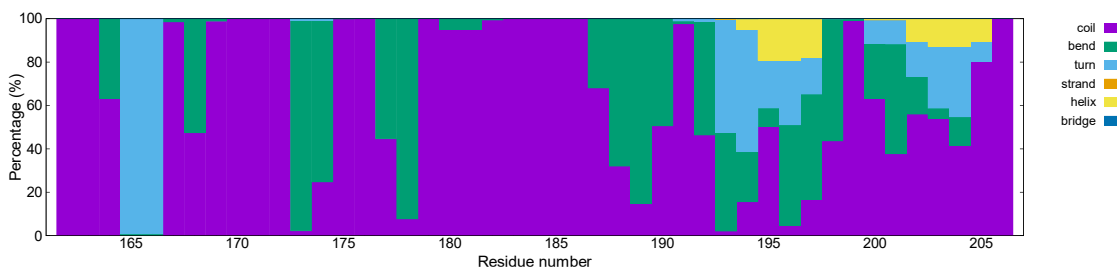


Figure S6.4 Percentage of frames in the last 400 ns of the ID domain with every type of secondary structure.

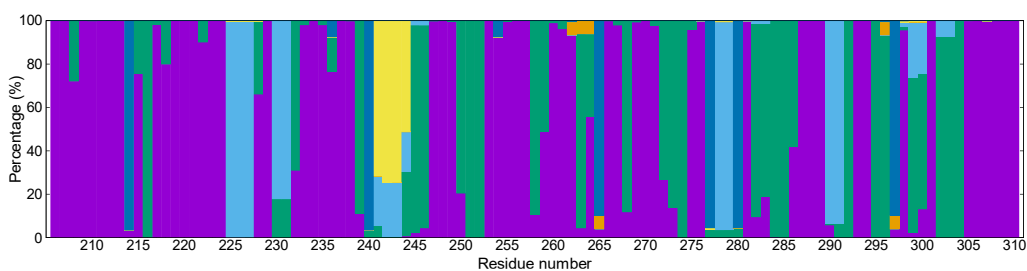


Figure S6.5 Percentage of frames in the last 400 ns of the TRD domain with every type of secondary structure.

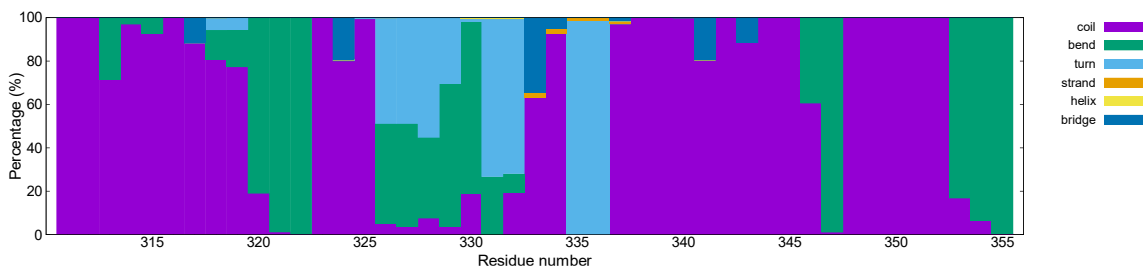


Figure S6.6 Percentage of frames in the last 400 ns of the CTD α domain with every type of secondary structure.

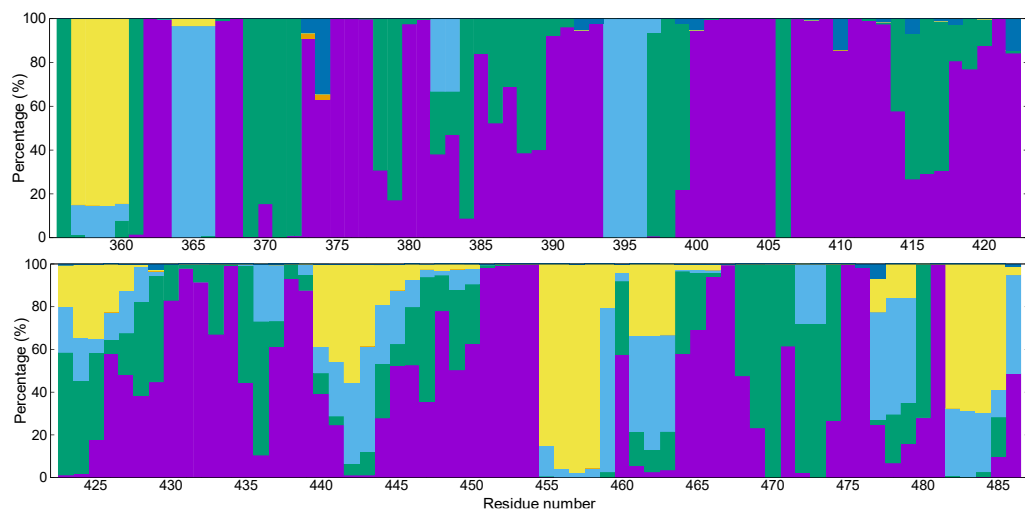


Figure S6.7 Percentage of frames in the last 400 ns of the CTD β domain with every type of secondary structure.

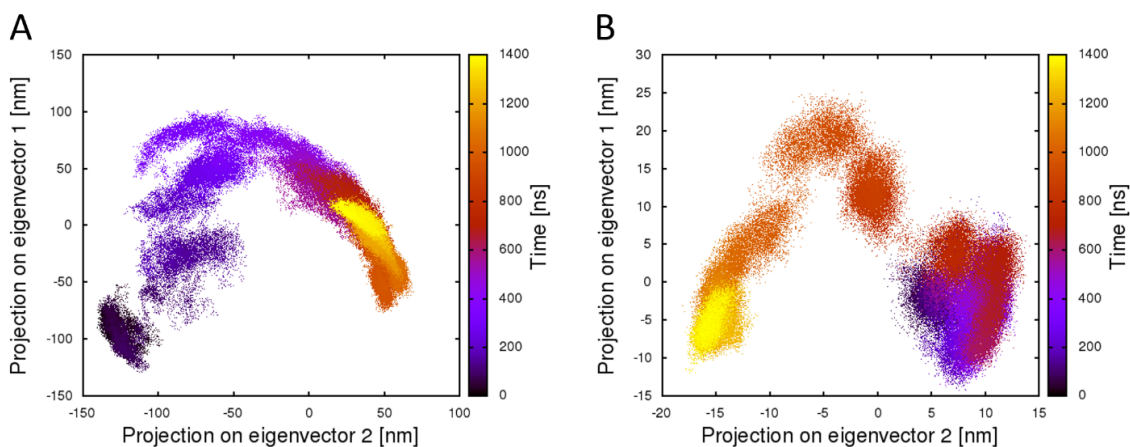


Figure S6.8 Principal Component Analysis. Projection of the MeCP2_1 simulation on the first two eigenvectors, coloured by time. (A) Projection of the entire protein. (B) Projection of the TRD domain. The protein samples the largest portion of this 2D-space during the first 600 ns of the simulation. In contrast, the TRD domain begins to sample more conformational space in the second half of the simulation.

Table S6.4 Relative solvent accessible surface area (rSA) of residue W104 in the four most populated clusters of the all-atom simulation.

<i>Cluster #</i>	<i>weight</i>	<i>rASA</i>
1	23,559	7.4
2	11,556	9.6
3	2,920	6.6
4	1,189	11.4
Weighted average:		8.1

Table S6.5 Relative solvent accessible surface area (rSA) of residue R111 in the four most populated clusters of the all-atom simulation.

<i>Cluster #</i>	<i>weight</i>	<i>rASA</i>
1	23,559	12.7
2	11,556	37.4
3	2,920	5.8
4	1,189	44.3
Weighted average:		20.4

Table S6.6 Relative solvent accessible surface area (rSA) of residue R133 in the four most populated clusters of the all-atom simulation.

<i>Cluster #</i>	<i>weight</i>	<i>rASA</i>
1	23,559	50.3
2	11,556	56.7
3	2,920	65.5
4	1,189	63.0
	Weighted average:	53.7

Table S6.7 Salt bridges in the last 400 ns of the full-length all-atom MeCP2_1 simulation.

Interacting residues	Protein domains
<i>Glu11 - Lys27</i>	NTD – NTD
<i>Asp15 - Arg162</i>	NTD – MBD
<i>Asp17 - Lys135</i>	NTD – MBD
<i>Lys22 - Glu214</i>	NTD – TRD
<i>Glu55 - Lys109</i>	NTD – MBD
<i>Glu55 - Arg111</i>	NTD – MBD
<i>Glu55 - Arg133</i>	NTD – MBD
<i>Glu66 - Arg91</i>	NTD – MBD

<i>Glu66 – Arg85</i>	NTD – MBD
<i>Glu76 – Arg85</i>	NTD – MBD
<i>Asp97 – Lys171</i>	MBD – ID
<i>Arg111 – Asp121</i>	MBD – MBD
<i>Asp154 – Arg167</i>	MBD – ID
<i>Asp156 – Arg167</i>	MBD – ID
<i>Arg168 – Glu205</i>	ID – ID
<i>Glu235 – Lys254</i>	TRD – TRD
<i>Glu235 – Lys347</i>	TRD – CTD α
<i>Arg253 – Glu315</i>	TRD – CTD α
<i>Arg253 – Glu365</i>	TRD – CTD β
<i>Arg255 – Glu258</i>	TRD – TRD
<i>Asp260 – Lys337</i>	TRD – CTD α
<i>Lys266 – Glu282</i>	TRD – TRD
<i>Arg270 – Glu394</i>	TRD – CTD β
<i>Lys271 – Glu404</i>	TRD – CTD β
<i>Glu290 – Lys307</i>	TRD – TRD
<i>Arg294 – Glu318</i>	TRD – CTD α
<i>Glu298 – Lys307</i>	TRD – TRD

<i>Lys304 – Asp407</i>	TRD – CTD β
<i>Lys321 – Glu397</i>	CTD α – CTD β
<i>Arg344 – Glu365</i>	CTD α – CTD β
<i>Arg420 – Glu473</i>	CTD β – CTD β
<i>Glu432 – Arg453</i>	CTD β – CTD β
<i>Arg453 – Glu457</i>	CTD β – CTD β
<i>Glu455 – Arg458</i>	CTD β – CTD β
<i>Glu457 – Arg471</i>	CTD β – CTD β

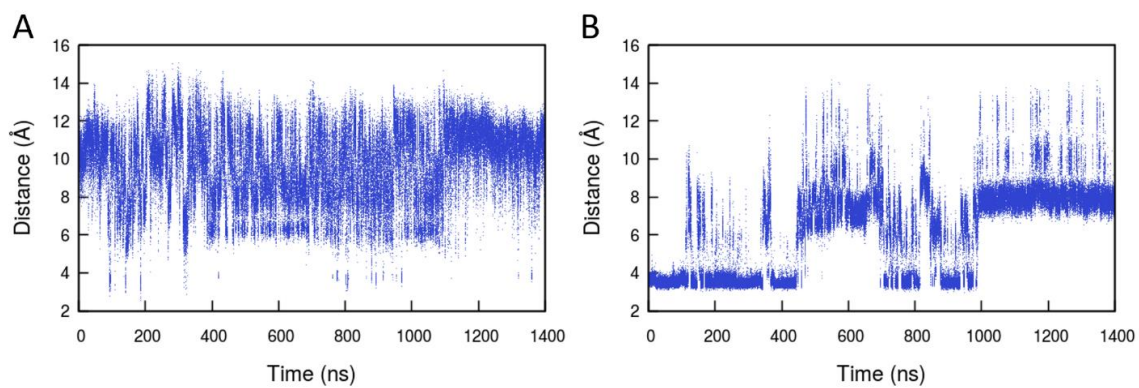
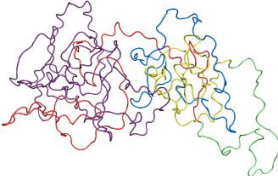
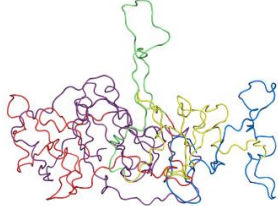
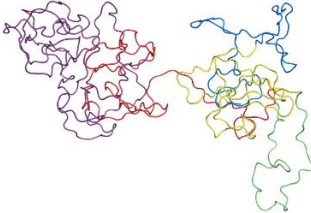
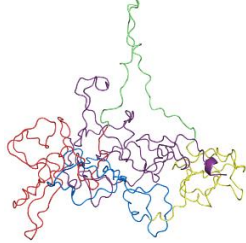



Figure S6.9 Salt bridge interaction between residues Lys 119 and Asp 121 (A) and between residues Arg 133 and Glu 137 (B).

Table S6.8 Clusters sampled in the coarse-grained simulations. The first eight clusters contain 95.3% of the sampled structures.

<i>Cluster #</i>	<i># Structures</i>	<i>Sampled in</i>	<i>Representative structure</i>
1	22,293	CG1, CG2, CG4	
2	5,810	CG1, CG2, CG3, CG4	
3	4,986	CG1	
4	3,492	CG1, CG3	
5	3,027	CG1, CG4	

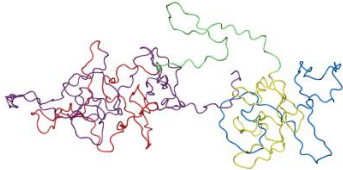
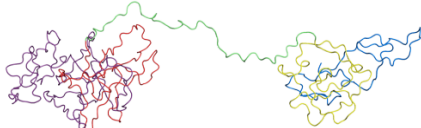
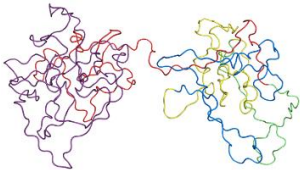


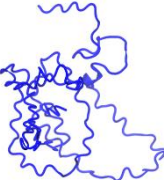
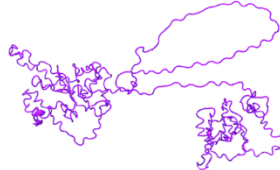
6	2,768	CG4	
7	1,257	CG4	
8	1,165	CG1	

Table S6.9 Templates used to generate models MeCP2_2 and MeCP2_3. The blue templates were taken from the coarse-grained simulation CG1, the green template was generated with Pymol⁵², the red template was taken from an all-atom simulation of the NTD+MBD domains⁶⁶ and the purple template was taken from model MeCP2_2.

	Template 1	Template 2	Template 3	Final Model
MeCP2_2				

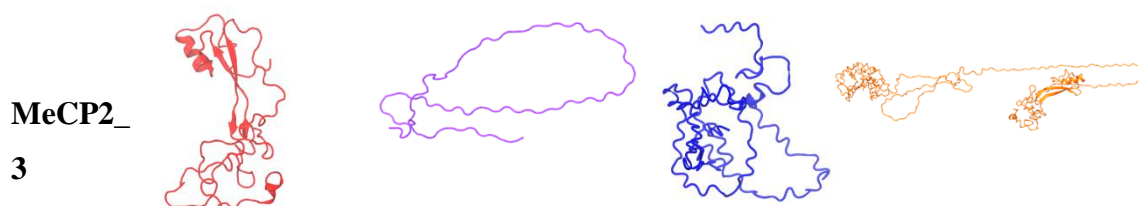


Table S6.10 Conformations sampled by the ID+TRD domains simulations. The five most populated clusters are shown for each simulation.

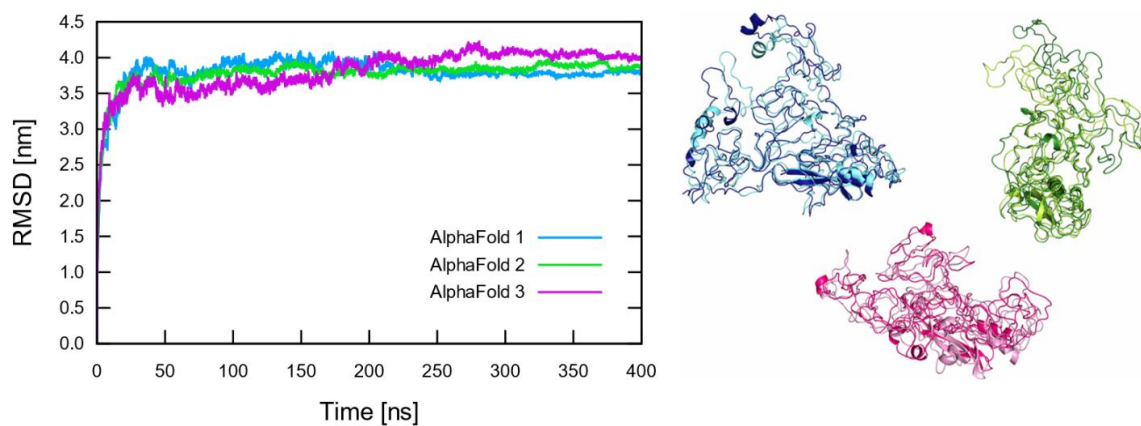
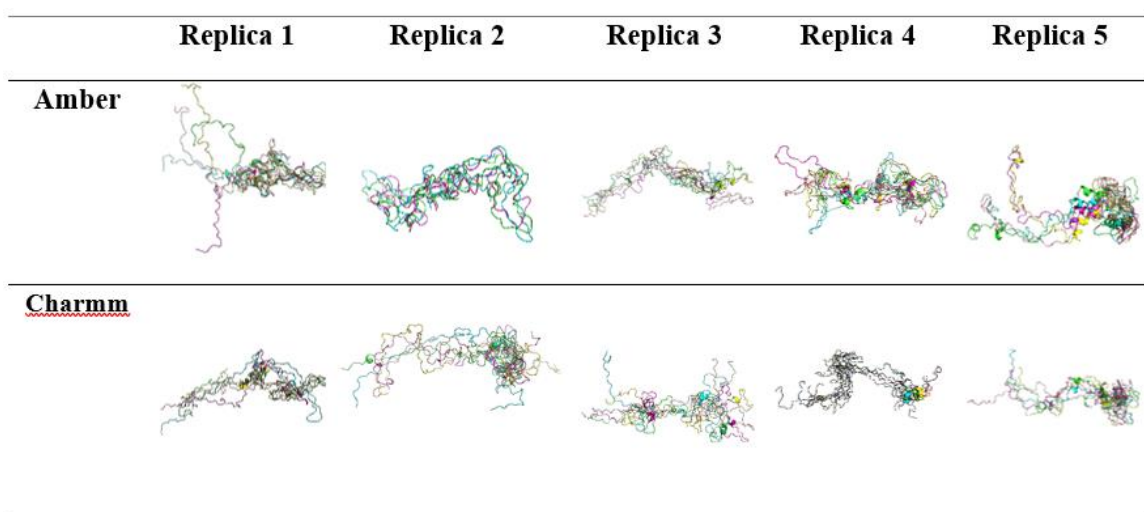


Figure S6.10 RMSD of the three AlphaFold simulations. Right: Alignment of the most sampled structure from 200 to 300 ns (light colours) and the most sampled structure

in the last 100 ns (dark colours) of each replica. Both structures are very similar to each other in the three simulations, ratifying the convergence of the simulations.

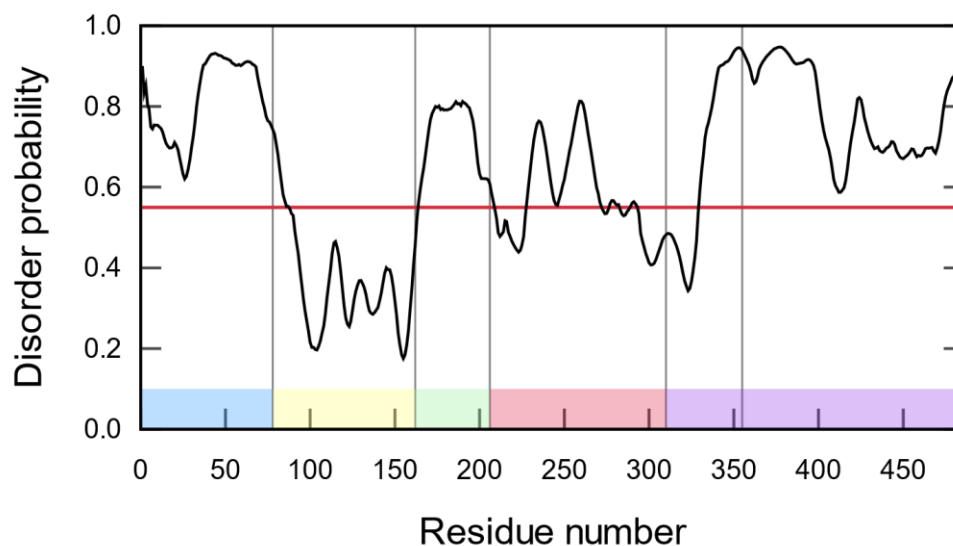


Figure S6.11 Prediction of protein disorder for MeCP2. Residues beyond the red threshold line are predicted to be disordered. The different domains are marked following the color code in Fig 1.

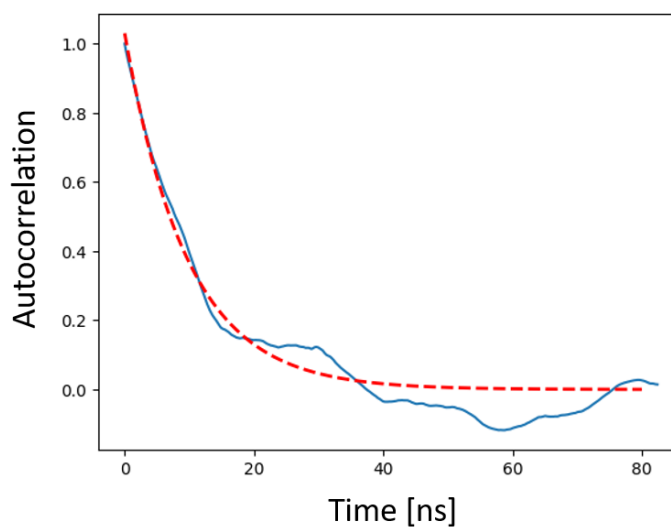


Figure S6.12 Autocorrelation of the radius of gyration of MeCP2_1. Red: exponential fit $y = a \cdot e^{-bt}$ with $a = 1.03027$ and $b = 10.34$ ns.

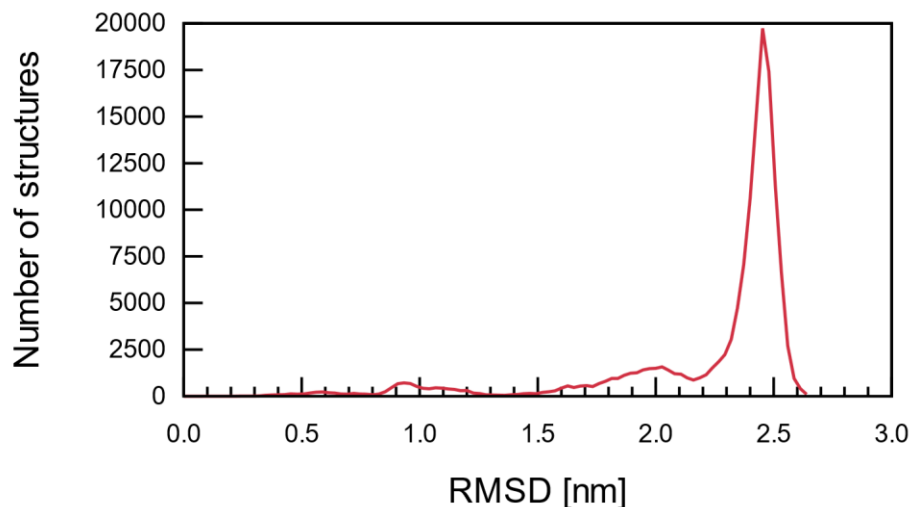


Figure S6.13 Distribution of structures sampled in the MeCP2_1 simulation.

6.7 Acknowledgments

CCG thanks the Province of Ontario Trillium Scholarship Program and Mitacs for their Globalink Research Award. MK thanks the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Canada Research Chairs Program. JH acknowledges support from the French National Research Agency under grant LABEX DYNAMO (ANR-11-LABX-0011), and from the Laboratoire International Associé CNRS/UIUC. Computing facilities were provided by SHARCNET (www.sharcnet.ca), Compute Canada (www.computecanada.ca).

6.8 References

- (1) Wakefield, R. I. D.; Smith, B. O.; Nan, X.; Free, A.; Soteriou, A.; Uhrin, D.; Bird, A. P.; Barlow, P. N. The Solution Structure of the Domain from MeCP2 That Binds to Methylated DNA. *J. Mol. Biol.* **1999**, *291* (5), 1055–1065. <https://doi.org/10.1006/jmbi.1999.3023>.
- (2) Hite, K. C.; Adams, V. H.; Hansen, J. C. Recent Advances in MeCP2 Structure and

- Function. *Biochem. Cell Biol.* **2009**, *87* (1), 219–227. <https://doi.org/10.1139/O08-115>.
- (3) Hansen, J. C.; Wexler, B. B.; Rogers, D. J.; Hite, K. C.; Panchenko, T.; Ajith, S.; Black, B. E. DNA Binding Restricts the Intrinsic Conformational Flexibility of Methyl CpG Binding Protein 2 (MeCP2). *J. Biol. Chem.* **2011**, *286* (21), 18938–18948. <https://doi.org/10.1074/jbc.M111.234609>.
- (4) Hite, K. C.; Kalashnikova, A. A.; Hansen, J. C. Coil-to-Helix Transitions in Intrinsically Disordered Methyl CpG Binding Protein 2 and Its Isolated Domains. *Protein Sci.* **2012**, *21* (4), 531–538. <https://doi.org/10.1002/pro.2037>.
- (5) Kucukkal, T. G.; Alexov, E. Structural, Dynamical, and Energetical Consequences of Rett Syndrome Mutation R133C in MeCP2. *Comput. Math. Methods Med.* **2015**, *2015*.
- (6) Yang, Y.; Kucukkal, T. G.; Li, J.; Alexov, E.; Cao, W. Binding Analysis of Methyl-CpG Binding Domain of MeCP2 and Rett Syndrome Mutations. *ACS Chem. Biol.* **2016**, *11* (10), 2706–2715. <https://doi.org/10.1021/acscchembio.6b00450>.
- (7) Kurcinski, M.; Kolinski, A.; Kmiecik, S. Mechanism of Folding and Binding of an Intrinsically Disordered Protein as Revealed by Ab Initio Simulations. *J. Chem. Theory Comput.* **2014**, *10* (6), 2224–2231. <https://doi.org/10.1021/ct500287c>.
- (8) Nicolau-Junior, N.; Giuliatti, S. Modeling and Molecular Dynamics of the Intrinsically Disordered E7 Proteins from High- and Low-Risk Types of Human Papillomavirus. *J. Mol. Model.* **2013**, *19* (9), 4025–4037. <https://doi.org/10.1007/s00894-013-1915-8>.
- (9) Šali, A.; Blundell, T. L. Comparative Protein Modelling by Satisfaction of Spatial Restraints. *J. Mol. Biol.* **1993**, *234*, 779–815.
- (10) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic Local Alignment Search Tool. *J. Mol. Biol.* **1990**, *215* (3), 403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2).

- (11) Leinonen, R.; Garcia Diez, F.; Binns, D.; Fleischmann, W.; Lopez, R.; Apweiler, R. UniProt Archive. *Bioinformatics* **2004**, *20* (17), 3236–3237. <https://doi.org/10.1093/bioinformatics/bth191>.
- (12) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Židek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596*, 583–589. <https://doi.org/10.1038/s41586-021-03819-2>.
- (13) James Abraham, M.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS : High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, *2*, 19–25. <https://doi.org/10.1016/j.softx.2015.06.001>.
- (14) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D. Comparison of Simple Potential Functions for Simulating Liquid Water Liquid Water. *J. Chem. Phys.* **1983**, *79*, 926–935.
- (15) Aliev, A. E.; Kulke, M.; Khaneja, H. S.; Chudasama, V.; Sheppard, T. D.; Lanigan, R. M. Motional Timescale Predictions by Molecular Dynamics Simulations: Case Study Using Proline and Hydroxyproline Sidechain Dynamics. *Proteins* **2014**, *82* (2), 195–215. <https://doi.org/10.1002/prot.24350>.
- (16) Liu, H.; Song, D.; Lu, H.; Luo, R.; Chen, H. F. Intrinsically Disordered Protein-Specific Force Field CHARMM36IDPSFF. *Chem. Biol. Drug Des.* **2018**, *92* (4), 1722–1735. <https://doi.org/10.1111/cbdd.13342>.
- (17) Samantray, S.; Yin, F.; Kav, B.; Strodel, B. Different Force Fields Give Rise to Different Amyloid Aggregation Pathways in Molecular Dynamics Simulations. *J. Chem. Inf. Model.* **2020**, *60* (12), 6462–6475.

<https://doi.org/10.1021/acs.jcim.0c01063>.

- (18) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An $N \cdot \log(N)$ Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98* (12), 10089–10092. <https://doi.org/10.1063/1.464397>.
- (19) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* **1995**, *103* (19), 8577–8593. <https://doi.org/10.1063/1.470117>.
- (20) Bussi, G.; Donadio, D.; Parrinello, M. Canonical Sampling through Velocity Rescaling. *J. Chem. Phys.* **2007**, *126* (1), 1–7. <https://doi.org/10.1063/1.2408420>.
- (21) Parrinello, M.; Rahman, A. Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method. *J. Appl. Phys.* **1981**, *52* (12), 7182–7190. <https://doi.org/10.1063/1.328693>.
- (22) Hess, B. P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. *J. Chem. Theory Comput.* **2008**, *4* (1), 116–122. <https://doi.org/10.1021/ct700200b>.
- (23) Michaud-agrawal, N.; Denning, E. J.; Woolf, T. B.; Beckstein, O. Software News and Updates MDAnalysis: A Toolkit for the Analysis of Molecular Dynamics Simulations. *J. Comput. Chem.* **2011**, *32* (10), 2319–2327. <https://doi.org/10.1002/jcc>.
- (24) Gowers, R.; Linke, M.; Barnoud, J.; Reddy, T.; Melo, M.; Seyler, S.; Domański, J.; Dotson, D.; Buchoux, S.; Kenney, I.; Beckstein, O. MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. *Proc. 15th Python Sci. Conf.* **2016**, No. Scipy, 98–105. <https://doi.org/10.25080/majora-629e541a-00e>.
- (25) Berau, T.; Deserno, M. Generic Coarse-Grained Model for Protein Folding and Aggregation. *J. Chem. Phys.* **2009**, *130* (23). <https://doi.org/10.1063/1.3152842>.
- (26) Haaga, J.; Gunton, J. D.; Buckles, C. N.; Rickman, J. M. Early Stage Aggregation of a Coarse-Grained Model of Polyglutamine. *J. Chem. Phys.* **2018**, *148* (4).

<https://doi.org/10.1063/1.5010888>.

- (27) Berau, T.; Globisch, C.; Deserno, M.; Peter, C. Coarse-Grained and Atomistic Simulations of the Salt-Stable Cowpea Chlorotic Mottle Virus (SS-CCMV) Subunit 26-49: β -Barrel Stability of the Hexamer and Pentamer Geometries. *J. Chem. Theory Comput.* **2012**, 8 (10), 3750–3758. <https://doi.org/10.1021/ct200888u>.
- (28) Berau, T.; Bennett, W. F. D.; Pfaendtner, J.; Deserno, M.; Karttunen, M. Folding and Insertion Thermodynamics of the Transmembrane WALP Peptide. *J. Chem. Phys.* **2015**, 143 (24). <https://doi.org/10.1063/1.4935487>.
- (29) Rutter, G. O.; Brown, A. H.; Quigley, D.; Walsh, T. R.; Allen, M. P. Testing the Transferability of a Coarse-Grained Model to Intrinsically Disordered Proteins. *Phys. Chem. Chem. Phys.* **2015**, 17 (47), 31741–31749. <https://doi.org/10.1039/c5cp05652g>.
- (30) Daura, X.; Gademann, K.; Jaun, B.; Seebach, D.; Van Gunsteren, W. F.; Mark, A. E. Peptide Folding: When Simulation Meets Experiment. *Angew. Chemie - Int. Ed.* **1999**, 38 (1–2), 236–240. [https://doi.org/10.1002/\(sici\)1521-3773\(19990115\)38:1/2<236::aid-anie236>3.0.co;2-m](https://doi.org/10.1002/(sici)1521-3773(19990115)38:1/2<236::aid-anie236>3.0.co;2-m).
- (31) Adams, V. H.; McBryant, S. J.; Wade, P. A.; Woodcock, C. L.; Hansen, J. C. Intrinsic Disorder and Autonomous Domain Function in the Multifunctional Nuclear Protein, MeCP2. *J. Biol. Chem.* **2007**, 282 (20), 15057–15064. <https://doi.org/10.1074/jbc.M700855200>.
- (32) Kabsch, W.; Sander, C. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-bonded and Geometrical Features. *Biopolymers* **1983**, 22 (12), 2577–2637. <https://doi.org/10.1002/bip.360221211>.
- (33) Joosten, R. P.; Te Beek, T. A. H.; Krieger, E.; Hekkelman, M. L.; Hooft, R. W. W.; Schneider, R.; Sander, C.; Vriend, G. A Series of PDB Related Databases for Everyday Needs. *Nucleic Acids Res.* **2011**, 39 (SUPPL. 1), 411–419. <https://doi.org/10.1093/nar/gkq1105>.

- (34) Hendrich, B.; Bird, A. Identification and Characterization of a Family of Mammalian Methyl-CpG Binding Proteins. *Mol. Cell. Biol.* **1998**, *18* (11), 6538–6547. <https://doi.org/10.1128/mcb.18.11.6538>.
- (35) Klose, R. J.; Sarraf, S. A.; Schmiedeberg, L.; McDermott, S. M.; Stancheva, I.; Bird, A. P. DNA Binding Selectivity of MeCP2 Due to a Requirement for A/T Sequences Adjacent to Methyl-CpG. *Mol. Cell* **2005**, *19* (5), 667–678. <https://doi.org/10.1016/j.molcel.2005.07.021>.
- (36) Ghosh, R. P.; Nikitina, T.; Horowitz-Scherer, R. A.; Gierasch, L. M.; Uversky, V. N.; Hite, K.; Hansen, J. C.; Woodcock, C. L. Unique Physical Properties and Interactions of the Domains of Methylated DNA Binding Protein 2. *Biochemistry* **2010**, *49* (20), 4395–4410. <https://doi.org/10.1021/bi9019753>.
- (37) Ghosh, R. P.; Horowitz-Scherer, R. A.; Nikitina, T.; Gierasch, L. M.; Woodcock, C. L. Rett Syndrome-Causing Mutations in Human MeCP2 Result in Diverse Structural Changes That Impact Folding and DNA Interactions. *J. Biol. Chem.* **2008**, *283* (29), 20523–20534. <https://doi.org/10.1074/jbc.M803021200>.
- (38) Heinig, M.; Frishman, D. STRIDE: A Web Server for Secondary Structure Assignment from Known Atomic Coordinates of Proteins. *Nucleic Acids Res.* **2004**, *32* (WEB SERVER ISS.), 500–502. <https://doi.org/10.1093/nar/gkh429>.
- (39) Kim, H.; Park, H. Prediction of Protein Relative Solvent Accessibility with Support Vector Machines and Long-Range Interaction 3D Local Descriptor. *Proteins Struct. Funct. Genet.* **2004**, *54* (3), 557–562. <https://doi.org/10.1002/prot.10602>.
- (40) Gress, A.; Kalinina, O. V. SphereCon - A Method for Precise Estimation of Residue Relative Solvent Accessible Area from Limited Structural Information. *Bioinformatics* **2020**, *36* (11), 3372–3378. <https://doi.org/10.1093/bioinformatics/btaa159>.
- (41) Ho, K. L.; McNae, I. W.; Schmiedeberg, L.; Klose, R. J.; Bird, A. P.; Walkinshaw, M. D. MeCP2 Binding to DNA Depends upon Hydration at Methyl-CpG. *Mol. Cell*

- 2008**, 29 (4), 525–531. <https://doi.org/10.1016/j.molcel.2007.12.028>.
- (42) Lei, M.; Tempel, W.; Chen, S.; Liu, K.; Min, J. Plasticity at the DNA Recognition Site of the MeCP2 MCG-Binding Domain. *Biochim. Biophys. Acta - Gene Regul. Mech.* **2019**, 1862 (9), 194409. <https://doi.org/10.1016/j.bbagr.2019.194409>.
- (43) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graph.* **1996**, 14 (1), 33–38. [https://doi.org/10.1016/0263-7855\(96\)00018-5](https://doi.org/10.1016/0263-7855(96)00018-5).
- (44) Schrödinger, L. The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC. 2015.
- (45) Bonomi, M.; Branduardi, D.; Bussi, G.; Camilloni, C.; Provasi, D.; Raiteri, P.; Donadio, D.; Marinelli, F.; Pietrucci, F.; Broglia, R. A.; Parrinello, M. PLUMED: A Portable Plugin for Free-Energy Calculations with Molecular Dynamics. *Comput. Phys. Commun.* **2009**, 180 (10), 1961–1972. <https://doi.org/10.1016/j.cpc.2009.05.011>.
- (46) Pietrucci, F.; Laio, A. A Collective Variable for the Efficient Exploration of Protein Beta-Sheet Structures: Application to SH3 and GB1. *J. Chem. Theory Comput.* **2009**, 5 (9), 2197–2201. <https://doi.org/10.1021/ct900202f>.
- (47) Šolc, K. Shape of a Random-Flight Chain. *J. Chem. Phys.* **1971**, 55 (1).
- (48) Tolmachev, D. A.; Boyko, O. S.; Lukasheva, N. V.; Martinez-Seara, H.; Karttunen, M. Overbinding and Qualitative and Quantitative Changes Caused by Simple Na⁺ and K⁺ Ions in Polyelectrolyte Simulations: Comparison of Force Fields with and without NBFIX and ECC Corrections. *J. Chem. Theory Comput.* **2020**, 16 (1), 677–687. <https://doi.org/10.1021/acs.jctc.9b00813>.
- (49) Rauscher, S.; Gapsys, V.; Gajda, M. J.; Zweckstetter, M.; De Groot, B. L.; Grubmüller, H. Structural Ensembles of Intrinsically Disordered Proteins Depend Strongly on Force Field: A Comparison to Experiment. *J. Chem. Theory Comput.* **2015**, 11 (11), 5513–5524. <https://doi.org/10.1021/acs.jctc.5b00736>.

- (50) Chang, M.; Wilson, C. J.; Karunatileke, N. C.; Moselhy, M. H.; Karttunen, M.; Choy, W. Y. Exploring the Conformational Landscape of the Neh4 and Neh5 Domains of Nrf2 Using Two Different Force Fields and Circular Dichroism. *J. Chem. Theory Comput.* **2021**, *17* (5), 3145–3156. <https://doi.org/10.1021/acs.jctc.0c01243>.
- (51) Tunyasuvunakool, K.; Adler, J.; Wu, Z.; Green, T.; Zielinski, M.; Židek, A.; Bridgland, A.; Cowie, A.; Meyer, C.; Laydon, A.; Velankar, S.; Kleywegt, G. J.; Bateman, A.; Evans, R.; Pritzel, A.; Figurnov, M.; Ronneberger, O.; Bates, R.; Kohl, S. A. A.; Potapenko, A.; Ballard, A. J.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Clancy, E.; Reiman, D.; Petersen, S.; Senior, A. W.; Kavukcuoglu, K.; Birney, E.; Kohli, P.; Jumper, J.; Hassabis, D. Highly Accurate Protein Structure Prediction for the Human Proteome. *Nature* **2021**, *596* (7873), 590–596. <https://doi.org/10.1038/s41586-021-03828-1>.
- (52) Ruff, K. M.; Pappu, R. V. AlphaFold and Implications for Intrinsically Disordered Proteins. *J. Mol. Biol.* **2021**, *433* (20), 167208. <https://doi.org/10.1016/j.jmb.2021.167208>.
- (53) Wilson, C. J.; Choy, W.; Karttunen, M. AlphaFold2 : A Role for Disordered Protein Prediction? *bioRxiv* **2021**. <https://doi.org/https://doi.org/10.1101/2021.09.27.461910>.
- (54) Strodel, B. Energy Landscapes of Protein Aggregation and Conformation Switching in Intrinsically Disordered Proteins. *J. Mol. Biol.* **2021**, No. xxxx. <https://doi.org/10.1016/j.jmb.2021.167182>.
- (55) Kodera, N.; Noshiro, D.; Dora, S. K.; Mori, T.; Habchi, J.; Blocquel, D.; Gruet, A.; Dosnon, M.; Salladini, E.; Bignon, C.; Fujioka, Y.; Oda, T.; Noda, N. N.; Sato, M.; Lotti, M.; Mizuguchi, M.; Longhi, S.; Ando, T. Structural and Dynamics Analysis of Intrinsically Disordered Proteins by High-Speed Atomic Force Microscopy. *Nat. Nanotechnol.* **2021**, *16* (2), 181–189. <https://doi.org/10.1038/s41565-020-00798-9>.
- (56) Solanki, A.; Neupane, K.; Woodside, M. T. Single-Molecule Force Spectroscopy of

Rapidly Fluctuating, Marginally Stable Structures in the Intrinsically Disordered Protein α -Synuclein. *Phys. Rev. Lett.* **2014**, *112* (15), 1–6. <https://doi.org/10.1103/PhysRevLett.112.158103>.

- (57) Cheng, S.; Cetinkaya, M.; Gräter, F. How Sequence Determines Elasticity of Disordered Proteins. *Biophys. J.* **2010**, *99* (12), 3863–3869. <https://doi.org/10.1016/j.bpj.2010.10.011>.
- (58) Ando, T.; Uchihashi, T.; Scheuring, S. Filming Biomolecular Processes by High-Speed Atomic Force Microscopy. *Chem. Rev.* **2014**, *114* (6), 3120–3188. <https://doi.org/10.1021/cr4003837>.

7 Conclusions and future directions

7.1 Conclusions

7.1.1 Conclusions

Using a computational approach, we characterized the structure and dynamics of two biologically relevant proteins: a globular protein and an IDP. This work illustrates the challenges faced when dealing with an IDPs, as well as the importance of having long simulations. The primary conclusions of this thesis are presented below.

7.1.2 All-atom MD simulations can be enough to provide insight into globular proteins

Most of the structures of globular proteins have been solved, either by experimental methods or by homology modeling with high confidence in the prediction¹. This makes, in principle, performing all-atom MD simulations very straightforward. Furthermore, often there is no lack of experimental data to which to compare the results obtained *in silico*. The challenge lies in that oftentimes there is so much information already available, that one has to be creative in order to provide new insights, standard analyses are not enough. Chapter 4 showed how state-of-the-art techniques such as residue interaction network analysis can be successfully applied to explain the mechanisms that give rise to different behaviours in two highly similar homologous enzymes. It also underlines the need for long simulations. As the results showed, some of the previously reported results from shorter trajectories were transient and no longer observed after the first microsecond of simulation.

7.1.3 MD simulations can be a complimentary technique to experimental procedures

High-speed atomic force microscopy (HS-AFM) is one of the few techniques that can be used to evaluate the structure and dynamics of intrinsically disordered proteins². An important question that arises when using this technique is whether the molecular

behaviour observed in the protein under study is affected by the tip-sample and surface-sample interactions. It has been shown that the transferred energy from the AFM tip to sample is partitioned amongst all degrees of freedom of the molecule, so that the transferred energy per degree of freedom is negligibly small; importantly, it dissipates quickly, over a time much shorter than $1 \mu\text{s}$. However, there is no direct way to assess the impact of surface-sample interaction. By performing MD simulations of the MBD domain of MeCP2 in solution and in the presence of a surface, we were able to show that the surface-sample interaction did not affect the secondary structure of the protein, validating the experimental results. The surface used in the simulation had the same surface charge density as the surface in the experimental setup.

7.1.4 MD simulations can provide new insights when it's difficult to obtain experimental data

The structural characterization of a large IDP such as MeCP2 is a challenge. Indeed, our results show that no single method is sufficient on its own for predicting the conformational ensemble of MeCP2. A combination of all-atom and coarse-grained simulations, as well as backmapping to atomic structures, and extensive simulation times were needed. Thankfully, we do have some structural information against which we could validate our models. The all-atom model MeCP2_1 was in good agreement with all available experimental data, and the coarse-grained simulations sampled a structure similar to those observed in HS-AFM experiments. Our simulations help to put together a more complete picture where experiments had only looked at individual features and they provide predictions for new experiments.

7.2 Future directions

We studied two native TIM enzymes (TcTIM and TbTIM) as well as a chimeric protein in chapter 4. We examined their electrostatic interactions and found some significant differences in the hydrogen bonds of the three proteins. This led us to hypothesize that the

thermal stability of the chimeric protein is higher than TbTIM but lower than TcTIM, which would need to be validated by experiments.

In chapter 6, we characterized the structure and dynamics of the full-length MeCP2 protein. Our results can be the basis of further studies such as studying disease-associated mutations in their structural context or its interactions with its biological partners.

7.3 References

- (1) Tunyasuvunakool, K.; Adler, J.; Wu, Z.; Green, T.; Zielinski, M.; Židek, A.; Bridgland, A.; Cowie, A.; Meyer, C.; Laydon, A.; Velankar, S.; Kleywegt, G. J.; Bateman, A.; Evans, R.; Pritzel, A.; Figurnov, M.; Ronneberger, O.; Bates, R.; Kohl, S. A. A.; Potapenko, A.; Ballard, A. J.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Clancy, E.; Reiman, D.; Petersen, S.; Senior, A. W.; Kavukcuoglu, K.; Birney, E.; Kohli, P.; Jumper, J.; Hassabis, D. Highly Accurate Protein Structure Prediction for the Human Proteome. *Nature* **2021**, *596* (7873), 590–596. <https://doi.org/10.1038/s41586-021-03828-1>.
- (2) Ando, T.; Uchihashi, T.; Fukuma, T. High-Speed Atomic Force Microscopy for Nano-Visualization of Dynamic Biomolecular Processes. *Prog. Surf. Sci.* **2008**, *83* (7–9), 337–437. <https://doi.org/10.1016/j.progsurf.2008.09.001>.
- (3) Ando, T. High-Speed Atomic Force Microscopy and Its Future Prospects. *Biophys. Rev.* **2018**, *10* (2), 285–292. <https://doi.org/10.1007/s12551-017-0356-5>.

Curriculum Vitae

Name: Cecilia Chavez Garcia

Post-secondary Education and Degrees:

University of Western Ontario
London, Ontario, Canada
2017-Current Ph.D.

National Autonomous University of Mexico
Mexico City, Mexico
2015-2017 M.Sc. with honors

National Autonomous University of Mexico
Mexico City, Mexico
2009-2014 B.Sc.

Honours and Awards:

APS & ICTP-SAIFR Young Physicists Forum on Biological
Physics travel award
São Paulo, Brazil, 2020

Mitacs Globalink Research Award
Paris, France, 2019

Chemical Biophysics symposium travel award
Toronto, Canada, 2018

Western Graduate Research Scholarship
2017-2021

Ontario Trillium Scholarship
2017-2021

Publications:

1. Chávez-García, C.; Aguayo-Ortiz, R.; Dominguez, L. Quantifying Correlations between Mutational Sites in the Catalytic Subunit of γ -Secretase. *J. Mol. Graph. Model.* 2019, 88, 221–227. <https://doi.org/10.1016/j.jmgm.2019.02.002>.

2. Rohoullah, F.; Sowlati-Hashjin, S.; Chávez-García, C.; Mitra, A., Hossein Karimi-Jafarif, M.; Karttunen, M. Identification of Catechins Binding Pockets in Monomeric A β 42 Through Ensemble Docking and MD Simulations. *To be submitted*
3. Chávez-García, C.; Karttunen, M. Highly similar sequence and structure yet different biophysical behaviour: A computational study of two triosephosphate isomerases. *Submitted (JCIM)*
4. Kodera, N.; Kalashnikova, A.; Porter-Goff, M. E.; Musselman, C.; Chávez-García, C.; Karttunen, M.; Demeler, B.; Kutateladze, T.; Ando, T.; Hansen, J. Structure and dynamics of the Rett syndrome protein, MeCP2. *Submitted (Nucleic Acids Res)*
5. Chávez-García, C.; Hénin, J.; Karttunen, M. A multiscale computational study of the conformation of the full-length intrinsically disordered protein MeCP2. *Submitted (JCIM)*