Deriving Protein Structures Efficiently by Integrating Experimental Data into Biomolecular Simulations

Zur Erlangung des akademischen Grads eines

DOKTORS DER NATURWISSENSCHAFTEN (Dr. rer. nat.)

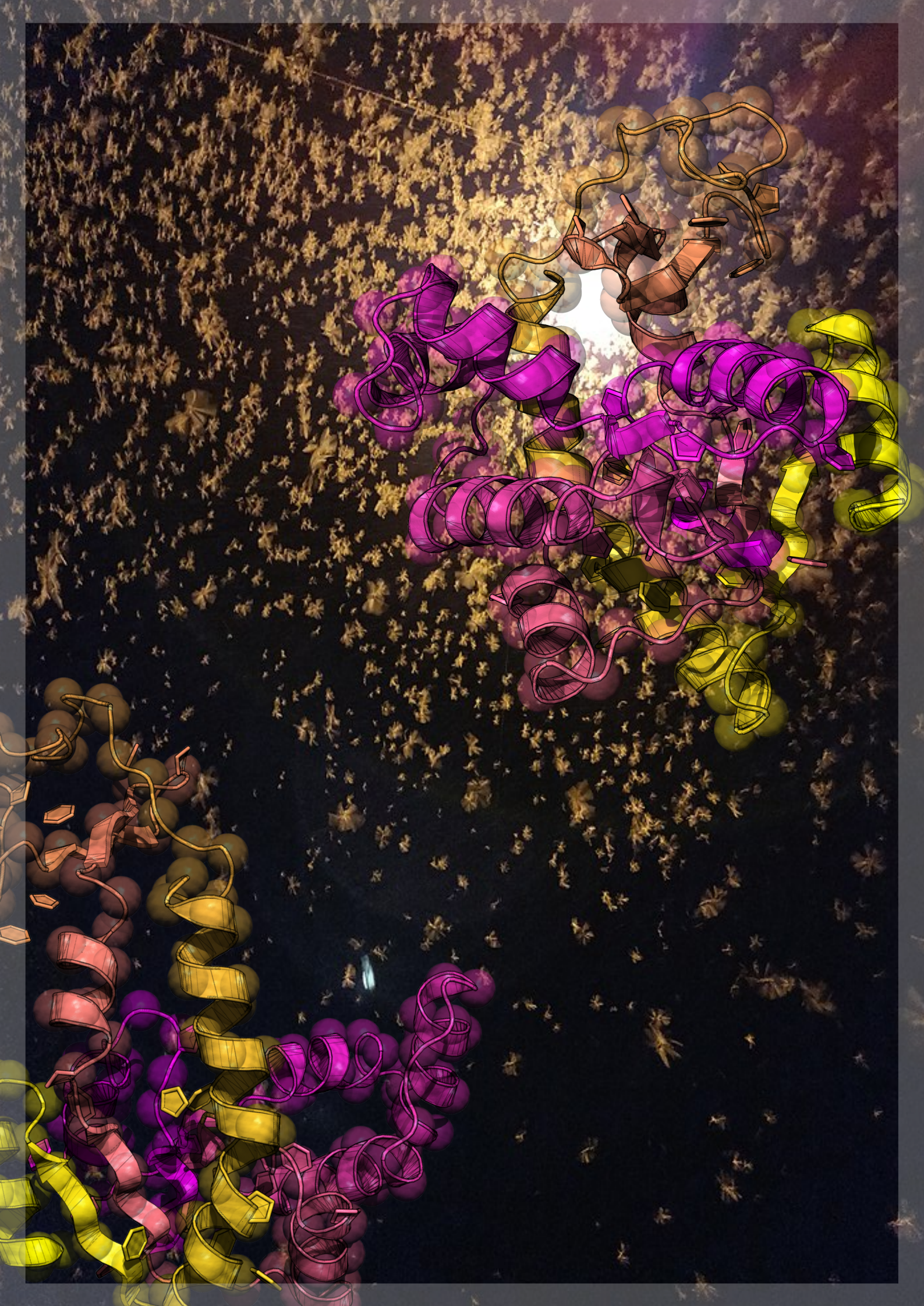
von der KIT-Fakultät für Physik des
Karlsruher Instituts für Technologie
angenommene

DISSERTATION

von

Marie Weiel-Potyagaylo, M.Sc.

Tag der mündlichen Prüfung: 9. Juli 2021
Erster Gutachter: Prof. Dr. Wolfgang Wenzel
Zweiter Gutachter: Prof. Dr. Alexander Schug



About the cover:

"Winged termites swarming" by T. R. Shankar Raman used under CC BY-SA 4.0 license, via Wikimedia Commons. Protein structures visualized in PyMOL[1]. Licensed by Marie Weiel-Potyagaylo under CC BY-SA 4.0.

"There's so many shades of black."

The Raconteurs

"If You Don't Like Cool Quotes Check Your GMXRC File."

Sincerely yours, the Sysadmin.

Abstract

Proteins are the molecular nanomachines in biological cells. They are vital to all known forms of life, ranging from single-cell organisms to complex beings like humans, and fulfill various functions, such as transporting oxygen and building hair. However, protein malfunction is associated with severe diseases such as Alzheimer's and Parkinson's. The development of effective treatments requires a comprehensive understanding of their biological function and structural dynamics, which we still lack despite massive advancements of molecular imaging and structure prediction.

Experimentally, proteins can only be observed indirectly and the measured data is often incomplete or ambiguous. Here, physics-based modeling of biomolecular dynamics can fill the gap. Data-assisted molecular dynamics simulations have emerged as a new paradigm to combine the structural puzzle pieces from various sources in silico and obtain a complete picture of protein structure and dynamics in atomic detail. Such simulations incorporate the experimental data as an integral component in a biased physical model, thereby eliminating ambiguities with their complementary theoretical knowledge.

In this thesis, I explore the capabilities and limits of structure-based models as a tool to access the sparse information content of structural protein data. These models efficiently describe the behavior emerging from a protein's evolutionarily optimized native topology. I introduce `XSBM`, a structure-based simulation framework for systematically deriving protein structures in accordance with the data from small-angle X-ray scattering experiments with a focus on minimizing computational demands. As a data-assisted method, its performance crucially depends on simulation parameters, where the key challenge is balancing experimental information and physical knowledge. I show how computational intelligence can be used to efficiently explore such parameter spaces, determine functional parameter sets, and optimize the performance of complex physics-based simulation techniques. To this end, I introduce `FLAPS`, a data-driven solution for fully automated and reproducible parameter searches for biomolecular simulations. Inspired by the emergent behavior of natural bird flocks and fish schools, `FLAPS` is a self-adapting particle-swarm based optimizer that solves the problem of weighting various quality features in multi-response optimization in a more general context.

Together with recent advances in structure prediction through artificial intelligence, performance-optimized data-assisted simulations as presented in this thesis can help push our understanding of the intricate relation between protein structure and function. Such computational methods can contextualize and fit the available puzzle pieces of structural information together, thus deepening our understanding of proteins as the ultimate building blocks of life.

Zusammenfassung

Proteine sind molekulare Nanomaschinen in biologischen Zellen. Sie sind wesentliche Bausteine aller bekannten Lebensformen, von Einzellern bis hin zu Menschen, und erfüllen vielfältige Funktionen, wie beispielsweise den Sauerstofftransport im Blut oder als Bestandteil von Haaren. Störungen ihrer physiologischen Funktion können jedoch schwere degenerative Krankheiten wie Alzheimer und Parkinson verursachen. Die Entwicklung wirksamer Therapien für solche Proteinfehlfaltungserkrankungen erfordert ein tiefgreifendes Verständnis der molekularen Struktur und Dynamik von Proteinen. Da Proteine aufgrund ihrer lichtmikroskopisch nicht mehr auflösbaren Größe nur indirekt beobachtet werden können, sind experimentelle Strukturdaten meist uneindeutig. Dieses Problem lässt sich in silico mittels physikalischer Modellierung biomolekularer Dynamik lösen. In diesem Feld haben sich datengestützte Molekulardynamiksimulationen als neues Paradigma für das Zusammenfügen der einzelnen Datenbausteine zu einem schlüssigen Gesamtbild der enkodierten Proteinstruktur etabliert. Die Strukturdaten werden dabei als integraler Bestandteil in ein physikbasiertes Modell eingebunden.

In dieser Arbeit untersuche ich, wie sogenannte strukturbasierte Modelle verwendet werden können, um mehrdeutige Strukturdaten zu komplementieren und die enthaltenen Informationen zu extrahieren. Diese Modelle liefern eine effiziente Beschreibung der aus der evolutionär optimierten nativen Struktur eines Proteins resultierenden Dynamik. Mithilfe meiner systematischen Simulationsmethode `XSBM` können biologische Kleinwinkelröntgenstreudaten mit möglichst geringem Rechenaufwand als physikalische Proteinstrukturen interpretiert werden. Die Funktionalität solcher datengestützten Methoden hängt stark von den verwendeten Simulationsparametern ab. Eine große Herausforderung besteht darin, experimentelle Informationen und theoretisches Wissen in geeigneter Weise relativ zueinander zu gewichten. In dieser Arbeit zeige ich, wie die entsprechenden Simulationsparameterräume mit Computational-Intelligence-Verfahren effizient erkundet und funktionale Parameter ausgewählt werden können, um die Leistungsfähigkeit komplexer physikbasierter Simulationstechniken zu optimieren. Ich präsentiere `FLAPS`, eine datengetriebene metaheuristische Optimierungsmethode zur vollautomatischen, reproduzierbaren Parametersuche für biomolekulare Simulationen. `FLAPS` ist ein adaptiver partikelschwarmbasierter Algorithmus inspiriert vom Verhalten natürlicher Vogel- und Fischschwärme, der das Problem der relativen Gewichtung verschiedener Kriterien in der multivariaten Optimierung generell lösen kann.

Neben massiven Fortschritten in der Verwendung von künstlichen Intelligenzen zur Proteinstrukturvorhersage ermöglichen leistungsoptimierte datengestützte Simulationen detaillierte Einblicke in die komplexe Beziehung von biomolekularer Struktur, Dynamik und Funktion. Solche computergestützten Methoden können Zusammenhänge zwischen den einzelnen Puzzleteilen experimenteller Strukturinformationen herstellen und so unser Verständnis von Proteinen als den Grundbausteinen des Lebens vertiefen.

Acknowledgments

"Baby, It Ain't Over Till It's Over."

Lenny Kravitz

The microbiologist and Nobel laureate André Lwoff once noted that "the researcher's art is first of all to find himself a good boss". This is very true, and I want to express my greatest gratitude to my doctoral supervisor and mentor, Alexander Schug. He always granted me a lot of freedom and encouraged me to pursue my own ideas and work independently on my projects. With his infectious enthusiasm and creativity, he motivated me to keep going and always had an open ear whenever I felt lost in my work. I thank Wolfgang Wenzel for being my first supervisor and supporting me and my graduation with his constructive and uncomplicated manner.

My biggest thanks go out to Markus Götz, the best informal advisor and future boss I could have ever wished for, for guiding me and helping me find meaning in my scientific work. I thank Oskar for never getting tired of answering my questions and solving all my computer problems. Special thanks go out to Daniel Coquelin for taking the time to edit my scribblings and listening to me in the middle of the night on Teams. I felt comfortable in your company all the time.

Thanks go out to all Karlsruhe and Jülich group members, in particular to Jakob for ranting together in Zoom about the ugly fact of having to write a PhD thesis, to Julian for being the funniest and most laid-back office mate on earth, to Arthur for always being my friend and supplying me with delicious food, and to Ines for being the most reliable workmate and, of course, for going together on our extended conference trip to Israel.

I want to say thanks with all my heart to my family who have always believed in me and made me feel loved, especially to my adorable sister Madeleine for always being there for me. Thank you, Philipp, for being the person you are for me (and letting me hit you as hard as I can every time I felt rushed with my work).

List of Publications

This thesis contains peer-reviewed articles about the work conducted during my doctoral studies:

- **Weiel, M.**, Reinartz, I., and Schug, A., 2019. *Rapid interpretation of small-angle X-ray scattering data*. PLoS Computational Biology, doi: 10.1371/journal.pcbi.1006900.
- Reinartz, I., **Weiel, M.**, and Schug, A., 2020. *FRET Dyes Significantly Affect SAXS Intensities of Proteins*. Israel Journal of Chemistry, doi: 10.1002/ijch.202000007.
- Christiansen, A., **Weiel, M.**, Winkler, A., Schug, A., and Reinstein, J., 2021. *The Trimeric Major Capsid Protein of Mavirus is stabilized by its Interlocked N-termini Enabling Core Flexibility for Capsid Assembly*. Journal of Molecular Biology, doi: 10.1016/j.jmb.2021.166859.
- **Weiel, M.**, Götz, M., Klein, A., Coquelin, D., Floca, R., and Schug, A., 2021. *Dynamic particle swarm optimization of biomolecular simulation parameters with flexible objective functions*. Nature Machine Intelligence, doi: 10.1038/s42256-021-00366-3.

Other contributions carried out throughout my doctoral studies, not included in this thesis:

- Voronin, A., **Weiel, M.**, and Schug, A., 2020. *Including residual contact information into replica-exchange MD simulations significantly enriches native-like conformations*. PLoS ONE, doi: 10.1371/journal.pone.0242072.

Contents

Abbreviations

3D Three-Dimensional.

ADK Adenylate Kinase.

AF Alexa Fluor.

ATP Adenosine Triphosphate.

B680 Biotium Dye CF680R.

CASP Critical Assessment of Protein Structure Prediction.

CD Circular Dichroism.

CI-2 Chymotrypsin Inhibitor 2.

ClyA Cytolysin A.

Cryo EM Cryogenic Electron Microscopy.

CspTm Cold Shock Protein from *Thermotoga maritima*.

DJR Double Jelly Roll.

DSSP Define Secondary Structure of Proteins.

^{10}FNIII Tenth Type III Domain of Fibronectin.

FRET Förster Resonance Energy Transfer.

GdmCl Guanidinium Chloride.

GDT Global Distance Test.

HDX-MS Hydrogen-Deuterium Exchange Coupled to Mass Spectrometry.

IDP Intrinsically Disordered Protein.

Ile Isoleucine.

LAO protein Lysine-, Arginine-, Ornithine-Binding Protein.

MCP Major Capsid Protein.

MCP_3 Trimeric Mavirus Hexon.

MD Molecular Dynamics.

mRNA Messenger RNA.

NMR Nuclear Magnetic Resonance.

OF Objective Function.

PDB Protein Data Bank.

PSO Particle Swarm Optimization.

RMSD Root-Mean-Square Deviation.

RMSF Root-Mean-Square Fluctuation.

SANS Small-Angle Neutron Scattering.

SAXS Small-Angle X-Ray Scattering.

SBM Structure-Based Model.

SWAXS Small- and Wide-Angle X-Ray Scattering.

Trp Tryptophan.

UV Ultraviolet.

VHP Villin Headpiece.

$\mathbf{VHP}_{\mathbf{54}}^{\mathbf{74}}$ VHP Subregion between Residues 54 and 74.

XS-Guided MD X-Ray Scattering-Guided Molecular Dynamics.

For my dad.

"It seemed a good idea at first".

Gerrit Groenhof

1

Motivation

This work is on deriving realistic protein structures in silico by integrating experimental data into biomolecular simulations. In the first chapter, I will explain why I worked on this topic. I will paint a picture of proteins and their functional roles in our bodies as seen through the eyes of a physicist. In particular, I want to answer the following questions: What are proteins and what do they do? Why is understanding proteins on a molecular level so important and so difficult? To provide context for the questions I address in this work, I will first review our current knowledge about the physical description as well as the experimental and computational study of proteins, along with arising problems and possible solutions. I will close this chapter by defining the objectives of my work and providing an outline of its structure and scope.

Inside every cell of our bodies, myriads of molecular nanomachines are busy working. They are what enable our eyes to detect light, our neurons to fire, and the instructions in our genome to be read. These ubiquitous nanomachines are proteins. Proteins drive life forward. They are complex biomolecules made of linear chains of amino acids. Their various functions include transportation, cellular communication, structural support, cell motion, and energy balance. Currently, we know of about 200 million proteins, with another 30 million discovered every year. What a protein does and how it works largely depends on its three-dimensional (3D) structure. Determining what structure a protein folds into is known as the "protein folding problem". From an evolutionary perspective, this structure emerges from a "form follows function" principle. This indicates that, as illustrated in Fig. 1.1, a protein's biological function is inextricably linked to its structure and dynamics. A detailed understanding of this function consequently requires resolving the protein's 3D shape. The strong link between structure, dynamics, and function is known as the "structure-function paradigm" and is a fundamental principle in molecular biophysics.

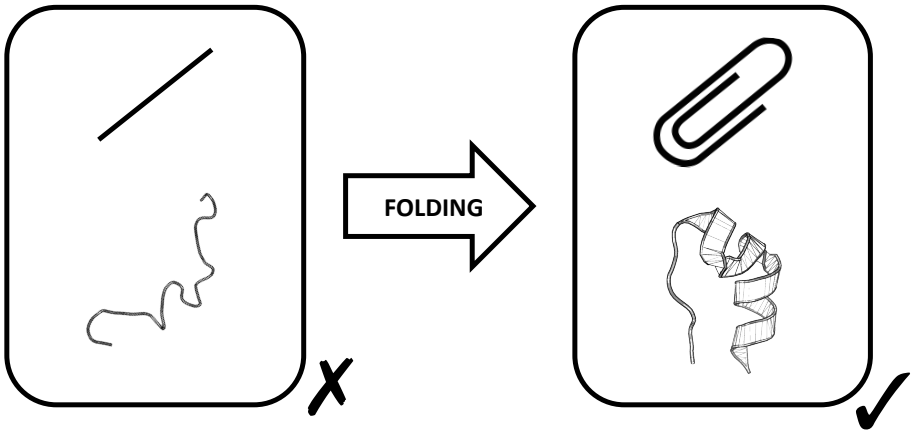


Figure 1.1. The "paper clip" analog. Functional protein structure has emerged as the result of evolutionary processes according to a "form follows function" principle. Similar to a paper clip, a protein is only functional when folded properly.

Why is this important? Proteins enable the complex biochemistry within cells via enzymatic activity and signaling processes, however their malfunction can lead to severe diseases. When healthy proteins lose their normal structure, they can form pathogenic amyloids that deposit around cells as fibrous plaques. These plaques interfere with the physiological function of tissues and organs and have been associated with more than 50 diseases, including Alzheimer's and Parkinson's[2,3], type-2 diabetes[4], and Creutzfeld-Jakob disease[5]. The ability to predict a protein's structure from its sequence is invaluable for understanding how abnormal molecular interactions cause human diseases. Recently, such efforts have proven to be instrumental in fighting the ongoing COVID-19 pandemic. Deciphering the structure-function paradigm unlocks the door to solving many of our biggest challenges, such as developing effective medical treatments, finding enzymes to break down industrial waste, and perhaps to eventually unraveling the mysteries of life. As the biophysicist and Nobel laureate Francis Crick put it, "almost all aspects of life are engineered at the molecular level, and without understanding molecules we can only have a very sketchy understanding of life itself"[1].

How can we observe proteins? Protein structure determination has been researched extensively. To date, more than 177 000 experimentally resolved structures are deposited in the Protein Data Bank[6] (PDB)[2]. Because proteins are nanoscale, their structures cannot be observed directly using an optical microscope. Yet, pictures are a key to understanding and scientific breakthroughs often triggered by visualizing systems invisible to the naked eye. Studying proteins requires indirect imaging techniques that capture not only their static structures but also their function-related conformational dynamics. Various complementary methods exist, including X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, electron microscopy, and biological solution scattering. The increasing significance of advanced molecular imaging is highlighted by the 2017 Nobel Prize in Chemistry, awarded to the three physicists Jacques Dubochet, Joachim Frank, and Richard Henderson "for developing cryo-electron microscopy for the high-resolution structure determination of biomolecules in solution"[7]. Experimentally determined structures have allowed deep insights into protein function and recent literature is filled with illustrations of molecular systems, ranging from proteins causing antibiotic resistance to the spiky coat of the coronavirus shown in Fig. 1.2.

[1] From Francis Crick, *What Mad Pursuit: A Personal View of Scientific Discovery* (1988), 61.
[2] https://www.rcsb.org

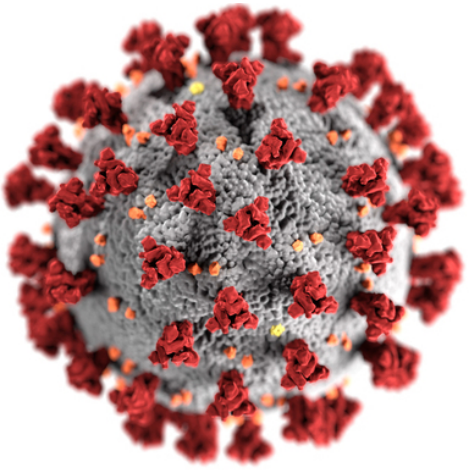


Figure 1.2. The coronavirus. Alissa Eckert and Dan Higgins, medical illustrators with the United States Centers for Disease Control and Prevention, created a "beauty shot" of the coronavirus to bring it to the public's attention. Red: spike proteins, gray: lipid bilayer envelope, yellow: envelope proteins, orange: membrane proteins. Cara Giaimo (2020-04-01). "The Spiky Blob Seen Around the World". The New York Times. Retrieved 2020-04-19.

Why is this insufficient? Experimental protein structure determination typically requires costly equipment and involves extensive trial and error such that resolving a single structure can take years of work. However, the sequential brute-force sampling of all conformations of a protein would take longer than the age of the universe, even at a picosecond rate. Therefore, we only know the exact 3D structures for a tiny fraction of all known proteins. In his famous acceptance speech for the 1972 Nobel Prize in Chemistry, Christian Anfinsen postulated that a protein's shape should be fully determined by its amino acid sequence. This sparked a half-century quest for predicting a protein's 3D structure from only its 1D sequence. Different methods for in silico structure prediction have been proposed, including physics-based, knowledge-based, and data-driven approaches. In 1994, scientists working on the protein folding problem founded the CASP (Critical Assessment of Protein Structure Prediction[3]) community forum. Their biennial contest for blind structure prediction provides researchers the opportunity to validate their computational predictions against ground-truth experimental data and thus serves as an independent assessor of the state of the art. In 2016, the Google affiliate DeepMind turned its AI expertise towards the protein folding problem and for the first time dedicated enormous resources to a real-world research question. In the 2020 CASP competition, their deep-learning based system AlphaFold[8] consistently predicted highly accurate molecular models considered as comparable with experimentally determined structures. Andriy Kryshtafovych, a scientific adjudicator in CASP, described the achievement as "truly remarkable" and claimed the protein folding problem to be "largely solved".

Why do we still need experiments? AlphaFold can only predict static structures of single-chain peptides. Even though proteins are generally thought to adopt unique structures determined by their sequences, they are by no means strictly static. Dynamic conformational transitions occur over a wide range of time and length scales and are tightly coupled to a protein's physiological function. In addition, many proteins function in the form of composite complexes. Important open questions in the field are how multiple peptide chains associate into protein complexes and how proteins interact with each other as well as with DNA, RNA, and other small molecules. Experiments are an irreplaceable tool to not only form a link to real-world observations and validate computational structure prediction techniques, but also to observe a protein's functional dynamics. However, traditional methods such as X-ray crystallography and NMR spectroscopy are limited in their application. X-ray crystallography

[3] https://predictioncenter.org/index.cgi

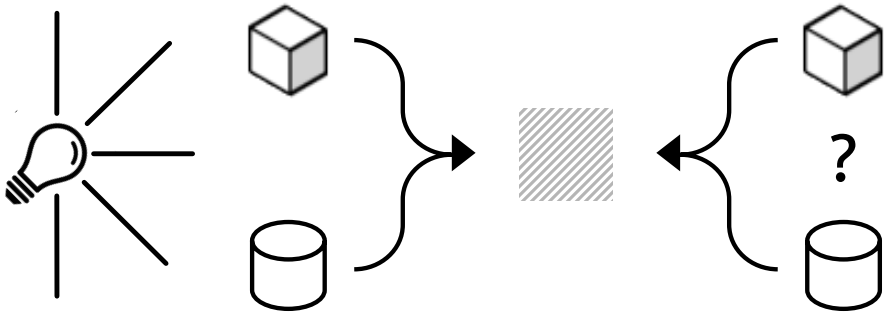


Figure 1.3. The "shadow theater" analog. Indirect methods for structural analysis typically involve an inevitable loss of information during the measurement process. Similar to reconstructing a unique 3D molecular structure from low-information experimental data, recovering a 3D geometrical body from its 2D shadow only is inherently ambiguous.

involves elaborate preparation of protein crystals and NMR spectroscopy is only suited for relatively small proteins. A practical and increasingly popular alternative is biological solution scattering, where dissolved proteins are irradiated by X-rays and the scattering intensity is recorded. Such studies can be conducted under various conditions and provide time-resolved information on structural dynamics[9,10].

What is the primary obstacle for the usage of experimental data? As protein structures are not directly observable, experimental data are often ambiguous, incomplete, or poorly resolved. The sparse information in the data is insufficient to determine a molecular structure's degrees of freedom exhaustively. In biological solution scattering, the desired molecular electron density is the Fourier transform of the experimentally inaccessible, complex-valued scattering amplitude. Recovering a 3D structural model from the measured 1D scattering intensity, that is the absolute amplitude squared, is an ill-posed inverse problem. Simply speaking, this can be thought of as reconstructing a 3D geometrical body from its 2D shadow only, an inherently ambiguous task as illustrated in Fig. 1.3. This implies that the accurate interpretation of indirect experimental data hinges on complementary physical, stereo-chemical, or structural knowledge.

How can we resolve this ambiguity? The potentially most powerful approach is to complement the data with molecular dynamics (MD)[11–15]. Resting on the physical theory of classical mechanics, computational MD simulations provide a time-dependent description of a protein's motions on the atomic level. Static protein structures from, e.g., AlphaFold or experimental methods can be extended by a view of the protein's function-related dynamics. Data-assisted MD incorporates the structural information from experimental data as an integral component into a biased physical model, where the experimentally derived information must be properly balanced with the theoretical knowledge. This balance is determined by the degree of confidence in the measured data versus the physics-based model. Data-assisted approaches have emerged as a new paradigm to interpret structural data on proteins in the form of molecular models, and various applications highlight the potential of combining experimental expertise and physics-based modeling[12,16–18]. As classical MD primarily aims to describe a system's dynamics accurately, such simulations are considerably complex. Their computational costs make them infeasible for simulating function-related slow or large-scale motions and limit their application to small systems. While simulated times are in the order of microseconds with a femtosecond time step, biologically relevant time scales are in the order of milliseconds to seconds, exceeding the capacity of available multipurpose high-performance computing systems.

How can we overcome the time-scale challenge in MD? A brute-force approach to tackle this is the usage of highly specialized supercomputers. In 2008, D. E. Shaw Research launched Anton, a massively parallel special-purpose system for explicit-solvent MD, where they could simulate folding and unfolding transitions of small proteins on a millisecond time scale for the first time. A more efficient and elegant alternative is reducing a molecular system's degrees of freedom by simplifying its interactions in the physical model. To minimize the computational demands and access biological time scales, I focus on so-called structure-based models (SBMs)[19–23]. SBMs probe the dynamics arising from a folded protein's unique geometry realistically. Originally developed to study protein folding, they provide rich information about function-related processes involving major structural changes. Based on energy landscape theory[24,25] and the principle of minimal frustration[26], SBMs can sample large conformational ensembles at considerably lower computational costs without loss of information on the system's essential characteristics. While maintaining full molecular flexibility, they provide a more efficient yet rougher description of a protein's dynamics with fewer parameters than the more fine-grained force fields in classical MD.

My contributions. In this work, I use physico-empirical SBMs as an efficient tool to access and complement the limited structural information in indirect experimental data. My research goal is to provide a self-contained pipeline for systematically deriving realistic protein structures with due consideration of the data. Using the example of biological solution scattering, I explore how such data can be best incorporated into SBMs. As my first main contribution, I present `XSBM`, a data-assisted structure-based framework for rapid interpretation of scattering intensities. I demonstrate my method's efficacy by refining structures towards scattering data for three well-characterized proteins using minimal computational resources and time[27]. An inherent problem of such simulations is their dependence on the non-trivial choice of MD parameters[16,17,27], where simple yet inefficient grid search or accurate yet laborious Bayesian inference is typically used. Targeting a fully automated parameter search for biomolecular simulations, I propose a fundamentally different approach. Inspired by the emergent behavior of natural bird flocks and fish schools, I introduce `FLAPS`, a self-learning swarm-based optimizer, as my second main contribution. To evaluate a data-assisted simulation's quality in terms of physical structures matching the data, I present a new type of flexible and robust objective function. As a showcase example, I successfully apply `FLAPS` to find functional parameters for my `XSBM` simulations, where I present results for two well-characterized proteins.

The structure of my thesis is outlined below. To create a consistent logical flow throughout, I present my own contributions and results not en bloc but embed them within their respective content-related contexts. To provide a clear overview of my work, I refer to my own contributions explicitly as "projects".

PART I - BACKGROUND AND FUNDAMENTALS

In the first part, I provide the required biophysical background knowledge.

- **Chapter 2** gives an introduction to proteins with a special focus on the biomolecular structure-function paradigm and the energy landscape theory of protein folding.
- **Chapter 3** covers experimental methods for protein structure analysis relevant in the context of this work.

PART II - COMPUTATIONAL METHODS AND METHOD DEVELOPMENT

In the second part, I introduce common methodologies used in computational biophysics which serve as a basis for the methods I developed in this work. At this point, I present the key concepts of my two main projects, i.e., `XSBM` and `FLAPS`, along with exemplary applications and corresponding results.

- **Chapter 4** gives an overview of the computational methods used in this work. I introduce the basics of MD and the theoretical concepts underlying minimalist SBMs. To underpin the explanatory power of the latter, I present two of my auxiliary projects as application examples. I show results from my structure-based simulations of the Mavirus virophage's capsomer[28] and a computational study on the mutual influence of solution scattering and Förster resonance energy transfer[29] measurements in combined applications.
- **Chapter 5** introduces the general concept of data-assisted biomolecular simulations. At this point, I present my first main project: `XSBM`, an efficient structure-based method to interpret solution scattering data within biomolecular simulations. I show results from an exemplary application of my `XSBM` method to three well-characterized protein systems.
- **Chapter 6** presents my second main project: `FLAPS`, a bio-inspired method for optimizing MD parameters of data-assisted biomolecular simulations using computational intelligence. I show my method's capability by applying it to my `XSBM` simulations and present results for two well-characterized protein systems.

PART III - CONCLUSIONS

In the third part, I recapitulate the outcomes of my work.

- **Chapter 7** gives a summarizing discussion as well as an outlook for future directions, studies, and applications.

PART I

BACKGROUND AND FUNDAMENTALS

"Let's Unzip And Let's Unfold."

Red Hot Chili Peppers

2

Proteins or "Form Follows Function"

This chapter covers the biological and theoretical basics of this work. Sec. 2.1 gives a biophysical introduction to proteins with a special focus on the "structure-function paradigm". Energy landscape theory, a statistical framework for protein folding, is explained in Sec. 2.2 along with the principle of minimal frustration. These concepts are the foundation of structure-based models, a coarse-grained type of interaction potential used for the simulations in my work.

Proteins are the macromolecular workhorses of biological cells and the main functional components of living organisms. Their various tasks include oxygen transport, DNA replication, signal transduction, enzymatic catalysis, force generation, structural stability, and energy balance. A protein's function is inextricably linked to its structure and dynamics. Therefore, understanding this function requires resolving the protein's 3D fold on the atomic level.

2.1 Protein Structure

A protein is one or multiple polypeptide chain(s) of linearly linked α-amino acids. An amino acid is an organic compound that contains a carboxyl functional group, COOH, an amino functional group, NH_2, and a side chain, R, defining its individual characteristics. α-amino acids have the amino group attached to the C_α atom next to the carboxyl group. In the protein backbone, neighboring amino acids are connected by peptide bonds between their carboxyl and amino groups (see Fig. 2.1). 21 different amino acids, classified into charged, polar, and hydrophobic, are specified in the genetic code of eukaryotes (see Fig. 2.2).

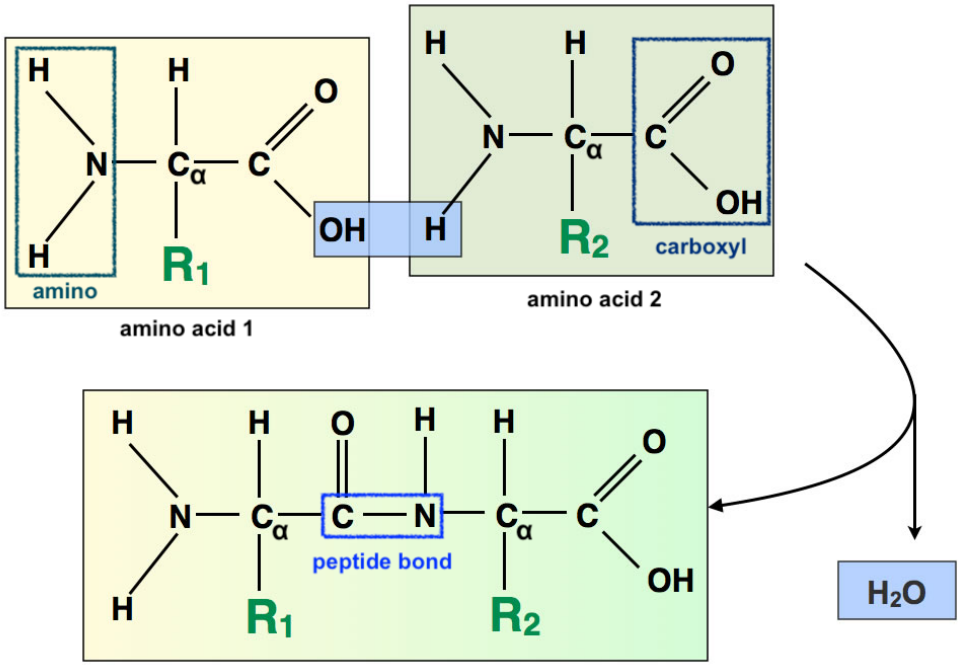


Figure 2.1. The peptide bond. A peptide bond is an amide-type covalent chemical bond that links two consecutive amino acids with side chains R_1 and R_2 in the protein backbone. The first amino acid's carboxyl group reacts with the second amino acid's amino group to form a peptide bond under the release of water, H_2O. Adapted from "Peptidbindung" created by Ulrich Helmich licensed under CC BY-NC-SA 4.0.

Protein structure is organized in four hierarchical levels. A protein's specific amino acid sequence is known as its primary structure. It is dictated by the nucleotide sequence of the protein's genes and uniquely encodes its 3D fold. Secondary structure refers to the local structural motifs defined by regularly repeating patterns of backbone bonding, which are normally stabilized by hydrogen bonds. Most common are the alpha helix and the beta sheet (see Fig. 2.3 c for cartoon representation). Tertiary structure is the entire protein's 3D fold. It is affected by nonlocal interactions such as the formation of a hydrophobic core and stabilized by salt bridges, hydrogen bonds, and disulfide bonds between cysteine residues. A protein's natural fold is also known as its native state. Quaternary structure is the spatial arrangement of multiple polypeptide chains within a functional protein complex.

In 1986, the famous physical chemist and Nobel laureate Linus Pauling said: "Much of what is understood in the field [of biology] is based on the structure of molecules and the properties of molecules in relation to their structure. If you have that basis, then biology isn't just a collection of disconnected facts."[1] This structure-function paradigm is a basic principle in biophysics: A protein's native 3D structure determines its biological function. Proteins are intrinsically dynamic. During their functional cycles, they perform structural changes and traverse between distinct conformational states. These structural rearrangements are tied to their physiological function and bioactivity. Resolving protein structures and dynamics and understanding protein folding from a random coil thus is very important to gain detailed insight into protein function.

[1] From an interview with Neil A. Campbell, in *Crossing the Boundaries of Science*, BioScience (Dec 1986), 36, No. 11, 737.

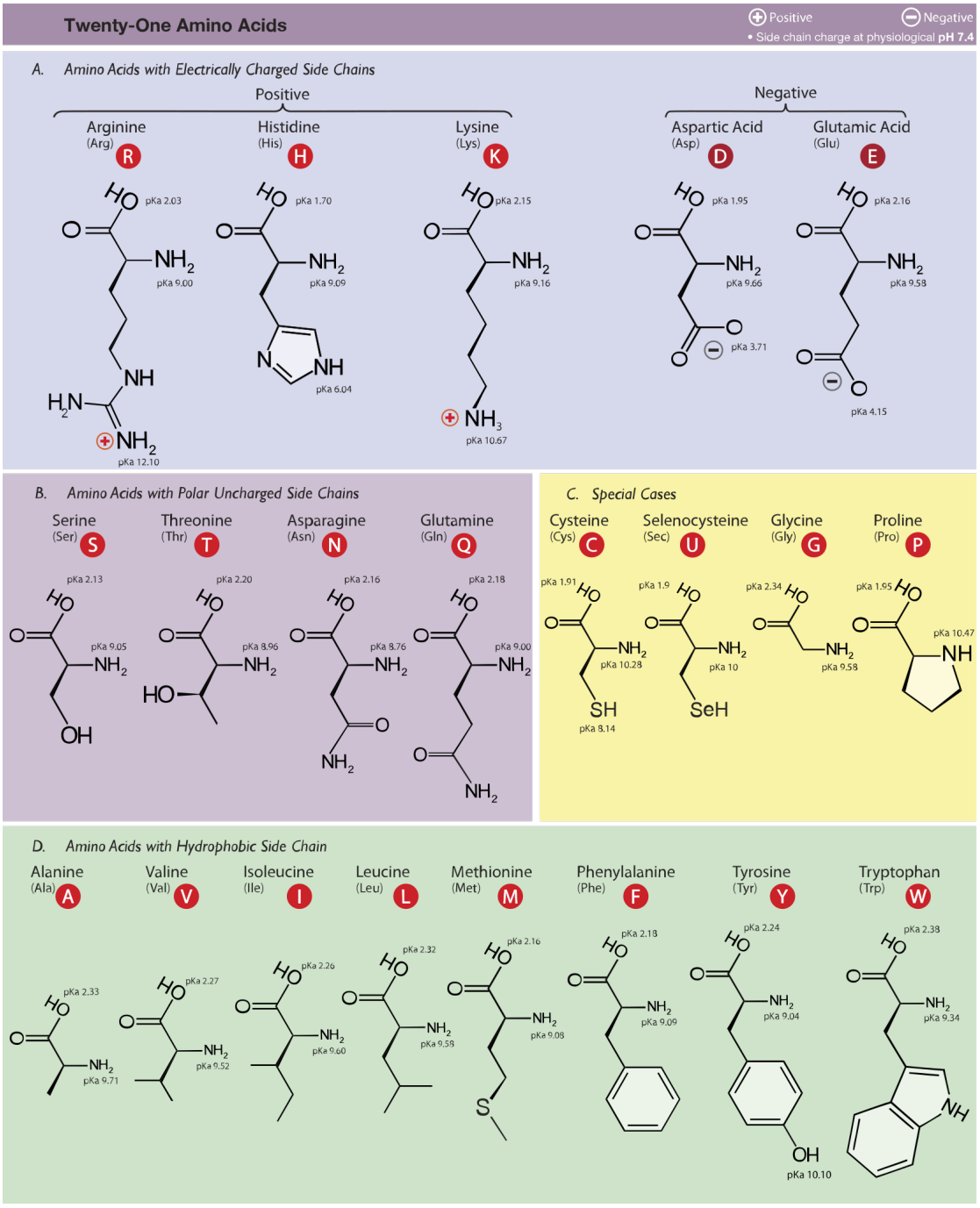


Figure 2.2. Classification of the 21 proteinogenic amino acids in eukaryotes. 21 proteinogenic amino acids are directly encoded in the eukaryotic genetic code. Both 1-letter and 3-letter codes are given. "Molecular structures of the 21 proteinogenic amino acids" by Dan Cojohari used under CC BY-SA 3.0 license, via Wikimedia Commons.

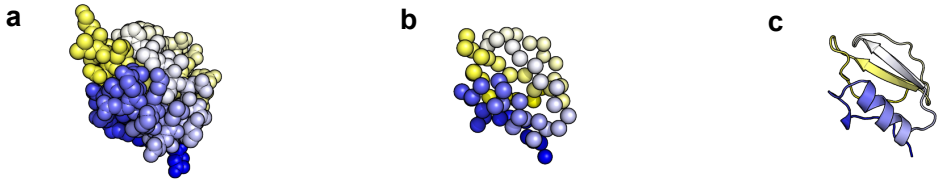


Figure 2.3. Different representations of protein structure. **a**. Sphere representation of all atoms. **b**. Sphere representation of C_α atoms. **c**. Cartoon representation of secondary structure with alpha helix (blue) and beta sheets (yellow and white). Visualized in PyMOL[1].

Anfinsen's dogma states that a protein's native fold is exclusively determined by its sequence, at least for small globular proteins in their usual physiological environment. The native fold then is a unique, stable, and kinetically accessible conformation with minimum free energy[30]. Different theories on actual folding mechanisms exist. In 1969, Levinthal observed that a polypeptide chain of length N with γ degrees of freedom per residue has an astronomic number of γ^N possible conformations[31]. Assuming $\gamma = 3$, one would obtain approximately 10^{48} conformations for a small protein with $N = 100$ residues. If a protein were to acquire its native structure by consecutively sampling its conformational space, the time span needed to find the energetic minimum would exceed the age of the universe. Most proteins, however, spontaneously fold on time scales of milliseconds to seconds[32]. To resolve Levinthal's apparent paradox, folding can alternatively be explained by the rapid formation of local interactions. This leads to an unambiguous pathway with well-defined intermediates between folded and unfolded states that guides the random-coil protein to its native conformation and thereby accelerates the folding process (see Fig. 2.4 b). In fact, such partially unfolded transition states could be observed experimentally. Further development of this approach based on kinetic evidence for the existence of several different folding pathways[33] yields the so-called energy landscape theory. This enhanced model takes into account various folding routes and the ensemble character of conformations. It allows for facilitated folding via an ensemble of multiple converging pathways, which as a whole are known as the transition state ensemble. This theory builds the foundation of the computational model used in this work.

2.2 Energy Landscape Theory

> "There is no such thing as free energy. Anyone who advocates it does not know what he is talking about."
>
> Alireza Haghighat

The energy landscape theory of protein folding gives a statistical description of a protein's potential surface[34]. This surface is a mapping of all possible conformations, or rather their degrees of freedom, to the respective Gibbs free energies[35]. It is also known as the protein's energy landscape and can be understood as a hypersurface in a multidimensional space, where each conformation is represented by a particular point on this surface.

Random amino acid sequences likely have conflicting interaction energies that cannot be minimized simultaneously. Such systems are referred to as geometrically frustrated. Small conformational changes can result in kinetic entrapment within local minima of their rugged energy landscape. This prohibits proper folding and thus, according to the structure-function paradigm, physiological function: The

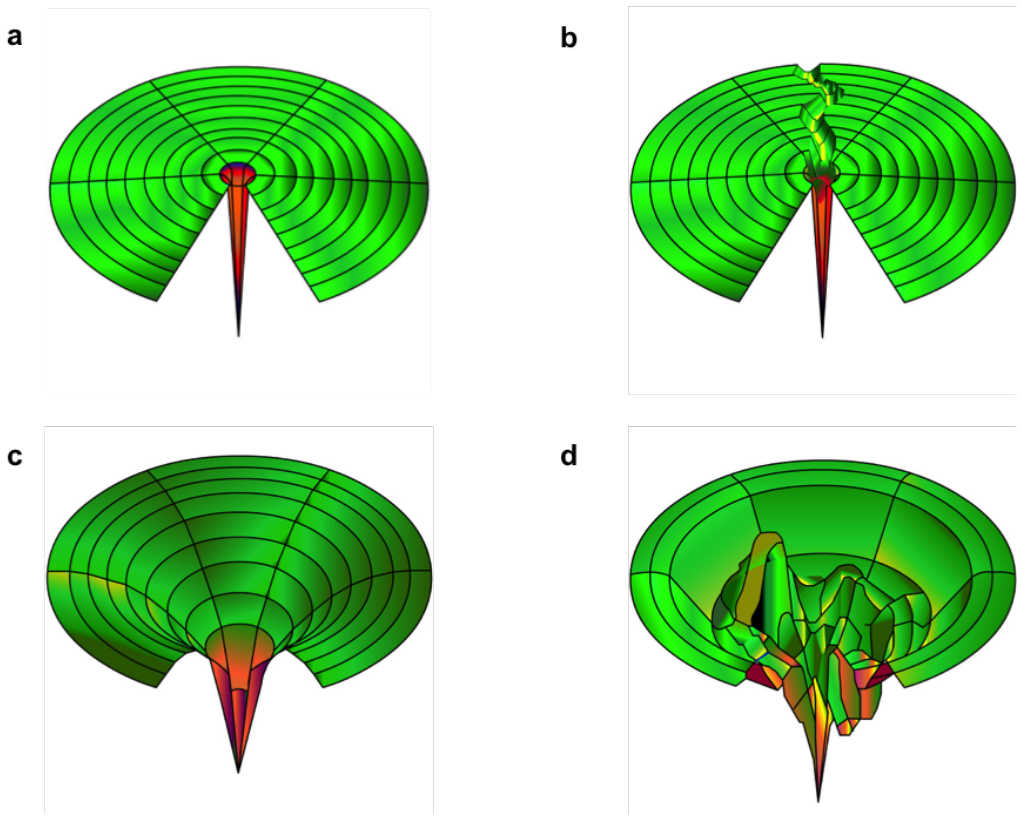


Figure 2.4. Protein folding energy landscapes. **a**. Levinthal's folding "golf court" with one global minimum, i.e., the native state, in an otherwise flat surface of energetically degenerate random-coil structures. **b**. Energy landscape with a single unambiguous folding pathway with *N* defined intermediate states between unfolded and folded conformation. **c**. Perfectly funneled folding landscape of a protein with fully unfrustrated interactions as used in structure-based models. **d**. Realistically rugged folding funnel with energetic roughness as used in energy landscape theory. Redrawn from Chemgapedia.

sequence would be strongly disfavored by natural selection[25]. The principle of minimal frustration states that the energetic frustration in naturally evolving amino acid sequences is as small as possible to achieve robust folding and stable conformations[19,20,25,26]. Accordingly, the energy landscape theory suggests that the most realistic model of a protein is a minimally frustrated heteropolymer with a rugged funnel-shaped potential surface biased towards the native state (see Fig. 2.4 d). The central assumption is that a protein folds through organizing an ensemble of structures rather than through a few defined intermediate states[34]. Folding is described as a diffusive exploration of the energy funnel with an overall drift from higher to lower energies until the protein finally arrives at its native state[19]. In this depiction, energy gains are directly connected to a loss in conformational entropy as, e.g., stabilizing bonds narrow the accessible conformational space (see Fig. 2.5). Mathematically, the funnel shape of a protein's potential surface can be derived from the statistical mechanics of spin glasses[20,26]. With the extreme example of a perfectly funneled downhill landscape (see Fig. 2.4 c), a fully unfrustrated protein can exhaustively be described by its native interactions[23,25]. This simplistic but powerful approach is employed in so-called native structure-based models (see Sec. 4.4), where a distinct funneling of the energy landscape without kinetic traps enables structure formation on short time scales.

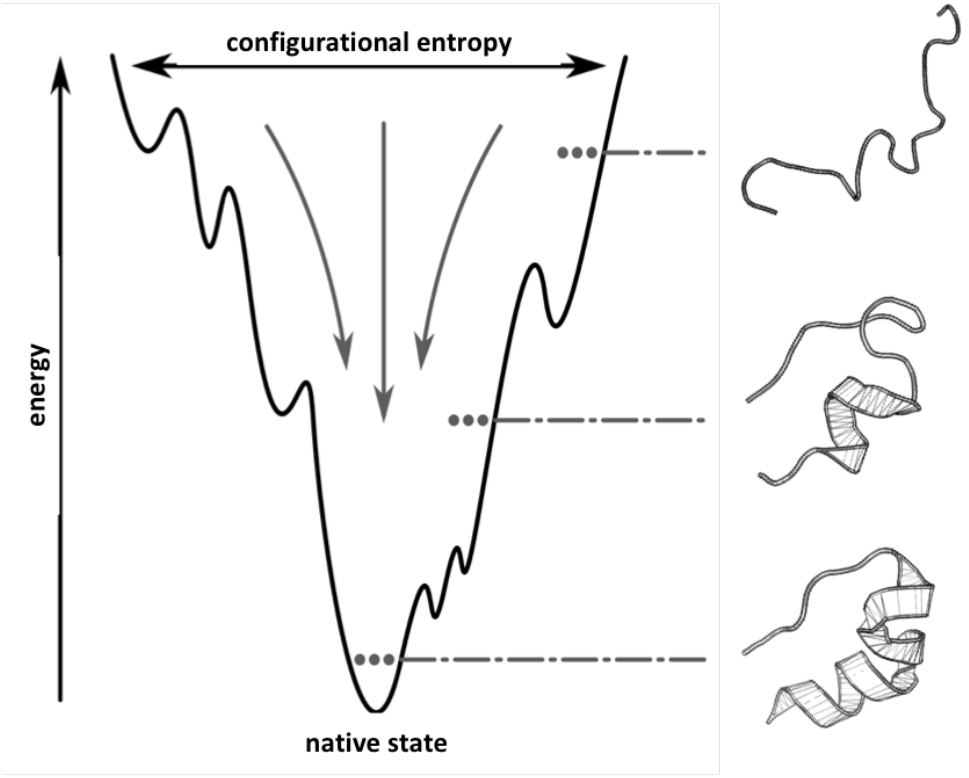


Figure 2.5. Schematic of an energy funnel in protein folding to illustrate the underlying principle of minimal frustration. High-energy unfolded conformations are located at the funnel's upper part. The native state corresponds to the global energetic minimum at the very bottom. With the protein approaching its native fold, the accessible conformational entropy, i.e., the funnel's width, decreases and the system is energetically stabilized by formation of native interactions. Redrawn from Ref.[23]. Protein structures visualized in PyMOL[1].

"I didn't think; I experimented."

Wilhelm Röntgen

3

Experimental Protein Structure Determination

This chapter covers experimental methods for protein structure determination that are relevant in the context of this work. Special focus is put on small-angle X-ray scattering, a practical low-resolution technique for structural analysis of biomolecules which is explained in Sec. 3.1. In Secs. 3.2 to 3.5, I introduce the experimental methods applied in a joint research project by my collaborators Alexander Christiansen and Jochen Reinstein from the Max Planck Institute for Medical Research in Heidelberg and Andreas Winkler from the Institute of Biochemistry at the Graz University of Technology. As explained in Sec. 4.6, we used fluorescence spectroscopy, circular dichroism spectroscopy, Förster resonance energy transfer spectroscopy, and hydrogen-deuterium exchange coupled to mass spectrometry to study the energetics and dynamics of a viral capsid protein in vitro.

Experimental structure determination is the procedure by which the 3D atomic coordinates of a protein are solved with an analytical technique. Many different techniques exist. Most common are X-ray crystallography, NMR spectroscopy, electron microscopy, and small-angle solution scattering. Naturally, each method has its advantages and limitations. X-ray crystallography provides (near) atomic resolution and has made the largest contribution to our understanding of protein structure. Protein crystallization, however, is often arduous and difficult and may produce artifacts in the molecular structure. Although NMR spectroscopy cannot provide such a high resolution, it has proven to be a valuable tool when crystallization is impossible or protein dynamics need to be studied. NMR requires the protein to be stable at room temperature for a comparably long time of data acquisition. The fast relaxing magnetization in large proteins thus limits NMR to smaller systems. Cryogenic electron microscopy (cryo EM) can be applied to study relatively large systems, such as protein complexes or even

cellular organelles. In comparison to crystallography and NMR, it requires less material and has a lower resolution. Solution scattering techniques, like small-angle X-ray scattering (SAXS) and small-angle neutron scattering (SANS), also provide structural information of limited resolution and are becoming increasingly popular. SAXS is easy to apply and suited for a wide variety of systems. Like in NMR, the measurements are performed in solution, which allows direct control of the experimental conditions. While NMR is only applicable to small proteins, SAXS has practically no size limitations[36]. In contrast to X-ray crystallography, which determines a single static structure only, SAXS provides information about steady-state structure as well as about kinetics on fast time scales[10,37]. The data can be used to create 3D low-resolution models or to fit high-resolution structures of separate domains into the SAXS envelope. Other methods for obtaining (local) structural information include mass spectrometry and various spectroscopic methods, such as fluorescence spectroscopy, circular dichroism spectroscopy, and Förster resonance energy transfer spectroscopy.

3.1 Small-Angle X-Ray Scattering

"X-rays will prove to be a hoax."

LORD KELVIN

My first main project was integrating SAXS data into structure-based protein simulations to obtain molecular models (see Sec. 5.2). Such scattering-guided simulations are an efficient tool to interpret the experimentally measured 1D data in terms of 3D structures. To provide the required experimental background knowledge, I thoroughly introduce biological SAXS below.

Small-angle X-ray scattering is an efficient tool for investigating nanostructures in matter. It is particularly suited for low-resolution characterization of disordered systems such as biomolecules in dilute solution[9,10,38]. Typically, the integrated intensity from elastic scattering of monochromatic X-rays by dissolved molecules is measured for small scattering angles (see Fig. 3.1). This yields information on the molecules' size, shape, and characteristic distances[9]. SAXS records the averaged solution scattering intensity over the entire conformational ensemble and all possible orientations of the dissolved molecules. To obtain the pure solute scattering, the intensity of the solvent is subtracted from that of the solution[9]. The net solute scattering intensity is related to the electron density difference between solute and solvent. Ideally, it is proportional to the spherically averaged scattering intensity from a single solute particle, I. Modeling a molecule as a collection of elementary scatterers, e.g., atoms or amino acids, I can be calculated via the Debye equation,[39]

$$I(q) = \sum_{i,j} f_i(q)\, f_j(q)\, \frac{\sin(q r_{ij})}{q r_{ij}}\,. \tag{3.1}$$

r_{ij} is the distance between two scatterers i and j with respective form factors f_i and f_j. $q = 4\pi \sin\theta/\lambda$ is the momentum transfer, where λ is the X-ray wavelength and 2θ the scattering angle. Each distance contributes a sinc-like term. Large distances correspond to high spatial frequencies. They contribute at small q, i.e., low resolution, where the solute's average molecular shape can be extracted. Short distances correspond to low frequencies and dominate the pattern at high q. Different parts of a SAXS intensity pattern thus contain information about different structural features of the solute molecules. An exemplary intensity pattern is shown in Fig. 3.2 along with q regions and their structural information contents. The Debye equation's derivation is presented in Appendix A.1.

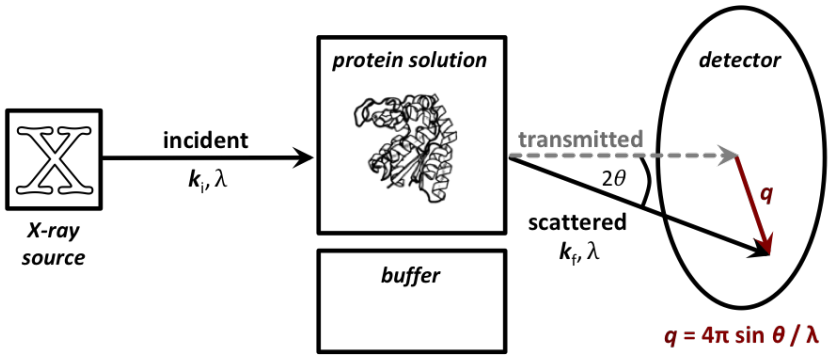


Figure 3.1. Experimental SAXS setup. The monochromatic X-ray beam illuminates a dilute solution of proteins, from which some of the X-ray photons scatter whereas the majority just passes through. Structural analyses use X-rays with an energy of approximately 10 keV[9]. The scattered intensity in the angular range between 0.1 and 5° is recorded with a flat X-ray detector placed behind the sample perpendicularly to the primary beam axis[10,36]. As biomolecules in dilute solution typically scatter weakly, biological SAXS is often conducted at high-intensity synchrotrons. Smaller laboratory systems also exist[9].

The signal-to-noise ratio of experimentally measured intensities decreases rapidly with increasing momentum transfer q. For small q, the intensity can be described by the Guinier approximation[40],

$$\lim_{q\to 0} I(q) = I(0) \exp\left[-\frac{q^2 R_g^2}{3}\right] . \tag{3.2}$$

R_g is the solute's radius of gyration, a measure of overall molecular size, which can be extracted from the curve slope in a logarithmic Guinier plot. The Guinier approximation is only valid for $qR_g < 1.3$ for globular proteins[38] and in an even smaller range for elongated structures.
Dividing out the decay of the scattering intensity in a so-called Kratky plot, $q^2 I(q)$ versus q, yields a tool to qualitatively assess the flexibility and/or degree of unfolding in a protein sample. It serves as an indicator of the protein's conformation, where a distinct bell-shaped peak points to compact, globular structures, or the folded ensemble, and a plateau at high q to highly flexible structures, or the unfolded ensemble. A combination of bell shape and plateau or a slowly decreasing plateau typically points to partially unfolded structures.

Solvent Contribution

SAXS is a contrast method, where the scattering signal is generated from a difference in the average electron densities of solute and solvent[36]. This is why solvent contributions to the scattering intensity in principle need to be taken into account. The net solute scattering depends on the excess electron density, or contrast, of solution and solvent, $\Delta\rho(\mathbf{r})$. For a single dissolved molecule, scattering amplitude and intensity read

$$\begin{aligned} A(\mathbf{q}) &= f(q)\, \mathfrak{F}[\Delta\rho(\mathbf{r})] = f(q) \int_V \Delta\rho(\mathbf{r}) \exp(i\mathbf{q}\cdot\mathbf{r})\, d\mathbf{r}\,, \\ I(\mathbf{q}) &= A(\mathbf{q}) \cdot A^*(\mathbf{q}) = |f(q)|^2 \iint_V \Delta\rho(\mathbf{r})\, \Delta\rho(\mathbf{r}') \exp\left(i\mathbf{q}\left(\mathbf{r}-\mathbf{r}'\right)\right) d\mathbf{r} d\mathbf{r}'\,, \end{aligned} \tag{3.3}$$

where $f(q)$ is the form factor. $I(\mathbf{q})$ is proportional to the Fourier transform of the electron density's correlation function, which is a measure of the probability of finding a scatterer at position $\mathbf{r}$ with another scatterer at position $\mathbf{r}' = \mathbf{0}$. The excess scattering of the solution with respect to the solvent thus provides information about fluctuations of the electron density and spatial correlations in the sample.

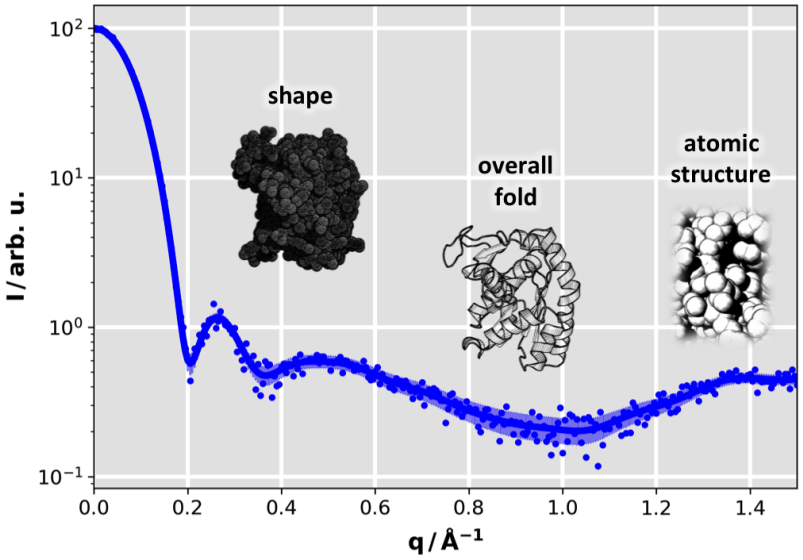


Figure 3.2. Typical X-ray solution scattering curve. Whereas low-resolution information about the particle's shape is encoded in the small-angle regime, i.e., at low q, details on the fold or even atomic structure are contained at wider angles.

In principle, the intensity has to be corrected for displaced-solvent effects. This can be done approximately by, e.g., suitably adapted form factors. Furthermore, the solvent density in the molecular solvation shell generally differs from the bulk value, which results in additional scattering terms[18]. A biomolecule in solution can be modeled as a particle with electron density $\rho(\mathbf{r})$ surrounded by a solvent with an average scattering density ρ_0. The solvation shell is represented by a border layer of thickness Δ and density ρ_b that may differ from ρ_0. The integrated SAXS intensity from such particles in dilute solution is proportional to the averaged scattering of a single particle. Considering displaced solvent and solvation shell yields[41]

$$I(q) = \left\langle \left| A(\mathbf{q}) - \rho_0 A_\mathrm{ex}(\mathbf{q}) + \delta\rho A_b(\mathbf{q}) \right|^2 \right\rangle_\Omega \tag{3.4}$$

with the scattering amplitude of the particle in vacuum, $A(\mathbf{q})$, the scattering amplitude from excluded volume, $A_\mathrm{ex}(\mathbf{q})$, and the scattering amplitude from the solvation shell, $A_\mathrm{b}(\mathbf{q})$, with $\delta\rho = \rho_\mathrm{b} - \rho_0$. $\langle\cdot\rangle_\Omega$ represents the average over all particle orientations with the solid angle in reciprocal space $\mathbf{q} = (q, \Omega)$. I refer the interested reader to Ref.[41] for a detailed explanation.

To sum up, $I(q)$ is proportional to the spatially averaged single-particle scattering for identical non-interacting particles as in ideal disordered systems such as monodisperse dilute solutions of purified biomolecules[9]. Information about the solute molecules' overall shape and internal structures can be extracted at a resolution of 50 Å down to 10 Å[38]. The number of independent data points in a SAXS curve is equal to the number of independent Shannon channels[38,42]. As a measured scattering intensity typically contains only tens of such points, its information content is insufficient to determine all degrees of freedom in a 3D molecular structure. The spatial averaging in SAXS due to the random orientations of the solute molecules results in an inherent loss of information. Various approaches have been developed to analyze 1D or 2D scattering data in terms of 3D models, with one of the most promising being scattering-data assisted molecular dynamics simulations. This concept is introduced in Chapter 5.

The experimental methods introduced below were used in a joint research project together with Alexander Christiansen and Jochen Reinstein from the Max Planck Institute for Medical Research in Heidelberg and Andreas Winkler from the Institute of Biochemistry at the Graz University of Technology to investigate the hexavalent capsomer of the Mavirus virophage (see Sec. 4.6).

3.2 Fluorescence Spectroscopy

Fluorescence spectroscopy is a type of electromagnetic spectroscopy and a powerful method to study biomolecular systems by analyzing their fluorescence. Fluorescence is the emission of light by a substance that has absorbed light or other electromagnetic radiation. In general, all systems have an electronic ground state with minimal energy and excited electronic states with higher energy. Electrons in certain molecular parts are excited by absorbing electromagnetic radiation, typically ultraviolet (UV) light, which causes them to emit photons, typically in the visible spectrum, while lowering their energy down to the ground state. Fluorescence spectroscopy analyzes the frequencies and relative intensities of the emitted light to, e.g., determine the structure of the system's energetic levels.

Biophysical studies of macromolecules use fluorescent molecules, or so-called fluorophores, as physical markers. Fluorophores can be either extrinsic, e.g., radioactive probes or dyes, or intrinsic, such as specific amino acids in proteins. Three proteinogenic amino acids are intrinsically fluorescent, i.e., phenylalanine, tyrosine, and tryptophan (Trp). A folded protein's fluorescence is a mixture of these aromatic amino acids' fluorescence, where Trp is dominating. Fluorescent amino acids are rare and a protein may have just one or a few Trp residues, which greatly facilitates the interpretation of the measured spectra. The intrinsic fluorescence of a protein is highly sensitive to the local Trp environment, in particular to its polarity. For example, denaturing a protein containing a single Trp in its hydrophobic core yields a red-shifted emission spectrum due to the exposure to an aqueous environment as opposed to a hydrophobic protein interior. Furthermore, the proximity of other residues strongly influences Trp fluorescence, e.g., spatially close protonated groups such as aspartic or glutamic acid may cause quenching. Energy transfer between Trp and other fluorescent amino acids is also possible. As Trp emission spectra often change in response to conformational transitions, subunit association, substrate binding, or denaturation, intrinsic fluorescence can be used as a diagnostic of a protein's conformational state without influencing the protein itself.

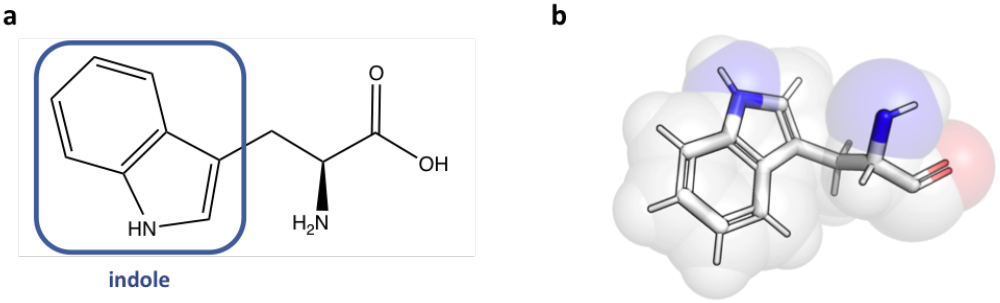


Figure 3.3. Tryptophan as the dominant intrinsic fluorophore in proteins. Trp is a non-polar aromatic amino acid with a side chain indole, which consists of a benzene six-ring fused to a pyrrole five-ring. **a**. Skeletal formula visualized in ChemDraw. **b**. Stick representation with overlaid transparent space-filling representation visualized in PyMOL[1].

3.3 Förster Resonance Energy Transfer

Förster resonance energy transfer (FRET) [43] is a non-radiative energy transfer between two light-sensitive molecules, where an electronically excited donor transfers energy to an acceptor via a non-radiative dipole-dipole coupling. The transfer efficiency depends on the sixth power of the donor-acceptor distance. That is why FRET is extremely sensitive to nanometer distance changes and also known as a "spectroscopic ruler" [44]. By labeling particular protein residues with appropriate dyes, different conformations become distinguishable (see Fig. 3.4). Structural transitions can be observed based on changes of the inter-dye distance. In this way, FRET provides access to time-resolved distance information on, e.g., folding dynamics [32], intermediate structures [45,46], and function-related conformational transitions [47].

Experimentally, the FRET efficiency E is measured. E is the quantum yield of the FRET transition, that is the occurrence probability of a FRET event per donor excitation event. It depends on the donor-acceptor distance R_{DA} (typically 1 to 10 nm) as [48]:

$$E = \left(1 + \left(\frac{R_{\mathrm{DA}}}{R_0}\right)^6\right)^{-1} \tag{3.5}$$

The Förster radius R_0 is the donor-acceptor distance with $E = 0.5$. It depends on (i) the spectral overlap of donor emission and acceptor absorption and (ii) the relative orientation of donor emission dipole moment and acceptor absorption dipole moment represented by the dipole orientation factor κ^2, where $R_0^6 \propto \kappa^2$ [48]. Usually, rotational dye diffusion is fast with respect to the lifetime of the excited state, yielding a constant value of $\overline{\kappa^2} = 2/3$ in the "isotropic averaging regime" [48].

3.4 Circular Dichroism

Circular polarization of an electromagnetic wave is a polarization state where the electromagnetic field has a constant magnitude and rotates at a constant rate in a plane orthogonal to the wave's traveling direction (see Fig. 3.5). Circular dichroism (CD) is a difference in the absorption of left-handed and right-handed circularly polarized light. Because of their dextrorotary and levorotary components, asymmetric biomolecules may absorb these lights to different extents and also have distinct refractive indices. This results in a rotation of the light wave's plane. CD spectroscopy is suited to study protein structure and folding [49] and measured in or near their molecular absorption bands. While the near-UV spectrum (> 250 nm) provides information about a protein's tertiary structure, the far-UV spectrum can reveal information about its secondary structure. Secondary structural elements such as alpha helices generate a distinct CD and have characteristic spectral signatures. Based on this, the molecular fractions in the alpha-helix conformation, the beta-sheet conformation, or the beta-turn conformation [49,50] can be estimated, which places informative constraints on a protein's secondary structure. In contrast to near-UV CD, the far-UV CD spectrum can be assigned to particular parts of a 3D molecular structure. It however is impossible to infer where the detected alpha helices are located in the protein or to predict how many there are. Although providing less specific information than X-ray crystallography or NMR spectroscopy, CD is a valuable tool for analyzing conformational changes or verifying that a protein is natively folded. It can be used to study how the secondary structure is influenced by various environmental factors, e.g., the temperature or the concentration of denaturants such as guanidinium chloride or urea. As a quick method that does not require large amounts of proteins or extensive data processing, CD can be used to probe proteins in solution under different solvent conditions, at varying temperature, pH, salinity, and in the presence of various cofactors. Furthermore, it can reveal

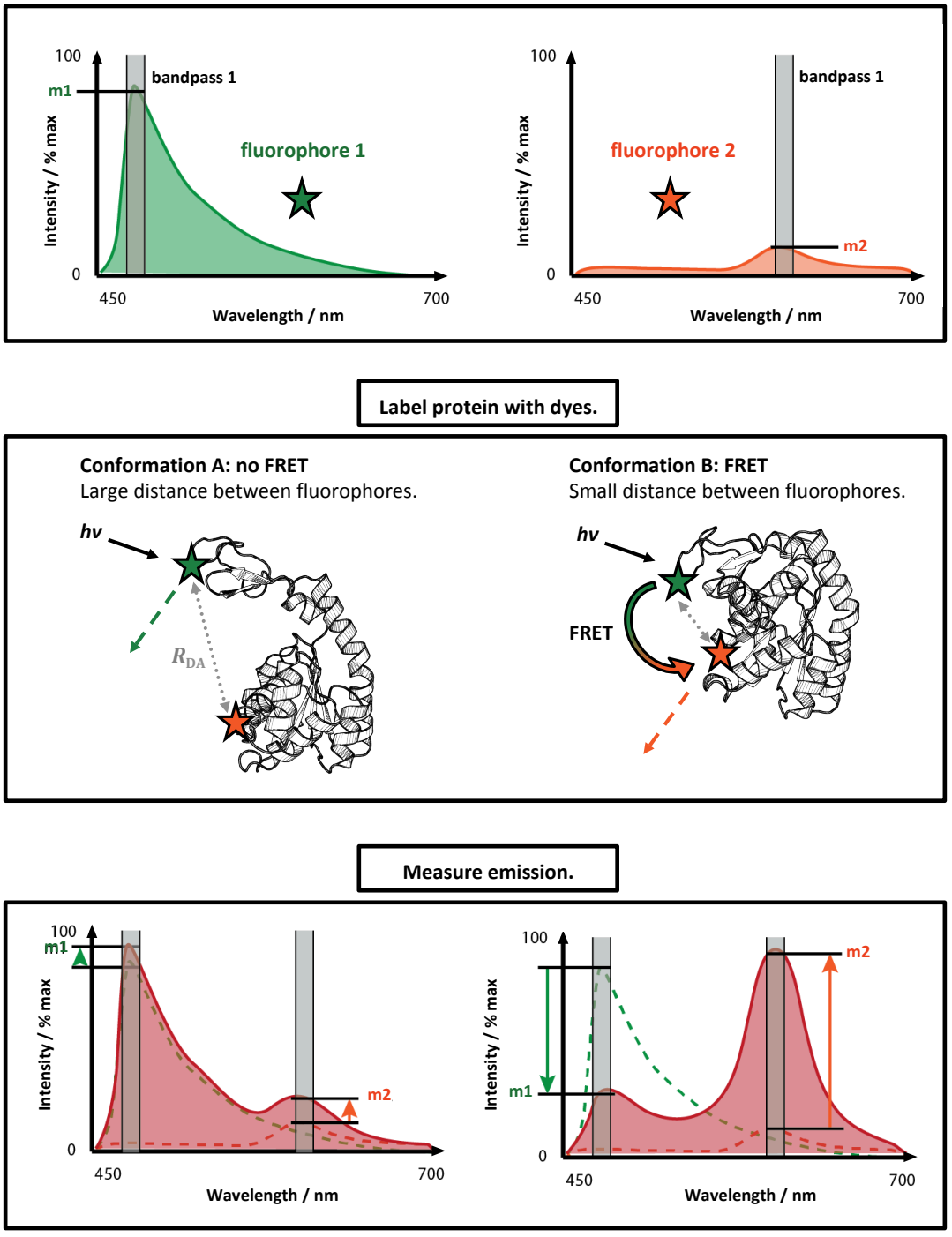


Figure 3.4. The functional principle of FRET spectroscopy. FRET can be used to measure distances within a protein by labeling specific molecular sites with fluorescent dyes, measuring their emission, and determining the distance-dependent energy transfer efficiency. The dyes must have overlapping emission spectra. Derived from "Concept of FRET" by Curtis Neveu, used under CC BY-SA 3.0, via Wikimedia Commons. Licensed under CC BY-SA 3.0 by Marie Weiel-Potyagaylo.

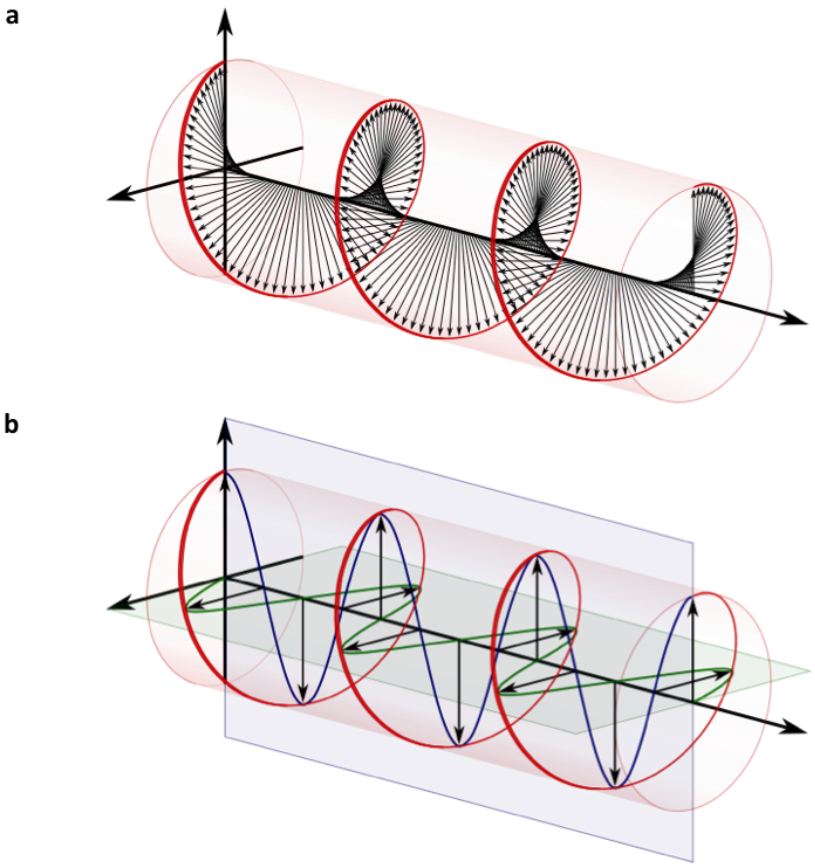


Figure 3.5. Circularly polarized light. Right-handed circularly polarized light (as defined from the receiver's perspective) illustrated without and with the use of components. **a**. The helix traced out by the light wave's electric field vector forms a right-handed screw. While its direction steadily changes in a circling manner, the electric field has a constant magnitude along its direction of propagation. The related magnetic field orthogonal to the electric field is proportional in its magnitude and not illustrated. Created by Dave3457, public domain. Retrieved from Wikimedia Commons. **b**. The phenomenon of polarization arises from the fact that light can be modeled as a 2D transverse wave. The two orthogonal electric-field component vectors of circularly polarized light are of equal magnitude and out of phase by exactly $\pi/2$. The horizontal and vertical components are illustrated in green and blue, respectively. The relative to the direction of propagation rightward horizontal component leads the vertical component by one-quarter wavelength. Created by Dave3457, public domain. Retrieved from Wikimedia Commons.

information on thermodynamic quantities that are not readily accessible otherwise, such as a molecule's enthalpy or its Gibbs free energy of denaturation.

3.5 Hydrogen-Deuterium Exchange Coupled to Mass Spectrometry

Hydrogen-deuterium exchange coupled to mass spectrometry (HDX-MS) is a versatile technique to monitor the structure and dynamics of proteins in solution[51]. Natively folded proteins are diluted into heavy water, allowing accessible amide hydrogens in the protein backbone to exchange for deuterium. This exchange can only occur for amides not engaged in hydrogen bonds, and how rapidly an amide exchanges reflects its solvent accessibility. For example, protein unfolding can expose natively buried

amide protons to the solvent and allow for exchange. With deuterium being a heavier isotope of hydrogen, such reactions can be monitored by measuring mass changes associated with the isotopic exchange. The exchange rates act as a proxy for protein structure and stability, with a more stable structure experiencing a slower rate of and hence greater protection from exchange than a less stable structure. Diverse factors can play a role, e.g., secondary structure, molecular contacts, and protein compaction or relaxation. Therefore, while HDX cannot analyze protein structures directly, it can provide localized information about a protein's structural and dynamic properties.

PART II

COMPUTATIONAL METHODS AND METHOD DEVELOPMENT

"I try to identify myself with the atoms. . . I ask what I would do if I were a carbon atom or a sodium atom."

Linus Pauling

4

Molecular Dynamics

This chapter covers the computational methods underlying this work. My simulations are built on molecular dynamics (MD). MD provides a classical-physics based description of molecular motion and gives in-depth insight into biomolecular function. In Sec. 4.4, I introduce so-called structure-based models (SBMs). Minimalist SBMs have been shown to be extremely effective in explaining fundamental questions of protein folding and function. At the same time, they are sufficiently efficient to run complex simulations on small computing systems or even laptops. In Sec. 4.6, I present my simulation study on the Mavirus virophage's capsomer protein as a fitting example of the explanatory power of SBMs. The capsomers self-assemble with high fidelity to a protective protein shell known as the virus' capsid. To better understand this process, I investigated the capsomer's dynamics and stability in structure-based simulations as a complement to experimental studies conducted at the Max Planck Institute for Medical Research in Heidelberg. Finally, I show another practical example of how SBMs can support and explain experimental data from combined applications of SAXS and FRET in Sec. 4.7.

Molecular dynamics (MD) is a computational simulation technique for studying the physical motions of nano- to microscale systems. Atomically resolved trajectories are derived from Newton's equations of motion for a system of interacting particles, thus providing insight into the system's dynamic evolution. The required forces are calculated from empirical interatomic potentials, or force fields. The system's coordinates are computed within successive femtosecond time steps, where the dynamics is determined via numerical integration over time. Assuming ergodicity, macroscopic properties can be extracted from MD simulations by averaging over a representative statistical ensemble in equilibrium or, strictly speaking, the corresponding time interval.

For a system of N atoms with masses m_i at positions $\{\mathbf{r}_i\}$, $i = 1, \ldots, N$, a force field can be expressed via a potential $V(\{\mathbf{r}_i\})$. At each simulated time step, conservative forces $\mathbf{F}$ are calculated as the potential's negative gradient with respect to the atomic positions:

$$\mathbf{F} = -\nabla V(\{\mathbf{r}_i\}) \tag{4.1}$$

The resulting system of Newton's equations of motion

$$m_i \frac{d^2\mathbf{r}_i}{dt^2} = \mathbf{F}_i \tag{4.2}$$

is solved simultaneously. The obtained positions are appended to a trajectory file, comprising individual frames for each discrete point in time.

Before an actual MD simulation, initial atomic positions $\{\mathbf{r}_i\}$ and velocities $\{\mathbf{v}_i\}$ have to be specified. The positions are usually taken from an experimentally resolved molecular structure. If initial velocities are unavailable, they are set randomly according to a Maxwell-Boltzmann distribution $p(v_i)$ at an absolute temperature T:

$$p(v_i) = \sqrt{\frac{m_i}{2\pi k_\mathrm{B} T}} \exp\left(-\frac{m_i v_i^2}{2k_\mathrm{B}T}\right) \tag{4.3}$$

k_B is the Boltzmann constant. In an N-particle system with N_DoF degrees of freedom, T is defined by the system's kinetic energy E_kin [52]:

$$T = \frac{2E_\mathrm{kin}}{N_\mathrm{DoF} k_\mathrm{B}} \quad \text{with} \quad E_\mathrm{kin} = \frac{1}{2}\sum_{i=1}^{N} m_i v_i^2 \tag{4.4}$$

In its simplest form, MD generates a microcanonical NVE ensemble with a fixed number of particles N, a constant volume V, and defined total energy E. In such isolated systems, there is no exchange of energy or particles with the environment. Quantities of interest typically arise from a canonical NVT ensemble with a constant temperature controlled via weak temperature coupling. A direct comparison with experiments requires a constant pressure P as in the isothermal-isobaric NPT ensemble. This is accomplished by additional pressure coupling. Detailed explanations of temperature and pressure coupling can be found in Ref. [53].

Throughout my work, I used the versatile MD engine `GROMACS` [52,54–62] to simulate Newton's equations of motion for biomolecular systems with hundreds to millions of particles. It is primarily designed for biomolecules such as proteins, lipids, and nucleic acids that have many complicated bonded interactions.

It is important to note that MD has some limitations [52]. The simulations are based on Newton's equations of motion, i.e., quantum mechanical behavior is not accounted for. Electrons are assumed to be in their ground state and electronic motions are not considered. The Born-Oppenheimer approximation is applied, i.e., it is assumed that the motion of atomic nuclei and electrons can be treated separately because the nuclei are much heavier than the electrons. Force fields are models and only approximate. Their parameters are derived from quantum mechanical calculations and modified to fit empirical data. Due to limited computational resources, long-range interactions are neglected above some predefined cutoff. To avoid undesirable real phase boundaries of a system with its environment, unnatural periodic boundary conditions are applied. In particular for small systems, this may provoke unphysical behavior.

4.1 Leap-Frog Integrator

An integrator is an algorithm for numerically solving differential equations. In MD, this task consists in calculating a trajectory from forces using Newton's equations of motion evaluated at discrete time steps. The leap-frog integrator is the `GROMACS` default method. It uses positions $\mathbf{r}_i$ at time t, forces $\mathbf{F}_i$ determined by the positions at time t, and velocities $\mathbf{v}_i$ at time $t - \frac{1}{2}\Delta t$ to update positions and velocities via:

$$\begin{aligned} \mathbf{r}_i\left(t + \Delta t\right) &= \mathbf{r}_i\left(t\right) + \Delta t \cdot \mathbf{v}_i\left(t + \frac{1}{2}\Delta t\right) \\ \mathbf{v}_i\left(t + \frac{1}{2}\Delta t\right) &= \mathbf{v}_i\left(t - \frac{1}{2}\Delta t\right) + \frac{\Delta t}{m_i}\mathbf{F}_i\left(t\right) \end{aligned} \tag{4.5}$$

Δt is the time step. The name "leap frog" is due to the fact that positions and velocities are updated at alternate time points.

4.2 Stochastic Dynamics

Realistic molecular systems are unlikely to exist in vacuum. Collisions with solvent or air molecules cause friction and perturb the system. Stochastic dynamics extends MD for these effects by mimicking the viscous aspect of a solvent. However, it does not fully model an implicit solvent as electrostatic screening and the hydrophobic effect are not considered. It uses a simplified model while accounting for the omitted degrees of freedom by incorporating friction and noise into Newton's equations of motion:

$$m_i \frac{d^2\mathbf{r}_i\left(t\right)}{dt^2} = \mathbf{F}_i\left(\mathbf{r}_i\left(t\right)\right) - \gamma_i m_i \frac{d\mathbf{r}_i\left(t\right)}{dt} + \mathbf{R}_i\left(t\right) \tag{4.6}$$

γ_i is a friction constant with $[\gamma_i] = \mathrm{ps}^{-1}$. $\mathbf{R}_i\left(t\right)$ is a noise process with

$$\left\langle \mathbf{R}_i\left(t\right)\, \mathbf{R}_j\left(t + \Delta t\right)\right\rangle = 2 m_i \gamma_i\, k_\mathrm{B} T\, \delta_{ij} \delta\left(\Delta t\right). \tag{4.7}$$

δ_{ij} and $\delta(\Delta t)$ are Kronecker and Dirac delta, respectively. The random force $\mathbf{R}_i\left(t\right)$ is modeled by a stationary Gaussian noise with zero mean. With $1/\gamma_i$ being much greater than the system's time scale, stochastic dynamics can be considered as MD with stochastic temperature coupling. The noise term inherently controls the system's temperature, thus approximating a canonical ensemble. In fact, γ_i is fixed by the temperature coupling time constant $\tau_T = m_i/\gamma_i$. For integration of Eq. 4.6, a third-order leap-frog algorithm is used[63].

4.3 Force Fields

"A protein is a set of coordinates."

A.P. Heiner

A force field is a computational model to estimate the forces between atoms in molecular systems. It is a set of equations defining interatomic potential energies along with atom-dependent parameters. These

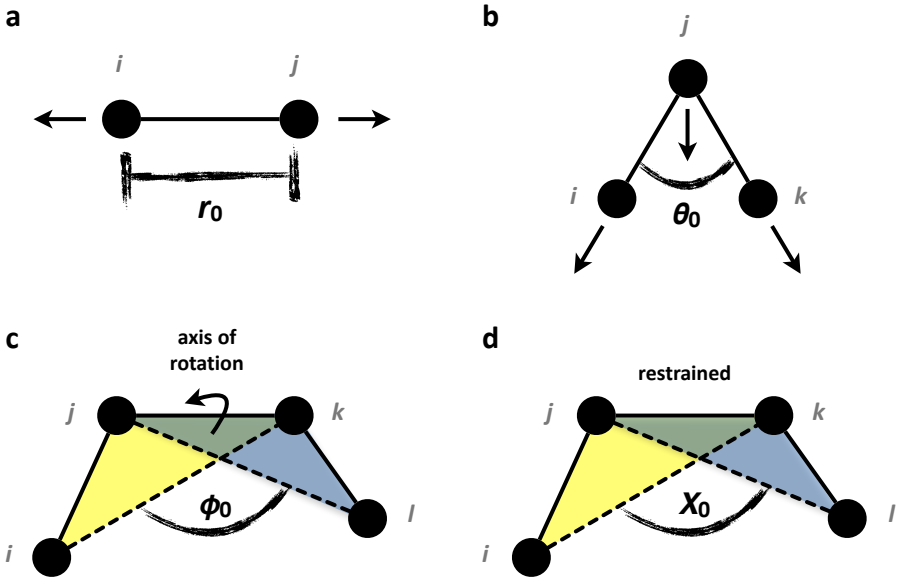


Figure 4.1. Bonded interactions. Atoms are depicted as black dots. **a**. Linear bond between two atoms with native distance r_0. **b**. Angle defined by three atoms with native value θ_0. **c**. Proper dihedral angle with native value ϕ_0 between planes defined by atoms (i, j, k) and (j, k, l). **d**. Improper dihedral angle with native value χ_0 between planes defined by atoms (i, j, k) and (j, k, l), e.g., to restrain atoms to a particular plane in aromatic rings.

parameters are derived from experiments in physics and chemistry and/or quantum mechanical calculations. Molecular force fields use the same concept as classical-physics force fields, with the difference that their parameters characterize the biomolecular energy landscape. The acting forces on every atom are computed as the potential's gradient with respect to the atom's positions[64].

The functional form of any MD potential comprises bonded terms for interactions of covalently bonded atoms and non-bonded terms modeling long-range electrostatic and van-der-Waals interactions:

$$\begin{aligned} V_{\text{total}} &= V_{\text{bonded}} + V_{\text{non-bonded}} \\ V_{\text{bonded}} &= V_{\text{bond}} + V_{\text{angle}} + V_{\text{dihedral}} \\ V_{\text{non-bonded}} &= V_{\text{electrostatic}} + V_{\text{van-der-Waals}} \end{aligned} \tag{4.8}$$

Bonded interactions include bond-stretching (2-body), bond-angle (3-body), and dihedral-angle (4-body) potentials (see Fig. 4.1). They are modeled by harmonic oscillators and do not allow bond breaking. With native bond length r^0_{ij}, native angle θ^0_{ijk}, and native dihedral angle ϕ^0_{ijkl} between two, three, and four atoms (i, j), (i, j, k), and (i, j, k, l), respectively, the different contributions read:

$$\begin{aligned} V_{\text{bond}}\left(r_{ij}\right) &= \frac{1}{2} k^{\text{b}}_{ij} (r_{ij} - r^0_{ij})^2 \\ V_{\text{angle}}\left(\theta_{ijk}\right) &= \frac{1}{2} k^{\text{a}}_{ijk} \left(\theta_{ijk} - \theta^0_{ijk}\right)^2 \\ V_{\text{dihedral}}(\phi_{ijkl}) &= \sum_n k^{\text{d}}_{ijkl} \left(1 + \cos\left(n\left(\phi_{ijkl} - \phi^0_{ijkl}\right)\right)\right) \end{aligned} \tag{4.9}$$

The dihedral potential V_{dihedral} is a generic periodic potential with multiplicity n. The force constants k^{b}_{ij}, k^{a}_{ijk}, and k^{d}_{ijkl} are parameters provided by the force field. Additionally, improper dihedral interactions are introduced to force atoms into a particular plane, e.g., in aromatic rings, or prevent undesirable chirality transitions. A harmonic potential with force constant k^{id}_{ijkl} and ground-state angle χ^0_{ijkl} is

included for improper dihedral angles χ_{ijkl} between four atoms (i, j, k, l):

$$V_{\text{impr.-dihedr.}}\left(\chi_{ijkl}\right) = \frac{1}{2} k_{ijkl}^{\text{id}} \left(\chi_{ijkl} - \chi_{ijkl}^{0}\right)^2 \tag{4.10}$$

The two types of dihedral angles are illustrated in Figs. 4.1 c and d.

Non-bonded terms affect atoms of greater sequence distance and are computationally most intensive. Very often, the interactions are limited to pairwise energies within predefined cutoffs. They cover a Lennard-Jones term

$$V_{\text{van-der-Waals}}\left(r_{ij}\right) = k_{\text{LJ}} \left[\left(\frac{\sigma_{ij}^{0}}{r_{ij}} \right)^{12} - 2 \cdot \left(\frac{\sigma_{ij}^{0}}{r_{ij}} \right)^{6} \right] \tag{4.11}$$

for the van-der-Waals interaction, where σ_{ij}^{0} is the excluded-volume radius, and a Coulomb term

$$V_{\text{electrostatic}}\left(r_{ij}\right) = \frac{q_i q_j}{4\pi\varepsilon_0\varepsilon_r r_{ij}} \tag{4.12}$$

for the electrostatic interaction. r_{ij} is the distance between two atoms i and j with charges q_i and q_j, ε_0 is the electric permittivity, and ε_r the dielectric constant.

Classical MD simulations consider an enormous number of atoms, including the protein's atoms and those of the solvent in an appropriately sized simulation box. For example, the system of adenylate kinase with 214 residues and 3341 atoms dissolved in a box of water comprises several ten thousand atoms. Simulating biologically relevant micro- to millisecond time scales with a femtosecond time step requires 10^9 to 10^{12} integration steps. Analyzing functionally relevant large-scale motions thus involves excessive computational costs. Numerous ambitions to minimize the computational demands by reducing the simulated systems' number of degrees of freedom exist. In my work, special focus is laid on minimal model representations by coarse-graining the proteins' physical interactions in the force field.

4.4 Structure-Based Models

"Nothing fucks you harder than time."

Ser Davos, the Onion Knight

An interesting alternative to MD using classical molecular-mechanics force fields is focusing on a system's essential characteristics by appropriately coarse-graining its resolution. Coarse-graining makes extended time and length scales accessible. One approach is coarsening a system's granularity by systematically merging groups of atoms into single beads. In C_α models, for example, one bead is introduced for each amino acid at its C_α position. Another option to improve sampling while decreasing computational demands is reducing the force field's complexity in so-called all-atom structure-based models (SBMs).

According to the energy landscape theory[24,25] (see Sec. 2.2), proteins have an evolutionarily smoothened free energy funnel biased towards their native state. Robust structure formation is induced by minimally frustrated native interactions, giving rise to the typical funnel shape[25,26]. SBMs provide a minimalist description of the dynamics arising from such funneled energy landscapes. While SBMs use the same bonded interactions as classical molecular mechanics, the crucial difference lies in the non-bonded interactions categorized as either native, and thus favorable, or non-native, and thus unfavorable. In the first place, long-range interactions between remote amino acids are governed by a protein's geometry[21,23].

Native interactions are assumed to be generally stabilizing whereas non-native interactions are only included to preserve excluded volume. The structural information of the native fold thus becomes a defining component and the structure-based potential explicitly represents the protein's geometry. The force field's complexity can be drastically decreased without loss of information about the system's key characteristics. In particular, SBMs enable thorough sampling of large conformational ensembles such as intrinsically disordered or unfolded systems. Conformational transitions involving multiple stable states can be modeled with so-called multi-Gō models[65]. Combining high efficiency and low computational costs with full protein flexibility, SBM simulations have shown to be in excellent accordance with experiments[11,29,66–68]. Successful applications cover a wide range of protein dynamics such as folding pathways[66,69–73] and kinetics[68]. SBMs are also employed for structure prediction[74–77], integrative structural modeling of experimental data from, e.g., SAXS[27] or cryo-EM[13], and investigation of transition state ensembles[65,78].

The Structure-Based Energy Function

The SBM's key component is its contact potential. Native contacts are defined by pair interactions between spatially close atoms of amino acids i and j in the native structure, where $i > j + 3$. Each contact is assigned a Lennard-Jones like potential with an attractive and a repulsive term. Other non-local interactions are purely repulsive to account for excluded-volume effects[21]. With native bond lengths r_0, bond angles ϑ_0, and proper and improper dihedral angles ϕ_0 and χ_0, the all-atom structure-based potential reads[79]:

$$\begin{aligned} V = & \sum_{\text{bonds}} K_\text{b} \left(r - r_0\right)^2 + \sum_{\text{angles}} K_\text{a} \left(\vartheta - \vartheta_0\right)^2 + \sum_{\substack{\text{improper}\\\text{dihedrals}}} K_\text{i} \left(\chi - \chi_0\right)^2 \\ & + \sum_{\text{dihedrals}} K_\text{d} \left[\left(1 - \cos\left(\phi - \phi_0\right)\right) + \frac{1}{2}\left(1 - \cos\left(3\left(\phi - \phi_0\right)\right)\right)\right] \\ & + \sum_{\text{contacts}} K_\text{c} \left[\left(\frac{\sigma_{ij}^0}{r_{ij}}\right)^{12} - 2 \cdot \left(\frac{\sigma_{ij}^0}{r_{ij}}\right)^{6}\right] \\ & + \sum_{\substack{\text{non-native}\\\text{contacts}}} K_\text{nc} \left(\frac{\tilde{\sigma}}{r_{ij}}\right)^{12} \end{aligned} \tag{4.13}$$

The energetic weights of bonds, angles, improper dihedral angles, and non-native contacts are $K_\text{b} = 20\,000\,\varepsilon/\text{nm}^2$, $K_\text{a} = 40\,\varepsilon/\text{deg}$, $K_\text{i} = 40\,\varepsilon/\text{deg}$, and $K_\text{nc} = 0.01\,\varepsilon$[80]. ε is the SBM's reduced energy unit and deg is arc degree. σ_{ij}^0 is the native distance of atom pair (i, j) in contact, r_{ij} the distance between atoms i and j, and $\tilde{\sigma} = 2.5\,\text{Å}$ the excluded volume for Pauli repulsion. As in classical molecular-mechanics force fields, bond stretching, bond angle, and improper dihedral angle potentials are modeled by harmonic oscillators. The dihedral potential illustrated in Fig. 4.2 allows for the occupation of isomeric conformations next to the native state. In Eq. 4.13, the system is stabilized by the energies for proper dihedral angles, E_d, and native contacts, E_c[80]. By construction, the overall stabilizing energy E_s is set to the system's number of heavy atoms N[80]:

$$E_\text{s} = \sum E_\text{c} + \sum E_\text{d} = N \tag{4.14}$$

Contact weight and proper dihedral weight are scaled so that the respective energies fulfill[80]

$$R_\text{c/d} = \frac{\sum E_\text{c}}{\sum E_\text{d}} = 2\,. \tag{4.15}$$

Dihedral angles with mutual middle atoms are grouped to prevent multiple counting. With N_d group

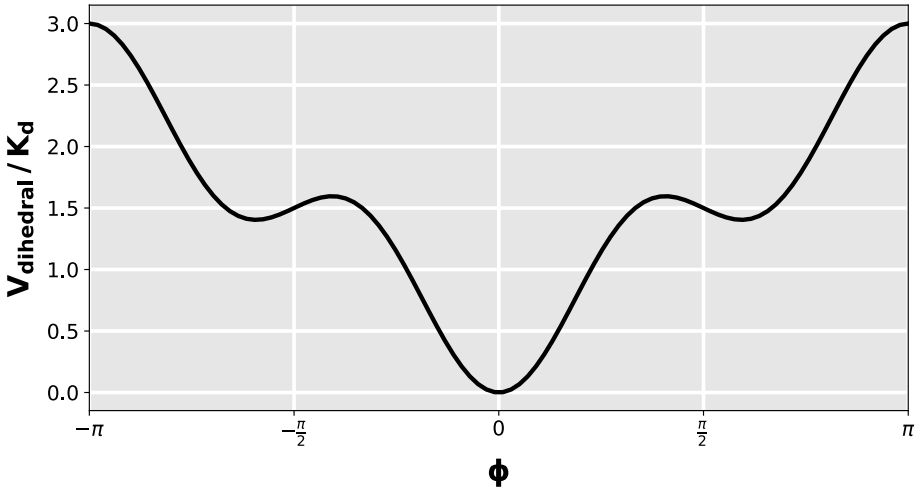


Figure 4.2. Symmetric dihedral potential used in structure-based models. The two local minima next to the native angle $\phi_0 = 0$ correspond to isomeric conformations.

members, which is also known as the angle's multiplicity, each dihedral angle is assigned an interaction strength of $1/N_\mathrm{d}$. Combining Eqs. 4.14 and 4.15 yields the dihedral and contact weight

$$\begin{aligned} K_\mathrm{d} &= \frac{N}{1+R_\mathrm{c/d}} \cdot \frac{1}{N_\mathrm{d}} \\ K_\mathrm{c} &= \frac{N}{N_\mathrm{c}} \cdot \frac{R_\mathrm{c/d}}{1+R_\mathrm{c/d}}. \end{aligned} \tag{4.16}$$

with the number of contacts N_c. Throughout my work, I used all-atom SBMs, which explicitly include all non-hydrogen atoms as unit beads with equal masses, radii, and force constants[79]. Solvent and hydrogen atoms are only treated implicitly[70,79].

Determination of Native Contacts

The structure-based potential strongly depends on contacts formed in the native state. This information is collected in the form of a contact map (see Fig. 4.3), a list of atom pairs which are natively in contact. A contact map is a reduced representation of the protein's 3D structure in form of a 2D binary matrix. It can be set up using either the cutoff or the shadow algorithm.

Cutoff algorithm. The cutoff algorithm defines native contacts between all pairs of atoms with a native distance less than a predefined cutoff C. For two atoms i and j, the element (i, j) is 1 if the atoms are closer than C, and 0 otherwise.

Shadow algorithm. The more elaborate shadow algorithm applies a heavy-atom cutoff distance in combination with geometric occlusion[21] (see Fig. 4.4). All atoms within a cutoff radius C around an atom i are considered to identify respective contacts and assigned a screening radius S. A virtual point-shaped light source is placed at the position of atom i. Atoms within the cutoff being shadowed are neglected, the remaining ones are defined to be in contact with atom i. If this criterion is satisfied vice versa, the atom pair is eventually declared a contact. To avoid mutual shadowing of overlaying bonded atoms, these atoms have a reduced screening radius S_neighbor. Contacts are only valid for atoms separated by at least three residues in the sequence to prevent forces conflicting with bonded interactions.

Table 4.1. MD units in GROMACS[61]. Bracketed digits give accuracy.

Symbol	Denotation	Unit
r	length	nm $= 10^{-9}$ m
m	mass	u (atomic mass unit) $= 1.660\,539\,040(20) \cdot 10^{-27}$ kg
t	time	ps $= 10^{-12}$ s
q	charge	e (electronic charge) $= 1.602\,176\,6208(98) \cdot 10^{-19}$ C
T	temperature	K

In this way, the algorithm determines contacts between atoms while inhibiting higher-order occluded interactions through other intermediates.

Units in Structure-Based Models

"It's Because of the Metric System."

PULP FICTION

GROMACS naturally uses a nm length scale, ps time scale, amu mass scale, and kJ/mol energy scale. The units are chosen to largely produce values in the order of 1 for all molecular quantities of interest (see Table 4.1)[61]. In SBMs, however, it is beneficial to use reduced units: length scale, time scale, mass scale, and energy scale are all 1. While the PDB's Å length scale can easily be converted into nm, the mass scale, time scale, and energy scale remain free. It is possible to determine a system-specific overall energy and mass scale from the structure and its dynamics and subsequently infer a time scale. Time scales can also be extracted by comparing simulation results with experimental observations using, e.g., folding rates or rotational correlation times[11]. However, there is no standard method for calculating "real" times from structure-based simulations. Due to the system's inherently accelerated dynamics, the "real" structure-based time unit is always longer than the GROMACS ps time scale[80]. The Boltzmann constant k_B sets an energy scale $\varepsilon = k_B T$ and is reported in reduced GROMACS units. Thus, GROMACS temperatures are specified in $k_B = 0.008\,314\,51$ reduced units. A reduced temperature of 1 is a GROMACS temperature of 120.2717[61]. For a detailed discussion on SBM units, see Ref.[80].

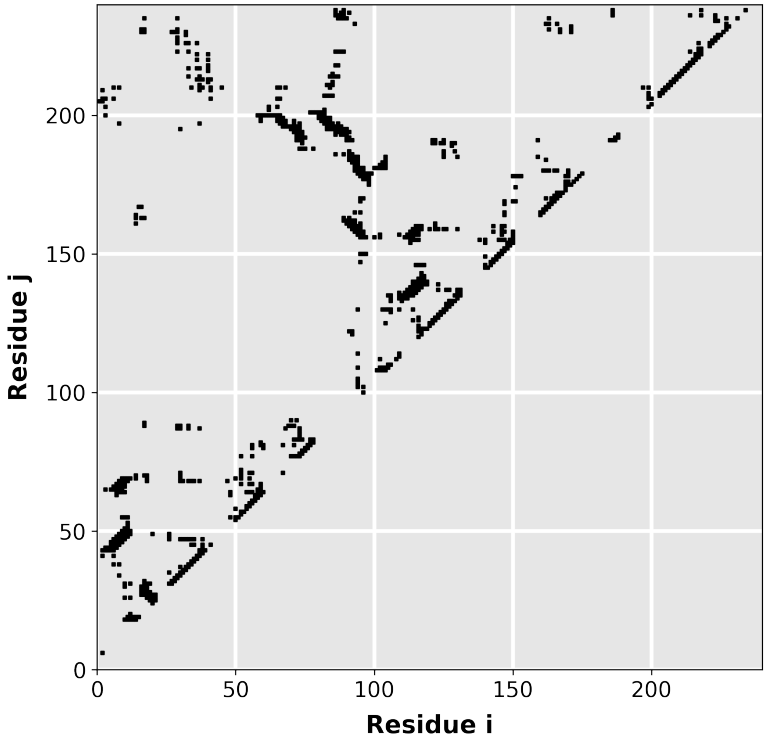


Figure 4.3. Contact map of lysine-, arginine-, ornithine-binding protein's apo conformation (PDB code 2LAO[81]). A contact map is a reduced representation of the protein's 3D structure in form of a 2D binary matrix. For two residues i and j, the element (i, j) is 1 if the residues are closer than a predefined cutoff distance C, and 0 otherwise. Helices correspond to strips directly adjacent to the main diagonal. Parallel beta-sheets correspond to parallels to the diagonal, antiparallel sheets can be identified with cross diagonals.

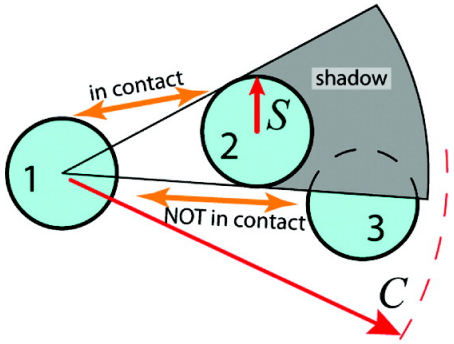


Figure 4.4. The shadow map algorithm's principle for atoms 1, 2, and 3 with cutoff radius C and screening radius S. While 1 and 2 are in contact, 1 and 3 are not as 3 is shadowed by 2. Reprinted with permission from Ref.[21]. Copyright © 2012 American Chemical Society.

4.5 Quantifying Structural Similarity of Proteins

Studying protein dynamics in biomolecular simulations requires quantitative measures of similarity between different conformations of a protein.

Global distance test. I use the global distance test (GDT)[82–84] to quantify differences between two conformations of a protein. The GDT is a more suitable similarity measure than the widely used root-mean-square deviation (RMSD), which is very sensitive to small numbers of locally displaced atoms in an otherwise accurate structure. To estimate how similar two superimposed structures are, the displacement of each C_α atom is compared to various distance cutoffs. Percentages P_x of C_α atoms with displacements below certain cutoffs of x Å are used to calculate the total score

$$\mathrm{GDT} = 0.25 \cdot (P_1 + P_2 + P_4 + P_8) \in [0, 100]\,. \tag{4.17}$$

In simple terms, it can be approximately thought of as the percentage of amino acids in one structure that lie within a threshold distance from their position in another structure. Higher GDT values indicate a stronger similarity between two models. Structures with a GDT greater than 50 are considered topologically accurate, and a score of around 90 in protein structure prediction is informally considered competitive with experimentally obtained results[85].

Root-mean square deviation. The more common RMSD is the minimal mass-weighted average distance between N atoms (usually backbone or C_α) of two superimposed structures over all possible spatial translations and rotations,

$$\mathrm{RMSD} = \min_{\text{trans, rot}} \sqrt{\frac{1}{M} \sum_{i=1}^{N} m_i \left|\left|\mathbf{r}_i - \mathbf{r}_{i,0}\right|\right|^2}\,, \tag{4.18}$$

where $M = \sum_{i=1}^{N} m_i$ with the mass m_i of atom i. $\mathbf{r}_i$ and $\mathbf{r}_{i,0}$ are the positions of atom i in the mobile and reference structure, respectively. Taking the native structure as a reference, the RMSD is low for folded and high for unfolded conformations.

4.6 PROJECT: Simulating the Mavirus Capsomer with Structure-Based Models

"You can observe a lot by just watching."

Yogi Berra

This work was done in collaboration with Alexander Christiansen and Jochen Reinstein from the Max Planck Institute for Medical Research, Heidelberg, and Andreas Winkler from Graz University of Technology. In this joint project, I did the computational simulation part as a complement to my collaborators' experiments. The following section is reproduced from our Journal of Molecular Biology article "The trimeric major capsid protein of Mavirus is stabilized by its interlocked N-termini to enable core flexibility for capsid assembly" (2021)[28].

A virus' genetic code is cocooned by a protective protein shell known as the capsid. The capsid's morphological subunits are called capsomers. In the simplest case, capsids are built from identical

monomeric capsomers. Very often, however, one capsomer consists of two to five different proteins. The oligomeric subunits of a composed capsomer are called protomers. The capsid can also be built from different types of capsomers. Upon a viral infection, the host's immune response depends on the recognition of the capsid's surface. This means that understanding the capsid has direct medical relevance for developing effective treatments for virus-induced diseases, such as the globally pandemic coronavirus disease 2019 caused by the severe acute respiratory syndrome coronavirus 2, SARS-CoV-2.

We studied Mavirus, a genus of double-stranded DNA virus infecting the marine phagotrophic flagellate *Cafeteria roenbergensis*. As a virophage, its replication depends on the replication factory of a co-infecting giant virus[86]. This parasitism often deactivates the giant virus; thus, virophages can improve the host organism's recovery and survival. Mavirus' icosahedral capsid is built from hexavalent capsomers (hexons) at the faces and pentavalent capsomers (pentons) at the vertices[87,88]. According to the biomolecular structure-function paradigm (see Sec. 2.1), the capsid's assembly and thus its structure and size are solely encoded in its capsomeric building blocks. Understanding their energetics and dynamics therefore is crucial for gaining insight into viral capsid assembly. This knowledge could be applied to manipulate capsids in order to develop tailor-made nanocontainers or influence the viral life cycle for therapeutic uses[89,90].

Mavirus is an ideal test system since a defined capsid can self-assemble from only its hexons [91]. To understand this assembly in detail, we investigated the hexon's folding, dynamics, and oligomerization properties. Its protomer is a double jelly roll (DJR) major capsid protein (MCP). The jelly roll is a common supersecondary structural motif in viral capsid proteins[92,93], where eight beta strands are arranged in two four-stranded antiparallel sheets. In DJR proteins, two single jelly rolls are connected by a short linker. X-ray crystallography revealed a trimeric Mavirus hexon (MCP_3) with tightly intertwined N-terminal arms[91]. Each arm wraps around the base of the other two protomers and seems to be locked within a clasp formed by the neighboring protomer's C-terminus. This structural motif apparently stabilizes the capsomer and thus the whole capsid.

To provide a detailed biophysical view of this mechanism, we studied the hexon's dynamics using a combination of in vitro kinetics and equilibrium approaches, HDX-MS, and in silico MD simulations[1]. We find that its highly intertwined N-termini establish a balance between stability and plasticity in the structure. The corresponding conformational barrier may prevent monomerization and at the same time facilitate structural flexibility in the core, thus stabilizing the hexon in an assembly-competent conformation. This could be a key component in promoting conformational sampling and hence a productive capsid assembly reaction. As similar mechanisms have been observed in other capsid hexons[94–96], our findings help to better understand viral stability and assembly in general.

Results

The N-terminal arms detach first followed by unfolding of the DJR core.

The hexon's symmetric crystal structure (PDB code 6G45[91]) is depicted in Fig. 4.5. Its intertwined N- and C-termini form three basal interprotomer beta sheets, where each N-terminus (residues 1 to 36) spans over both its neighboring protomers and each C-terminus (residues 489 to 504) forms a clasp around both N-termini from neighboring protomers.

[1] Experimental materials and methods are described in the Materials & Methods section of Ref.[28] in detail.

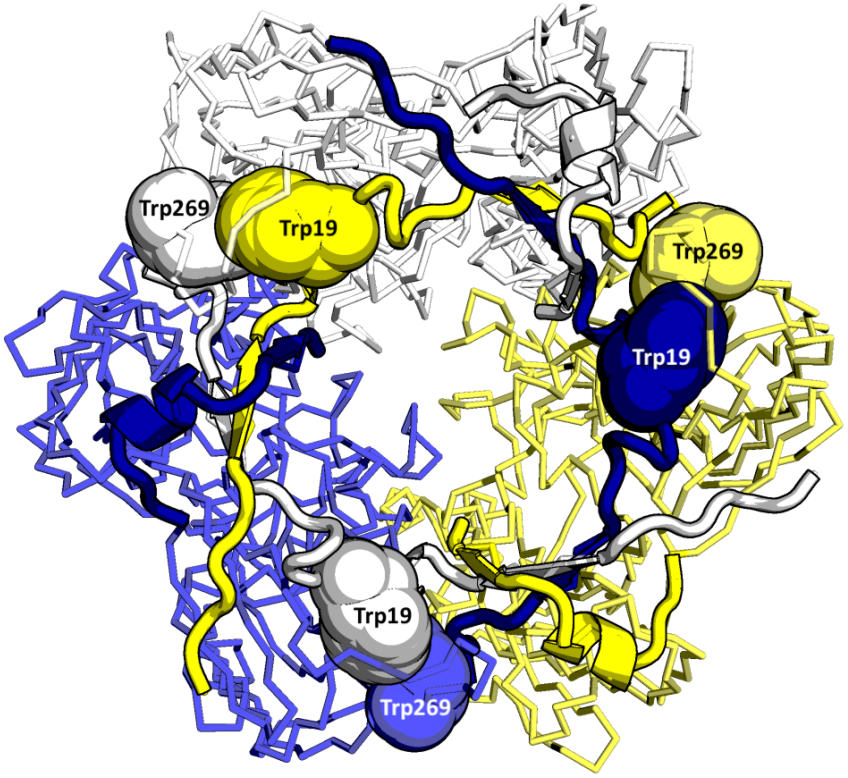


Figure 4.5. Structure and interactions at the basal interprotomer beta sheet of Mavirus' hexon (PDB code 6G45[91]). The protomers are colored white, yellow, and blue, respectively. Hexon with DJR core as ribbon and N- and C-terminal arms (residues 2 to 36 and 489 to 504, respectively) as cartoon. Spheres illustrate the interacting Trp19 and Trp269 residues. Trp19 is part of the N-terminal arm and in close proximity to Trp269 in the first DJR strand of a neighboring protomer. Visualized in PyMOL[1].

Fluorescence spectroscopy. Each protomer in the trimeric hexon has two intrinsically fluorescent Trp residues, one at position 19 (Trp19) in the N-terminal arm and another at position 269 (Trp269) in the first DJR's (DJR1) core. Its maximum-absorption wavelength is 280 nm. At 295 nm, its emission dominates the weaker tyrosine and phenylalanine fluorescence. Between 300 nm and 350 nm, Trp has a solvatochromic emission peak, i.e., its emission depends on its microenvironment's polarity. Consequently, its fluorescence can be used to probe the local environment. It provides information about the Trp's conformational state and, depending on its position in the sequence, about the whole protein's fold. Introducing a denaturant changes the Trp residues' microenvironment. Denaturing a protein with a Trp in its hydrophobic core yields a red-shifted emission spectrum due to the exposure of the non-polar Trp to an aqueous environment.

To probe the hexon's stability, we recorded fluorescence spectra (see Sec. 3.2) in absence and presence of the chaotropic denaturant guanidinium chloride (GdmCl). A chaotropic agent disrupts hydrogen bonds between water molecules and thus lowers the structural stability of natively folded proteins in the solution by weakening the hydrophobic effect. At 297 nm excitation, we observed a nearly triple increase and a slight red shift in the fluorescence intensity up to 1.0 M GdmCl, followed by a decrease and a further red shift up to 3.0 M GdmCl (see Fig. 4.6 a). This is in accordance with a denaturation curve at 340 nm emission, where we observed two transitions with a steep increase between 0.5 and 1.0 M, a plateau up

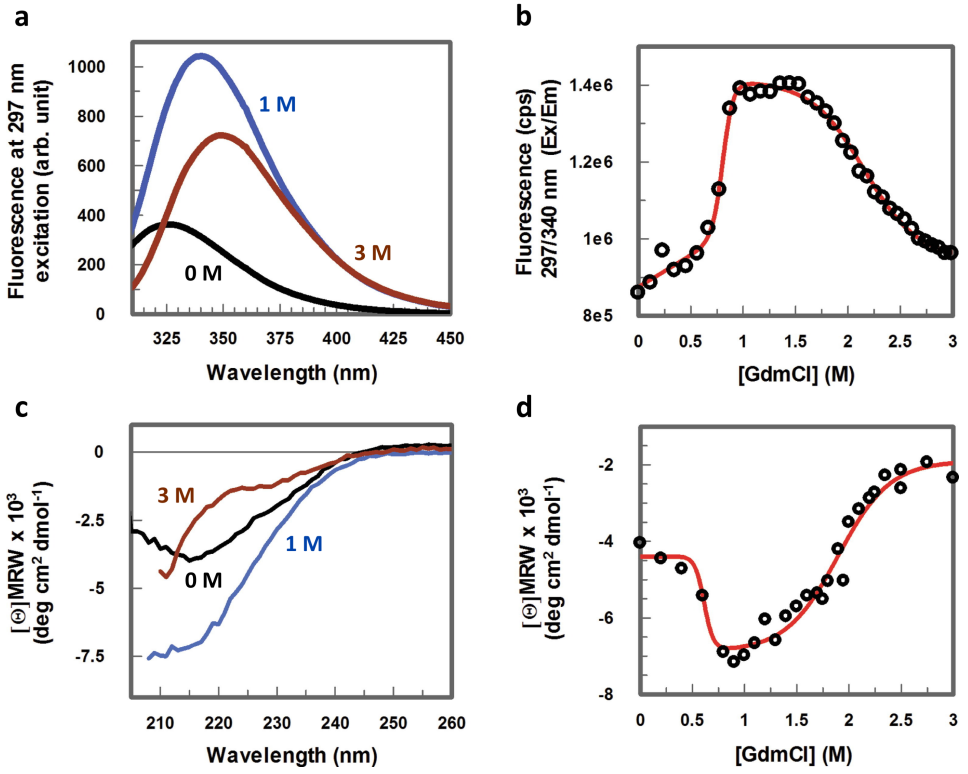


Figure 4.6. Hexon structure and stability. Fluorescence and CD spectra of MCP_3 at increasing GdmCl concentration, [GdmCl], and GdmCl titration curves at fixed wavelengths. Equilibrium spectra of 2.5 μM MCP at different [GdmCl] measured with **a**. fluorescence (297 nm excitation) and **c**. CD. Chemical denaturation followed at fixed wavelength for **b**. fluorescence at 340 nm for Trp emission (297 nm excitation) and **d**. CD at 217 nm. Reproduced with permission from Ref.[28] under CC BY-NC-ND 4.0 (relabeled from original).

to 1.5 M, and a decrease up to 3.0 M GdmCl (see Fig. 4.6 b). The second transition is associated with global unfolding. Complete denaturation leads to the exposure of Trp to a hydrophilic environment, inducing a decreased emission accompanied by the expected red shift. It is conspicuous that the Trp fluorescence at 3.0 M GdmCl is greater than at 0.0 M (see Fig. 4.6 b), indicating a significant quenching of the Trp fluorescence in the native state. The initial de-quenching between 0.0 and 1.0 M GdmCl is unusual for global unfolding and points to a different underlying process. As evident from Fig. 4.5, Trp269 in DJR1 lies within 6 to 7 Å from the N-terminal Trp19 of a neighboring protomer. Since Trp fluorescence is strongly influenced by the spatial proximity of other aromatic residues, this can cause unusual fluorescence phenomena[97,98]. If this is true for the Trp-Trp interaction in the Mavirus hexon, it can be used to track the N-terminal arms' movement with respect to the structural core. To validate the aromatic interaction between Trp19 and Trp269 as the cause of de-quenching in the first transition, we point-mutated either Trp to isoleucine (Ile). For both mutants, the first transition observed for the wildtype upon addition of GdmCl is lost (see Supplementary Fig. B.1 b). With increasing GdmCl concentration, the emission spectra are red-shifted and show a decrease (see Supplementary Fig. B.1 a), which is reminiscent of unfolding and comparable to the wildtype's second transition.

Circular dichroism spectroscopy. To test whether the Trp-Trp interaction can be used as an indicator of the N-terminal arms' relative position, we performed CD spectroscopy (see Sec. 3.4). CD is an excellent method to study a protein's secondary structure and folding properties[49] and mostly used to determine its degree of foldedness or if a mutation affects its conformation or stability. Aromatic residues with their bands in the near-UV region (> 250 nm) often exist in asymmetric environments and are particularly

suited to examine the effects of mutations. The far-UV CD spectrum can reveal information on a protein's secondary structure by estimating the molecular fraction in, e.g., the alpha-helix conformation, the beta-sheet conformation, or the beta-turn conformation[49,50].

Near-UV CD of the wildtype revealed a signal at 270 nm, indicating aromatic side chain interactions which were absent for both Trp mutants (see Supplementary Fig. B.1 d). This signal disappeared at 1 M GdmCl (see Fig. 4.6 d) and coincides with the first unfolding transition in the fluorescence data (see Fig. 4.6 b). Thus, the de-quenching observed during the first transition in the denaturation curve (see Fig. 4.6 b) is due to Trp-Trp quenching between the N-terminal arm and the DJR core. It serves as an intrinsic reporter for structural changes affecting the relative location of the N-terminal arms with respect to the hexon's core.

Next, we probed the hexon's global structural integrity. Well-defined antiparallel beta sheets typically produce negative bands at 218 nm and positive bands at 195 nm. Accordingly, the wildtype's CD spectrum reflects its beta-strand rich DJR structure (see Fig. 4.6 c, black). Adding GdmCl up to 1.0 M surprisingly resulted in an increased CD intensity accompanied by a shift in the maximum wavelength (see Fig. 4.6 c, blue), pointing to a structural rearrangement of the protein. Spectra at 3.0 M GdmCl showed a decreased CD signal and a further change in line with an unfolded random coil (see Fig. 4.6 c, red). A denaturation curve measured at 217 nm revealed two transitions with inflection points comparable to those observed in fluorescence (see Fig. 4.6 d). The Trp mutants' far-UV CD spectra were similar to those of the wildtype (see Supplementary Fig. B.1 c). This indicates that the mutations do not affect the hexon's overall fold and unfolding mechanism.

Combining the results from fluorescence and CD spectroscopy, we conclude that the hexon has a three-state unfolding mechanism, with the second transition corresponding to global unfolding. The first transition likely reports on a conformational change which can be assigned to the separation of the two Trp residues in neighboring protomers because of the N-terminal arms' detaching from the hexon's core.

Intertwined N-termini and C-terminal clasp prevent capsomer dissociation.

Förster resonance energy transfer spectroscopy. To evaluate whether the hexon is stabilized by the intertwined termini at its base, we investigated how GdmCl affects its oligomeric state using FRET (see Sec. 3.3). The efficiency of this non-radiative energy transfer between two light-sensitive molecules depends on the sixth power of their distance, making it extremely sensitive to nanometer distance changes[44]. Labeling specific residues with suitable dyes makes different conformations distinguishable.

To study the hexon with FRET spectroscopy, we double-labeled a cysteine mutant (MCP Asp277Cys) with Atto donor and acceptor dyes Atto488 and Atto594 (see Supplementary Fig. B.2 a). Upon excitation at the donor wavelength of 488 nm, the emission spectrum showed an additional FRET peak at the acceptor emission, confirming successful double-labeling (see Supplementary Fig. B.2 b). Addition of 1 M GdmCl resulted in a FRET decrease, i.e., an increased donor and a decreased acceptor intensity (see Supplementary Fig. B.2 b). The obtained FRET signal calculated as the ratio of emissions at 520 nm (donor) and 630 nm (acceptor) at 488 nm excitation as a function of the GdmCl concentration (see Supplementary Fig. B.2 c) revealed a single transition matching the first unfolding transition observed in CD and fluorescence spectroscopy. This indicates that the hexon monomerizes upon detaching of the termini from its base. Time-resolved FRET measurements of the double-labeled cysteine mutant showed an increasingly faster decrease in the FRET signal with increasing GdmCl concentration (see Fig. 4.7 a). We fitted the resulting kinetic traces with single exponentials. The derived rate constants showed an initial increase up to 1 M GdmCl and leveled off afterwards (see Fig. 4.7 c). We linearly extrapolated

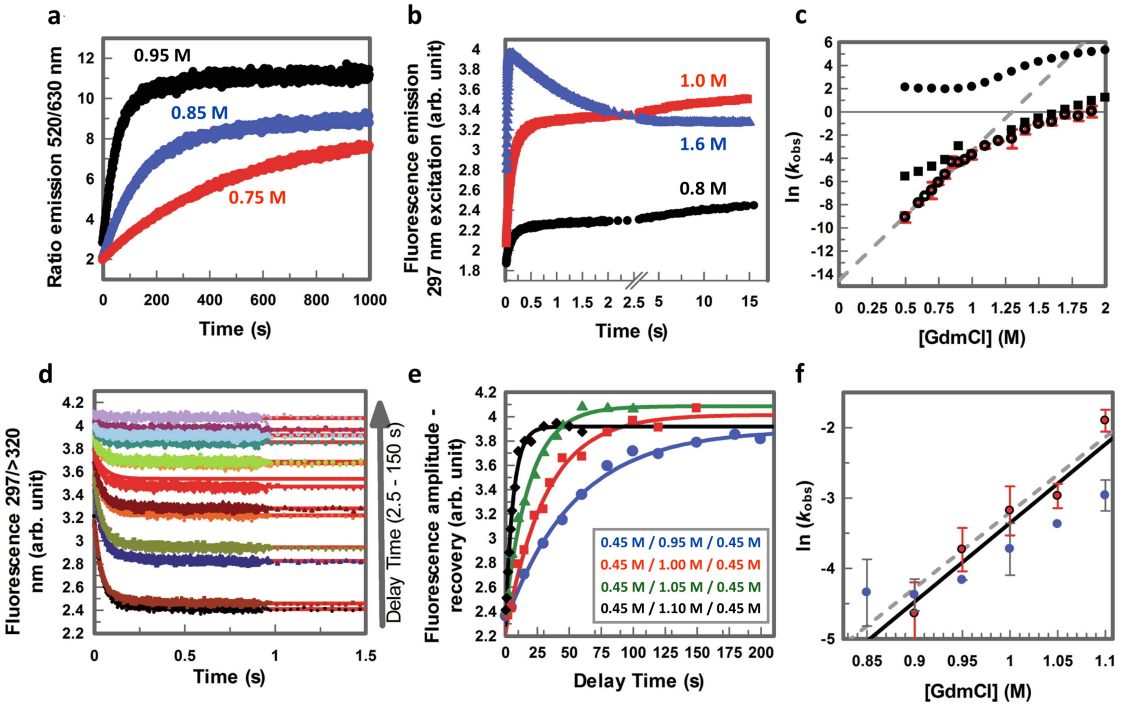


Figure 4.7. Kinetic FRET and stopped-flow single- and double-jump fluorescence measurements. **a**. Time-dependent FRET signal upon jumps to different [GdmCl] from pure buffer. **b**. Fluorescence unfolding traces upon Trp excitation at 297 nm induced by jumps from 0.0 M to different [GdmCl]. **c**. Rate constants from double-exponential fits to the Trp fluorescence traces (black circles: fast phase, black squares: slow phase) and mono-exponential fits to the FRET traces (open circles). Error bars from 2 to 4 repetitions at each [GdmCl]. **d**. Fluorescence traces of the second (backward) jump from 1 M to 0.45 M for different delay times after an initial forward jump from 0.45 M GdmCl to 1 M GdmCl. The fluorescence (297 nm excitation using a 320 nm long pass filter) was recorded after the second jump. The final MCP concentration was 1 μM. **e**. Double-jump fluorescence traces at varying [GdmCl] to show the progressive irreversibility of the N-terminal arms' separation. The final fluorescence for the backward jumps to 0.45 M GdmCl from various [GdmCl] after 5 s was recorded and plotted as a function of delay time. Solid lines are mono-exponential fits to the resulting traces. **f**. Replot of the rate constants obtained from **e**. for the hexon's N-terminal arms' movement's irreversibility as a function of [GdmCl]. A linear fit to the data points (dashed grey line) yields an extrapolated rate constant of $9 \cdot 10^{-7}\,s^{-1}$ at 0 M GdmCl (ln = -13.9). Averages and error bars from 2 to 3 repetitions. Red circles: double-jump rate constants; blue squares: FRET rate constants; solid black line: linear fit to FRET dissociation data from **c**. Reproduced with permission from Ref.[28] under CC BY-NC-ND 4.0 (relabeled from original).

a theoretical rate constant of disassembly without denaturant between 0.5 and 1.0 M GdmCl, which turned out to be about $5 \cdot 10^{-7}\,s^{-1}$. Accordingly, the trimeric hexon is very stable if its basal structure of N- and C-termini is intact.

Stopped-flow measurements. We elucidated the order of events during the separation of the N-terminal arms from the core (by Trp fluorescence) and the capsomer disassembly (by FRET) with stopped-flow measurements. A stopped-flow instrument is a rapid mixing device for studying the kinetics of fast reactions in solution. Such systems are typically modeled by conventional kinetic equations. After rapidly mixing the dissolved reagents, an observation cell is filled by a piston linked to a sensor triggering the measuring device. The flow is stopped suddenly. Coupled to either a CD or fluorescence spectrometer, stopped-flow instruments are often used to observe rapid protein unfolding and refolding.

For the wildtype hexon, stopped-flow Trp fluorescence showed a fast phase accounting for the bulk of the amplitude, a minor second phase with an increase up to 1 M GdmCl, and two distinct phases at higher GdmCl concentrations (see Fig. 4.7 b). The rate constants derived from exponential fits revealed two well-separated processes. The first phase is associated with the intensity increase and stayed constant between 0.5 and 1.0 M GdmCl (see Fig. 4.7 c, black dots) before increasing with the GdmCl concentration. The second phase (see Fig. 4.7 c, black squares) set in at 0.5 M and became more pronounced at higher concentrations. In contrast, both Trp19Ile and Trp269Ile mutants exhibited only a single phase at 297 nm excitation, coinciding with the wildtype's slow phase (see Supplementary Figs. B.3 a and b). The slow phase can consequently be associated with the unfolding of the DJR cores. Considering the reversibility of Trp de-quenching and capsomer dissociation, refolding was only partially successful after chemical denaturation (see Supplementary Fig. B.4). Only the second unfolding transition (denaturation) could be reversed on a minute time scale.

Double-jump kinetics. To probe if the first unfolding transition can be recovered on a faster time scale, we used double-jump kinetics measuring Trp fluorescence. The first jump from 0.45 M GdmCl (near the first unfolding transition) to a higher concentration allowed the fast Trp de-quenching reaction to proceed for a defined delay time t_1 (0.1 to 120 s). Subsequently, a second mixing reversed the GdmCl concentration to 0.45 M (see Fig. 4.7 d). Increasing t_1 decreased the initial fluorescence signal's recovery after the second jump. It was increasingly difficult to reverse the de-quenching and thus the separation of the N-terminal arm from the DJR core. We fitted the final fluorescence values after the double jump versus different delay times with a mono-exponential equation. The obtained trace showed the progressive irreversibility of the N-terminal separation (see Fig. 4.7 e). Repeated measurements with different intermediate GdmCl concentrations during the first mixing step revealed the time required to produce a reassembly-incompetent form to be shorter at higher concentrations (see Fig. 4.7 f). The corresponding rate constants increased linearly with the GdmCl concentration (see Fig. 4.7 f). While the de-quenching caused by the Trp separation could be reversed on a fast time scale, the hexon's disassembly as observed by FRET was not reversible. The rate constants for the irreversible movement of the N-terminal arms as observed via the Trp-based signal and those for monomerization from FRET experiments are very similar (see Fig. 4.7 f). We find that the inability to restore the N-terminal arms to their initial positions is founded in the hexon's irreversible dissociation. The denaturation of individual protomers could be reversed to a certain extent. We conclude that the intricate entanglement of the N-terminal arms and the C-terminal clasps across the hexon's base has to dissolve before capsomer dissociation.

Molecular simulations. To validate our experimentally derived working model, I investigated the N-terminal arms' role for the hexon using structure-based simulations (see Sec. 4.4). After determining native contacts with the shadow algorithm [21], I set up an all-atom SBM from the hexon's crystal structure (PDB code 6G45) [91] with `eSBMTools` [70,99]. Intra- and interprotomer contacts were weighted according to a 2:1 ratio. The hexon was simulated at constant temperatures T between 40 and 140 reduced units in `GROMACS` [100]. I calculated the per-residue root-mean-square fluctuations (RMSFs) of atomic positions as a measure of relative conformational flexibility in the hexon structure. The RMSFs at $T = 90$ indicate a high relative flexibility of the protomers' N-terminal arms (see Fig. 4.8 a). I further considered the C_α distance of the clasp-forming residues (valine 9 and proline 1505) as a function of simulated time. In the native clasp, these and neighboring residues induce the interaction between the N- and C-terminal parts of chains A and C, respectively. Its development over time shows repeated loosening and tightening of the clasp, i.e., opening and closing of the N-terminal arm (see Fig. 4.8 b). My simulations at $T = 106$ reveal a distinct three-state dissociation mechanism (see Fig. 4.8 d). Starting from the trimeric hexon clamped by three structure-spanning terminal clasps, the N-terminal arms disentangle first while the

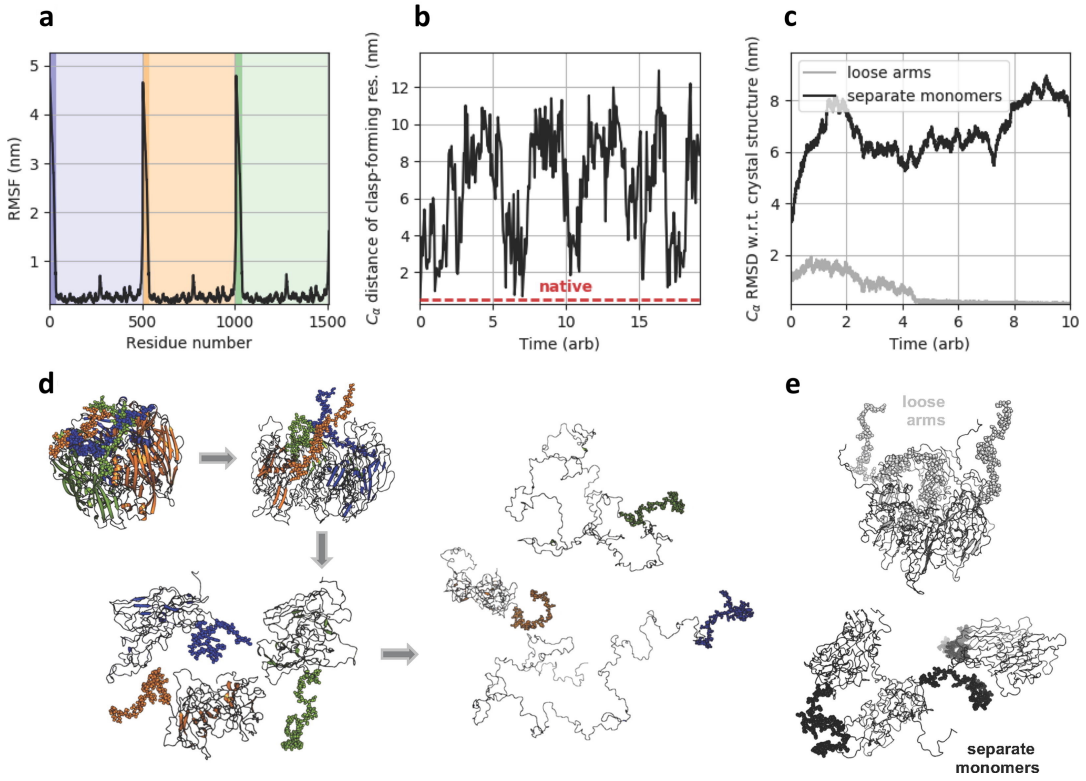


Figure 4.8. Hexon SBM simulation results. **a**. The average RMSFs of atomic positions per residue reveal a relatively high mobility of the protomers' N-termini. The background is shaded according to the individual protomers' colors in **d**. with N-terminal regions depicted slightly darker. **b**. The time-dependent C_α distance of the clasp-forming residues valine 9 and proline 1505 indicates a repeated loosening and tightening of the clasp, i.e., opening and closing of the N-terminal arm. The simulation used a reduced GROMACS temperature of 90. **c**. Time-dependent C_α RMSDs with respect to the crystal structure. The simulations initiated from an intact hexon structure with loose N-terminal arms (gray) and a fully dissociated structure of three separate monomers (black) as displayed in **e**. According to the minimizing gray curve, the loose arms could easily be folded back to the structurally intact trunk. Starting from separate monomers, the system did not re-converge to the crystal structure as indicated by the diverging black curve. **d**. Representative cartoon structures illustrating our model of the hexon's dissociation mechanism with N-terminal arms depicted as spheres. *First reaction:* Starting from the crystal structure, the N-terminal arms detach from the stable core. *Second reaction:* The complex separates into three structurally intact monomers. *Third reaction:* The monomers unfold completely. This simulation used a reduced GROMACS temperature of 106. Reproduced with permission from Ref.[28] under CC BY-NC-ND 4.0.

structural core remains relatively stable. Subsequently, the complex separates into three structurally intact protomers before the individual subunits finally unfold completely.

To probe the reversibility of this dissociation process, I manually extracted partially unfolded and dissociated starting structures from the constant-temperature trajectories (see Fig. 4.8 e). The dissociated starting structure was energy-minimized before the actual simulation. For the structure-based simulations of these systems, I additionally applied a simulated-annealing protocol. In simulated annealing, the system is heated to a temperature providing access to all conformations of interest and allowing to overcome any energy barriers between them[101]. Subsequently, the system is cooled down so that the highest energy conformations become gradually inaccessible but escaping from them to lower energy states is still possible. Ideally, the system is confined to its global minimum energy conformation as the temperature approaches its minimum. I decreased the temperature linearly from $T = 90$ to $T = 60$ after 90 % of the simulated time. Starting from a structurally intact hexon with loose N-termini, the

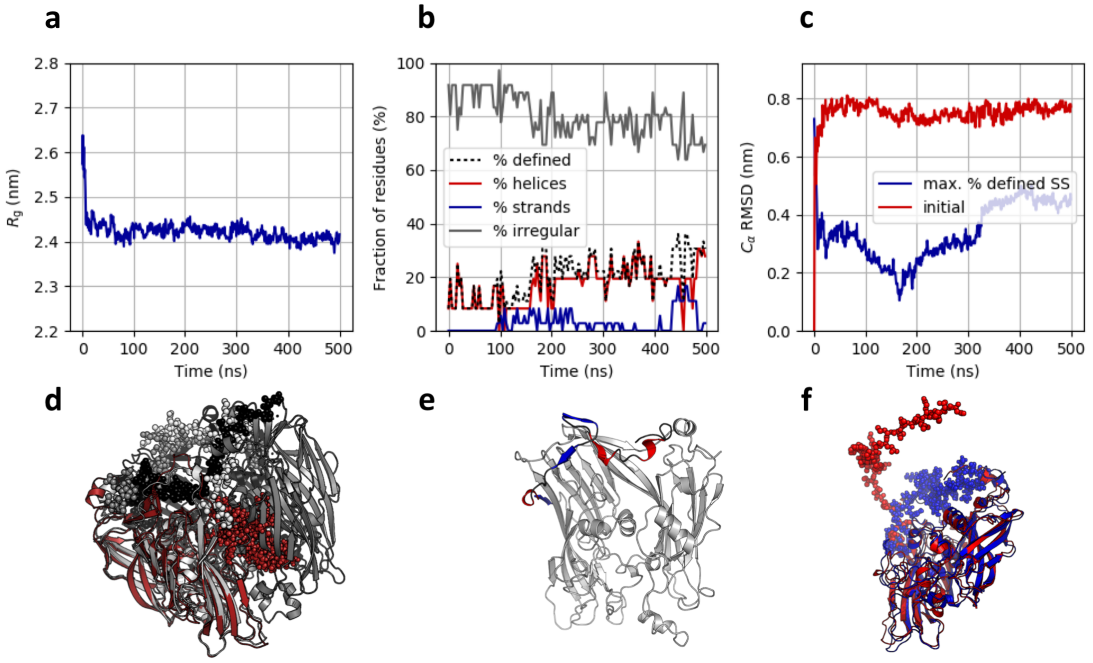


Figure 4.9. Protomer explicit-solvent MD simulation results. Results are shown for a temperature of 335 K. **a**. Radius of gyration R_g versus simulated time. The initially free N-terminus compacts and stably attaches to the protomer's trunk (see **f**.), resulting in a significantly reduced R_g over time. The trimerization interface is predicted to be spatially blocked, hindering a proper reassembly of the hexon as displayed in **d**. A representative structure of the equilibrated free monomer (red) is superimposed onto a protomer subunit of the trimeric hexon (gray). **b**. Fraction of secondary structure in the N-terminal arm versus simulated time. The structure with maximum fraction of distinct N-terminal secondary structure is shown in **e**. **c**. Time-dependent C_α RMSD with respect to the protomer as extracted from the crystal structure (red) and with respect to the structure with maximum fraction of defined N-terminal secondary structure (blue). The simulation persistently proceeded closer to the compacted state than to the state with elongated arm. **f**. Structure with maximum fraction of defined N-terminal secondary structure (blue) superimposed onto the protomer as extracted from the crystal structure (red). Reproduced with permission from Ref.[28] under CC BY-NC-ND 4.0.

arms could be folded back onto the structural core by gradually cooling the system down (see Fig. 4.8 c, grey). The system eventually reached its native state as evident from the time-dependent C_α RMSD with respect to the crystal structure. However, starting from the fully dissociated hexon, i.e., three separate but spatially close protomers, the system did not converge to the native state, as indicated by its diverging C_α RMSD with respect to the crystal structure (see Fig. 4.8 c, black).

To further investigate why the separate protomers cannot reassemble into the native hexon, I performed explicit-solvent MD simulations of an individual protomer (residues 2 to 504) with elongated N-terminal arm as extracted from the crystal structure (see Fig. 4.9 f, red). The system was preprocessed in a TIP3P water box with a minimum distance of 2 nm from the edges, charge-neutralized, energy-minimized, and successively equilibrated in the canonical and isothermal-isobaric ensemble. The simulations were run for 500 ns at temperatures $T = 300$, 314, 325, and 335 K and a pressure $p = 1$ bar. I used the AMBER99SB-ILDN[102] force field, the V-rescale temperature coupling scheme, a Parrinello-Rahman barostat, and Verlet neighbor searching. Bonds were constrained using the LINCS algorithm. Electrostatics were treated with the Particle Mesh Ewald method. In the simulations, the initially free N-terminus condensed and finally attached to the protomer's DJR core (see Figs. 4.9 a and f, blue), thus blocking the interprotomer interface and hindering reassembly (see Fig. 4.9 d). After the N-terminal's initial collapse, the structure persistently stayed closer to the compacted state than to the extended state

(see Fig. 4.9 c), which is also reflected by its decreasing radius of gyration R_g (see Fig. 4.9 a). I used the DSSP (Define Secondary Structure of Proteins) algorithm to investigate the N-terminal arm's secondary structure over the simulation. DSSP [103] is the standard method for assigning secondary structure to a protein's residues based on its atomic coordinates. It predicted a maximum (average) fraction of 47.2 % (approx. 20 %) of residues to have defined secondary structure (see Figs. 4.9 b and e, respectively).

High structural plasticity of the hexon's second DJR could be essential for capsid assembly.

Our spectroscopic and simulation-based experiments on the hexon's stability suggest that its basal N- and C-terminal structure prevents dissociation. We hypothesize that the hexon's core might be more flexible, probably to enable capsid assembly and structural adaptation to the required position in the capsid as observed for the closely related Sputnik virophage [104].

Hydrogen-deuterium exchange coupled to mass spectrometry. Using HDX-MS (see Sec. 3.5), we aimed at detecting accessible and mobile elements in the native hexon to identify regions relevant to its stability and eventually to capsid assembly. HDX-MS is a technique for monitoring the structure and dynamics of proteins dissolved in heavy water, where the rate of deuterium exchange acts as a surrogate for their structure and stability. A slower rate of exchange, and hence a higher degree of protection from the latter, points to a more stable structure [51]. We mapped the H/D exchange rates in the amide bonds for the individual protomeric fragments (see Supplementary Information of Ref. [28]). In line with our spectroscopic and simulation data, we observed a high accessibility for residues 1 to 36 and 495 to 505, corresponding to the N-terminal arm and the clasp-forming C-terminal part, respectively (see Fig. 4.10). The experimentally unresolved C-terminal residues 505 to 606 also showed a high exchange rate, suggesting that they are indeed unstructured [91]. Compared to DJR1, the second DJR (DJR2, residues 300 to 590) interestingly showed a higher conformational dynamics almost on the level of the N-terminal arm (see Fig. 4.10). This is reminiscent of the observed plasticity in the second DJR of the native Sputnik virophage. The locations of these highly dynamic elements are illustrated in Fig. 4.11 as a heat map projection of the exchange rates on the Mavirus hexon's structure. As these "hot" jelly roll regions mostly lie on the outer surface involved in the intercapsomer interactions, we suggest that the DJR2 plasticity in the Mavirus hexons is critical to a productive capsid assembly. At the same time, this plasticity forces the hexon to develop basally intertwined termini to prevent disassembly.

Discussion

Our results enable us to draw a model picture describing the hexavalent capsomer's inherent dynamics (see Fig. 4.12). In its trimeric form, the Mavirus hexon is stabilized by intertwined N-terminal arms and C-terminal clasps at its base. A displacement of the arms, as measured by Trp fluorescence de-quenching, allows the protomers to dissociate. As long as the trimer is intact, the arms can revert to their native position and the clasps can close again. However, after full dissociation of the protomers, a reassembly to the trimer is impossible. Explicit-solvent MD simulations of the protomer suggest that this behavior is due to the loosened N-terminal arm's re-attaching to the DJR core upon refolding, which blocks the trimerization interface. According to CD spectroscopy, the protomer's global structure changes as well, which may also prevent the association to a trimer and thus a productive capsid assembly. The hexon's stabilization mechanism has some analogies with domain swapping [105,106], a process where two or more identical monomers exchange structural elements and fold into dimers or multimers with protomeric subunits similar to the original monomer. The separate protomer in its compacted state as observed in explicit-solvent MD would have the exchanging N-terminal arm closely attached to its

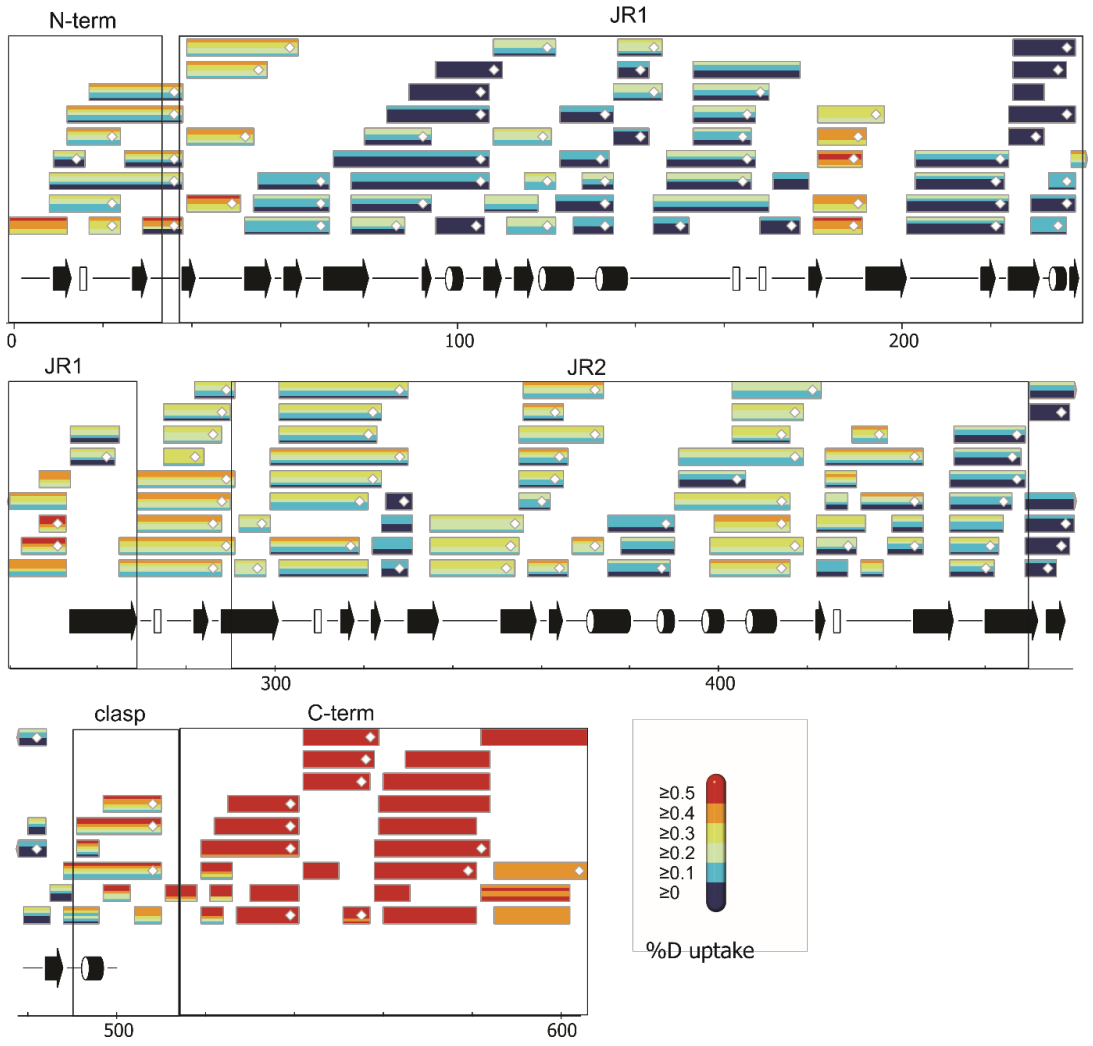


Figure 4.10. HDX-MS analysis of full-length MCP protomer. The individual peptides are represented as boxes with up to six different colors, which, from the bottom up, represent the deuteration times of 15 s, 45 s, 3 min, 15 min, 1 h, and 4 h. The colors give the relative deuterium uptake normalized to the total number of exchangeable amide positions per peptide as indicated in the legend. Based on the fully exchanging C-terminal peptides, a back-exchange of approx. 40 % is estimated and, considering that the first two amide positions back-exchange quickly, the scale was adjusted to represent the accessible deuterium uptake range. The colors thus indicate the individual peptides' secondary structure stability with red and blue corresponding to the lowest and highest stability, respectively. MS^2 confirmed peptides are marked with diamonds. Terminal arrows at the end of a box indicate continuation of the peptide in the previous or following line. Secondary structure elements were derived from a DSSP analysis of the crystal structure (PDB 6G45, chain A). Numbering according to wildtype MCP (Uniprot A0A1L4BK98). Reproduced with permission from Ref.[28] under CC BY-NC-ND 4.0.

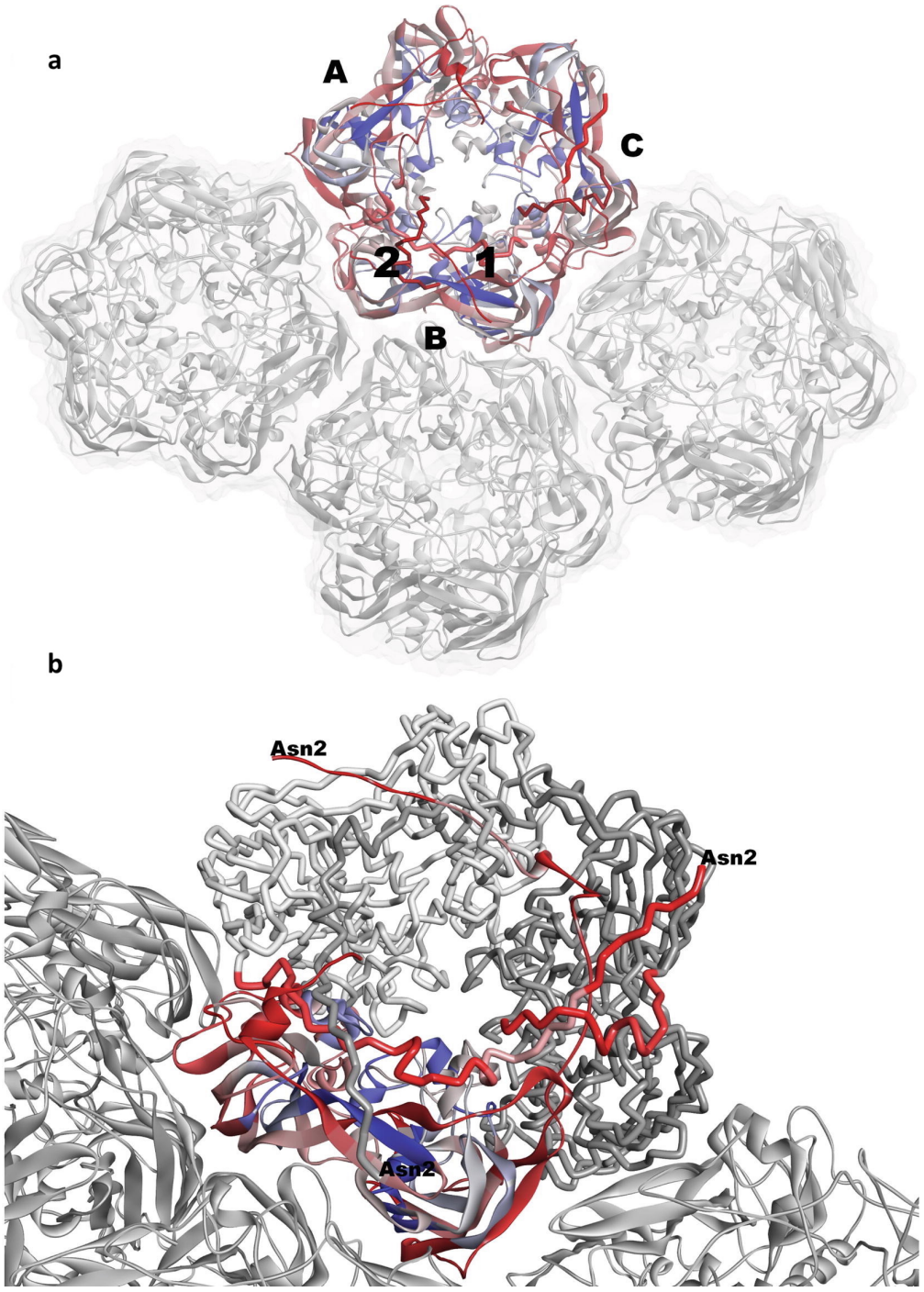


Figure 4.11. HDX-MS data projected onto the Mavirus hexon as a heat map. a. The Mavirus hexon[91] is depicted in the context of the Sputnik virophage's capsid[104] as seen from inside the capsid. Sputnik capsomers are colored light gray. The Mavirus hexon's ribbon structure is colored according to the HDX exchange data at 240 min with a blue-white-red scaling to indicate the increasing deuterium incorporation. The protomers are labeled A, B, and C, and the two jelly roll structures of protomer B are labeled 1 and 2. **b**. Interface-forming regions and dynamics of the two jelly roll structures of protomer B colored according to the HDX heat map as seen from inside the capsid. The N-terminal arm of protomer A is displayed as a tube and colored according to the HDX data as are the two clasp regions of protomers B and C. The N-termini of all three subunits are labeled with Asn2. Reproduced with permission from Ref.[28] under CC BY-NC-ND 4.0 (relabeled from original).

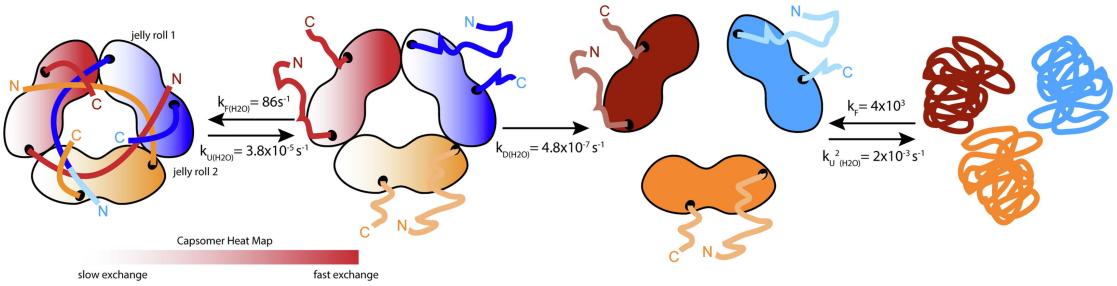


Figure 4.12. Quantitative model of the Mavirus capsomer's dynamics. Initially, the monomers are associated to a trimer with the N-terminal arms forming a stabilizing brace with neighboring protomers. *First reaction:* Under denaturing conditions, the N-terminal arms detach. The trimer is still intact. *Second reaction:* Without the brace, the protomers can dissociate irreversibly. *Third reaction:* The monomers can reversibly fold and unfold but do not re-associate to a trimer on their own. Reproduced with permission from Ref.[28] under CC BY-NC-ND 4.0.

own DJR core. It however remains unclear whether the 36 N-terminal residues constitute a domain or folding unit themselves. Neither in the crystal structure nor in the explicit-solvent MD did more than 20 % of the N-terminal residues adopt a discernible secondary structure. The question is why such an intricate interaction is necessary; after all, the domain exchange is associated with at least one energy barrier for a partial unfolding event[107]. The advantages of an internally highly symmetric oligomer, i.e., a lower chance of aggregation, a higher stability[108], and viral code and gene minimization, which is important due to a virus' limited genomic space[109], could be realized without an additional brace mechanism. Mavirus' minor capsid protein, for example, forms a pentamer from five single jelly rolls without any entanglement between its subunits. Still, it is substantially more stable, at least against thermal denaturation[91]. Stabilizing capsomers through a terminal brace-and-clasp mechanism as for the hexon depends on the protomers to be in an assembly-competent state. This requires at least one global structural rearrangement and potentially a second one for domain swapping that could not be achieved under in vitro conditions. In vivo, this is most likely facilitated by chaperones, co-translational modification, or non-viral helper proteins[110,111]. N-terminal interactions were also observed in studies on capsomers of other DJR-containing viruses, such as PRD1[95], Faustovirus[96], or Adenovirus[112], and considered relevant to their stability. Compared to PRD1, which is very stable against denaturation[113], Mavirus' hexavalent capsomers are considerably less so. This indicates that stabilizing the trimeric hexon might not be the sole purpose of its basally intertwined structure. The following observations are key aspects: (i) the N-terminal movement is not the rate-limiting step for trimer dissociation but considerably faster, and (ii) at intermediate GdmCl concentrations, the protomer's structure, as judged by CD, changes after dissociation. Since the Mavirus capsid can assemble in *Escherichia coli*, its assembly is obviously independent of specific co-factors provided by the native host[91]. This renders the capsomer's intrinsic properties much more important. Both C- and N-terminal parts are critical entities and have been shown to be crucial for the assembly of other virus capsids. For phage P22, interactions of the capsomers' N-termini with adjacent capsomers are essential for capsid assembly[114]. Likewise, the C-termini form contacts with neighboring capsomers for murine polyomavirus[115]. The bacteriophage T4 is similar to Mavirus in the sense that the N-terminus is important for MCP monomer-to-hexamer association, although it is cleaved off in this process[116]. For Mavirus itself, the C-termini are known to be important in supporting capsid assembly[91], whereas the N-termini have not been reported to be directly involved. Based on our results, we postulate that the N-terminus might also play a role in the assembly process. Especially the change in the protomer's overall structure upon capsomer dissociation as observed by CD spectroscopy points to different monomeric and oligomeric structures. In addition, explicit-solvent MD simulations also predicted an association-incompetent conformation of the monomeric protomer specifically involving the released N-terminal arm. As can be seen in HK97[117] or

PRD1[118], a productive capsid assembly often requires the hexons and pentons to undergo conformational changes, and a certain structural malleability would be a prerequisite for this. The Mavirus hexon balances the requirement of having a pre-organized trimer with the necessity of switchable interface structures for an efficient capsid assembly. For this purpose, it uses a special brace-and-clasp mechanism to prevent the trimeric system from falling apart. Our hypothesis is supported by the HDX-MS data, showing that the protomer's DJR2 has a high conformational flexibility comparable to that of the N-terminus. This could be a conserved mechanism in virophages as, e.g., DJR2 of the Sputnik virophage is strongly involved in intercapsomer interactions and also shows a high structural plasticity, while the intertwined arrangement of N- and C-termini at the capsomer's base is highly similar to that in Mavirus[104]. We showed that the main function of the brace-and-clasp mechanism of Mavirus is most likely not only the stabilization of its capsomer, as has been assumed from structural studies of related viruses. By preserving the capsomers' flexibility, this mechanism additionally allows for the structural malleability required for a productive capsid assembly. Our findings highlight the importance of studies on capsomer dynamics as a complement to structural evidence. They extend our understanding of the assembly of not only Mavirus and the closely related Sputnik virophage but also of related viruses, such as poxo- or the large family of adenoviruses.

4.7 PROJECT: Simulating the Interplay of FRET and SAXS with Structure-Based Models

This work was done in collaboration with Ines Reinartz from the Institute for Automation and Applied Informatics at Karlsruhe Institute of Technology. Based on her structure-based simulations of FRET-dye-labeled proteins[11], I helped to set up and design this study with particular focus on the SAXS part. I performed a proof of concept by showing that including FRET dyes indeed affects SAXS intensities back-calculated from simulations of proteins with and without dyes. The following section is reproduced from our Israel Journal of Chemistry article "FRET Dyes Significantly Affect SAXS Intensities of Proteins" (2020)[29]. The project was mainly conducted by Ines Reinartz and is already described in her doctoral thesis. In the context of my work, it shall further underline the explanatory power of structure-based approaches and serve as a practical application example.

SAXS and FRET (see Secs. 3.1 and 3.3, respectively) are experimental approaches to studying the dynamic aspect of the protein structure-function paradigm. Both methods are widely used for analyzing unfolded and intrinsically disordered proteins (IDPs)[119,120]. Despite lacking defined structure, such proteins fulfill diverse functions in the human body. The combined application of FRET and SAXS has sparked a debate on how to interpret FRET data against the backdrop of SAXS measurements, in particular for unstructured protein systems. While recent FRET studies imply IDPs to be compacted in water compared to high denaturant concentrations, this could not be observed with SAXS. Theoretical simulations predict the unfolded states of globular proteins to also become more compact with decreasing denaturant concentration. Whereas single-molecule FRET data support this prediction, SAXS data point to the opposite[121]. This is the SAXS-FRET controversy[122–124]. Different explanatory approaches to resolving these inconsistencies exist. The observed discrepancies could be explained by decoupled size and shape fluctuations, which allows for the conclusion that FRET and SAXS do not measure the same quantity[125]. Unstructured proteins may undergo a sequence-specific decoupling of the SAXS-accessible radius of gyration, R_g, and the FRET-accessible end-to-end distance, R_e[126]. Other studies find that the discrepancies mainly stem from the analysis methods[127]. While interpreting FRET data requires a model to relate R_g and R_e, the interpretation of SAXS data is basically model-free[128].

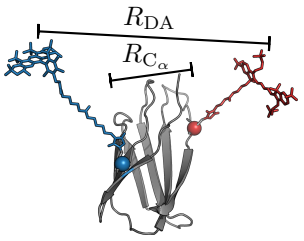


Figure 4.13. Dye-labeled tenth type III domain of fibronectin. The 94-residue tenth type III domain of fibronectin (^{10}FNIII, grey, PDB code 1TTG[129]) with Alexa Fluor dyes (blue and red) attached at residues 11 and 86. The corresponding C_α atoms are shown as spheres. R_{DA}, inter-dye distance, R_{C_α}, C_α distance. Reproduced from Ref.[29] under CC BY 4.0.

Still, we do not fully understand how to interrelate the quantities derived from FRET and SAXS. Central questions are how FRET dyes affect SAXS intensities in combined applications and to what extent R_g values obtained by either of the methods differ. We studied the interplay of FRET and SAXS within computational simulations. Molecular simulations are a powerful tool to analyze the influence of FRET dyes on SAXS intensities. They provide access to all variants of R_g. As explained in Sec. 3.3, FRET depends not only on the inter-dye distance but also on the dyes' dynamics. To investigate these aspects in atomic detail, we applied the structure-based method by Reinartz et al. for simulating FRET-dye-labeled proteins[11]. This technique requires only few parameters while maintaining full protein flexibility by including all heavy atoms of proteins, linkers, and dyes[11]. Such simulations allow us to calculate FRET efficiencies in agreement with experimentally measured values[11]. We calculated theoretical SAXS intensities from such simulations of four different proteins with and without dyes in folded and unfolded conformations. Including FRET dyes significantly impacts the intensities, especially for small proteins. We furthermore compared in-silico derived R_g variants commonly deduced from FRET and SAXS. The values obtained are different without and with dyes and depend on the chosen dye pair. The dyes seem to influence the dynamics of the unfolded systems. A systematic difference in SAXS- and FRET-derived R_g points to further mechanisms beyond existing explanation approaches.

Studied Systems

Proteins. The first test system is the 94-residue tenth type III domain of fibronectin (^{10}FNIII, PDB code 1TTG[129], see Fig. 4.13). Fibronectin is a dimeric glycoprotein of the extracellular matrix. It is involved in cell adhesion, growth, migration, and differentiation and plays a major role in wound healing and embryonic development[130]. Altered expression, degradation, and organization of this protein have been associated with severe diseases, including cancer and fibrosis[131].

Chymotrypsin inhibitor 2 (CI-2, PDB code 2CI2[132], see Fig. 4.14 a) is a well-understood 83-residue serine proteinase inhibitor from barley seeds and our second test system. Such small and well-characterized systems are particularly suited for studying protein-protein interactions and structure-function relationships. CI-2 was one of the first proteins whose folding/unfolding transition state was characterized in-depth with the protein engineering method[133,134]. NMR spectroscopy and hydrogen exchange were used to characterize its denatured state and folding[135–137].

We studied the 66-residue cold shock protein from *Thermotoga maritima* (CspTm, PDB code 1G6P[138], see Fig. 4.14 b). During cold shock, the efficiency of transcription and translation in DNA-based gene expression is reduced. This is due to the stabilized secondary structure of messenger RNA (mRNA).

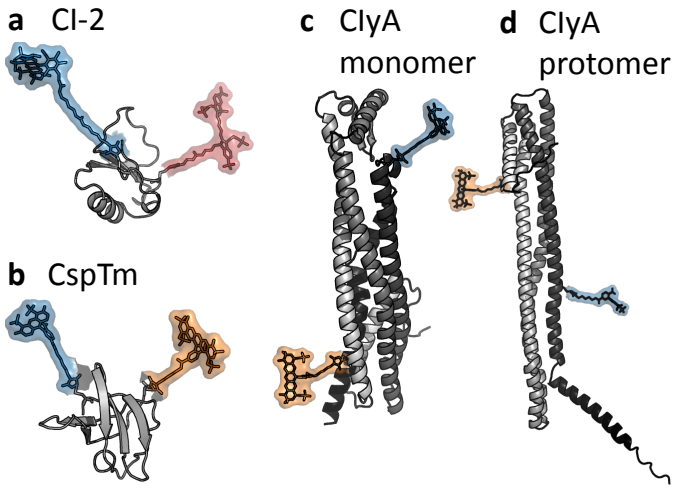


Figure 4.14. FRET-dye-labeled protein systems. **a**. CI-2 with AF dyes at positions 20 and 78. **b**. CspTm with AF dyes at positions 2 and 68. ClyA monomer **c**. and protomer **d**. with AF dyes at positions 56 and 252. Reproduced from Ref.[29] under CC BY 4.0.

mRNA is a single-stranded RNA molecule that has been copied from DNA in the process of transcription. It thus corresponds to the genetic sequence of a function-encoding DNA region, or gene. During translation, the transcribed mRNA is read by a ribosome to synthesize a protein from amino acids carried by the transfer RNA. To mitigate the reduced efficiency of these processes at low temperatures, many bacteria produce small cold shock proteins in response to a rapid temperature drop. These proteins counteract the formation of mRNA secondary structure by acting as nucleic-acid chaperones.

The last test system is cytolysin A (ClyA) of *Escherichia coli*, a common bacterium in the lower intestine of warm-blooded organisms. ClyA is a hemolytic toxin. It causes the rupture of red blood cells, which release their contents into the blood plasma in return. Before assembling into a dodecameric pore (PDB code 2WCD[139], see Supplementary Fig. C.2 c and d), the 303-residue ClyA protein switches from a monomer (PDB code 1QOY[140], see Fig. 4.14 c) to a protomer state (see Fig. 4.14 d).

Dyes. We used fluorescent dyes of the Alexa Fluor (AF) family[29,141]. These dyes are frequently applied as labels in fluorescence microscopy and named according to their approximated excitation maxima in nanometers. For the simulations involving three dyes, we additionally used the Biotium dye CF680R (B680)[142]. Figs. 4.13 and 4.14 show examples of the studied systems. All composite systems are depicted in Appendix C (see Supplementary Figs. C.1 and C.2). A detailed list of all systems as well as dye structures and parameters can be found in the Supplementary Information of Ref.[11].

Methods

Molecular simulations of dye-labeled proteins. We applied the structure-based protocol by Reinartz et al. for simulating dye-labeled proteins[11]. 3D dye structures were obtained from available chemical structures by quantum chemical calculations. The dyes were parameterized for inclusion into the SBM, where excluded-volume repulsion is the only interaction considered[11]. Finally, they were attached preferably orthogonally to the protein surface with linkers. The simulations were run in `GROMACS 4.5.4`[143] using the structure-based potential (see Eq. 4.13). Molecular dynamics parameters used are listed in Ref.[11].

Calculation of SAXS profiles from protein structures. Calculating accurate scattering patterns from atomic positions is essential to interpret SAXS data. We applied the widely used implicit-solvent method

`CRYSOL` to compute spherically averaged scattering patterns from protein structures[41] (see Sec. A.2). The solvation shell's excess density $\delta\rho$ may substantially influence the calculated SAXS curves and thus the derived radii of gyration, especially for unfolded proteins[144]. Instead of `CRYSOL`'s $0.03\,e\text{Å}^{-3}$ default, Henriques et al. recommend using lower values between $0.01\,e\text{Å}^{-3}$ and $0.02\,e\text{Å}^{-3}$. Comparing R_g values for different $\delta\rho$ showed that the latter only slightly affects SAXS-derived R_g for the considered proteins, which generally increase with $\delta\rho$ (see Supplementary Figs. C.7 to C.10)[29]. The overall trend discussed below is preserved among different values of $\delta\rho$.

Radius of gyration. The most popular structural feature derived from FRET and SAXS is a protein's radius of gyration, R_g, a measure of its overall molecular size. `GROMACS` calculates it from a molecular structure as[143]

$$R_{\mathrm{g,gmx}} = \sqrt{\frac{\sum_i r_i^2 m_i}{\sum_i m_i}}\,, \tag{4.19}$$

where atom i of mass m_i has a distance r_i to the molecular center of mass. To analyze the different R_g variants deduced from FRET and SAXS, we first calculated a reference $R_{\mathrm{g,gmx}}$ from the protein structure without dyes. Second, we considered the corresponding value $R_{\mathrm{g,gmx}}^{+\mathrm{dyes}}$ of the protein with dyes. The SAXS Guinier analysis (see Sec. 3.1) provides two additional values $R_{\mathrm{g,saxs}}$ and $R_{\mathrm{g,saxs}}^{+\mathrm{dyes}}$. To calculate the FRET-accessible R_g, we modeled the unfolded proteins as excluded-volume chains[11,29,145]. In order to emulate terminally attached dyes, we studied truncated systems derived from simulations of each full system, where we considered only atoms between the dye positions. We extracted the end-to-end C_α distance R_e (between the dye-labeled residues) to calculate the apparent R_g for excluded-volume chains[128],

$$R_{\mathrm{g,C}_\alpha}^{\mathrm{app}} = \sqrt{\frac{\langle R_\mathrm{e}^2\rangle}{6.26}}\,. \tag{4.20}$$

FRET is often assumed to deliver the distance between dye-labeled C_α atoms, thus ignoring the dyes' linkers. We modeled the linkers as chain extensions of additional sequence length L and rescaled the inter-dye distance R_DA to a C_α distance via[11,128]

$$f = \left(\frac{N_{\text{inter-dye}}}{N_{\text{inter-dye}} + L}\right)^{\nu}\,. \tag{4.21}$$

$N_{\text{inter-dye}}$ is the number of residues between the dye-labeled sites. $\nu = d_\mathrm{f}^{-1}$ is the scaling exponent, which can be derived from the fractal dimension, d_f, in a Porod plot of the SAXS intensity, $I(q) \propto q^{-d_\mathrm{f}}$, at higher q in the so-called "power-law regime" (see Supplementary Information of Ref.[29]). Thus:

$$R_{\mathrm{g},R_\mathrm{DA}}^{\mathrm{app}} = f \cdot \sqrt{\frac{\langle R_\mathrm{DA}^2\rangle}{6.26}} \tag{4.22}$$

Results

Influence of FRET dyes on R_g distributions. To investigate how FRET dyes influence a protein system, we studied the $R_{\mathrm{g,gmx}}$ distributions of each simulation's frames[29]. Exemplary distributions are shown in Supplementary Figs. C.3 to C.6 for ^{10}FNIII and ClyA monomer. As expected, the average $R_{\mathrm{g,gmx}}$ is greater with dyes than without in both folded and unfolded states. Naturally, this effect is more pronounced for the smaller ^{10}FNIII, where the dyes are relatively larger. Due to their generally larger conformational heterogeneity, this effect relativizes for the unfolded systems, resulting in broader distributions compared to the folded case.

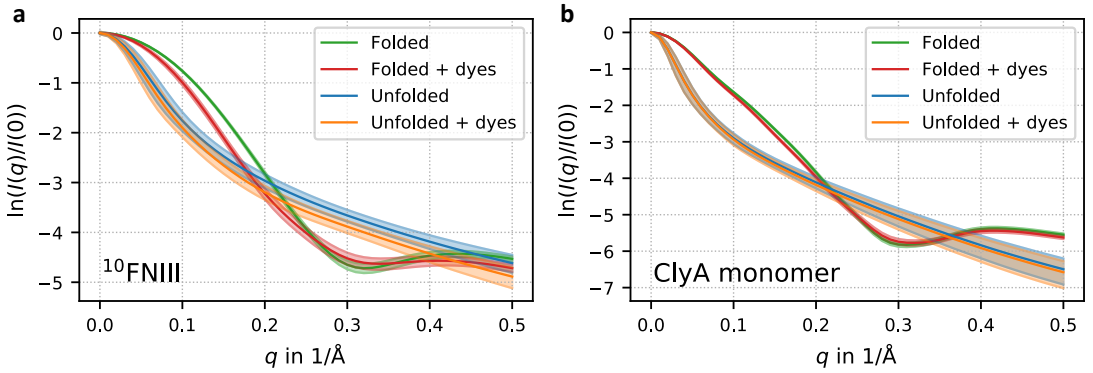


Figure 4.15. SAXS intensities without and with dyes for a. ^{10}FNIII and b. ClyA monomer with dyes at positions 56 and 252. Average intensities (solid lines) versus momentum transfer q along with each intensity-curve distribution's standard deviation over corresponding single-frame intensities calculated from individual simulation snapshots (shaded area). This standard deviation is a measure of conformational heterogeneity in the simulated ensemble (also see R_g distributions). It indicates the intensity distribution width at a particular q point rather than an actual error in the sense of statistical uncertainties or systematic deviations as they would occur in experimental data. This also evidences in the fact that the standard deviations are larger for the unfolded systems than for the folded systems. Reproduced from Ref.[29] under CC BY 4.0.

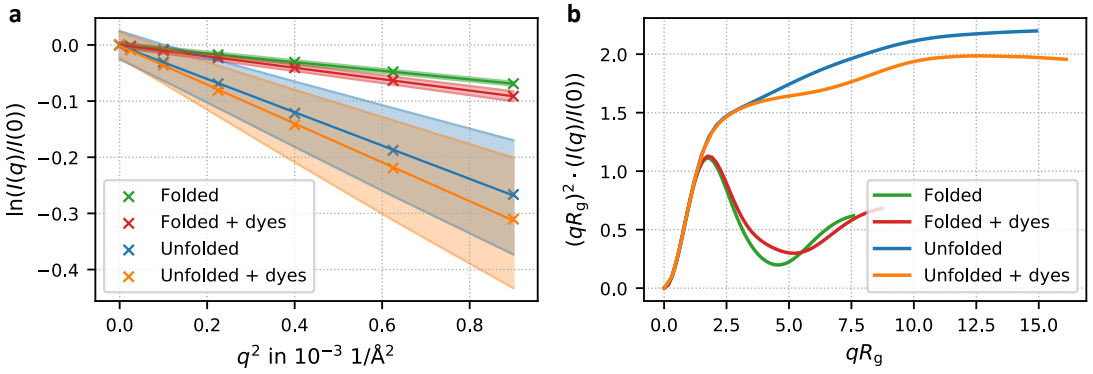


Figure 4.16. Guinier plot and dimensionless Kratky plot for ^{10}FNIII. a. The Guinier plot shows characteristic SAXS profiles in the Guinier region (crosses) along with linear fits (solid lines). Shaded areas indicate standard deviations, i.e., intensity distribution widths, propagated from Fig. 4.15 to the Guinier representation. R_g errors from the Guinier linear regression are in the order of $10^{-2} R_g$ to $10^{-4} R_g$[29]. **b**. The dimensionless Kratky plot gives information about the protein's conformation. Reproduced from Ref.[29] under CC BY 4.0.

Influence of FRET dyes on SAXS measurements. To examine the impact of FRET dyes on SAXS measurements, we computed a representative SAXS profile from 5 000 equally distributed structures for each simulated system. At each q point, the mean and standard deviation of the 5 000 individual intensities served as the ensemble-averaged SAXS profile and a dissimilarity measure of the different conformations' intensities, respectively. We compared the SAXS-derived $R_{g,saxs}$ from a Guinier analysis of the representative SAXS profiles to the average $R_{g,gmx}$ of each simulated ensemble. Because the dyes change both a system's size and shape, the SAXS intensities of folded ^{10}FNIII with and without dyes differ considerably (see Fig. 4.15 a). For the unfolded ensemble, we observed only a small difference, resulting from the more extended chain with dyes attached. The similar curve shapes point to a minor but still visible influence of the dyes. For the larger ClyA monomer, differences in the intensities without and with dyes are almost insignificant (see Fig. 4.15 b). Derived plots for ^{10}FNIII are shown in Fig. 4.16. The Guinier approximation (see Fig. 4.16 a) is only valid in the region where $\ln(I(q)/I(0))$ versus q^2 can be fitted linearly. We extracted R_g values from the slopes, which differ for folded and unfolded

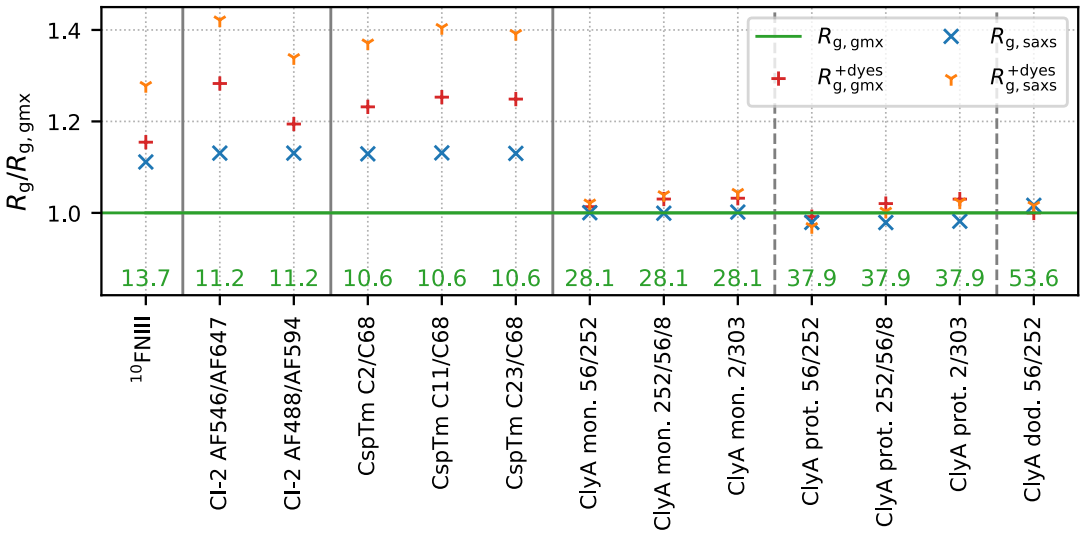


Figure 4.17. R_g variants for the folded systems with respect to $R_{g,gmx}$ (green line) given at the bottom in Å. We studied ^{10}FNIII, CI-2 with two different AF dye pairs, CspTm with AF dyes at three different labeling positions, and ClyA monomer, protomer, and dodecamer with two and three dyes at different labeling sites (see Ref.[29] for more details on the systems). R_g values calculated from atomic structures with dyes ($R^{+dyes}_{g,gmx}$, red) and SAXS-derived values without ($R_{g,saxs}$, blue) and with dyes ($R^{+dyes}_{g,saxs}$, orange) are shown. R_g errors from the Guinier linear regression can be found in the Supplementary Information of Ref.[29]. Reproduced from Ref.[29] under CC BY 4.0.

states as well as for the system with and without dyes. Fig. 4.16 gives a perfect example of how the Kratky analysis can be used to study protein folding. For the folded case, the narrower peak without dyes indicates a more compact structure compared to the dye-labeled system.

Comparison of R_g variants. Fig. 4.17 shows the ratios of all R_g variants to $R_{g,gmx}$ for the folded systems. As expected, dye-labeled systems appear to be larger, i.e., $R^{+dyes}_{g,gmx} > R_{g,gmx}$. The smaller the system, the more significant this effect is. With $R^{+dyes}_{g,saxs} > R_{g,saxs}$, SAXS-derived R_g show a similar shift. The only exception is ClyA protomer with AF dyes at positions 56/252, where $R^{+dyes}_{g,gmx} < R_{g,gmx}$ due to a dye-induced center-of-mass shift in favor of a reduced R_g (see Fig. 4.14). While different dye positions have only a minor influence on R_g, dye types seem to have a more pronounced effect. Apparently, SAXS-derived variants overestimate R_g for small systems, while underestimating it occasionally for larger systems. For the smaller proteins ^{10}FNIII, CI-2, and CspTm, SAXS-derived R_g are consistently larger than their respective references directly calculated from structural models (see Eq. 4.19). This could be caused by explicitly considering the solvation shell when calculating scattering profiles from structural models or result from neglecting hydrogen atoms in the molecular model. For ClyA monomer and protomer, however, all values are very close. SAXS-derived R_g are similar to or slightly smaller than corresponding $R_{g,gmx}$. This may result from a greater error in the Guinier analysis due to a very narrow Guinier region for elongated conformations (see Sec.3.1).

Fig. 4.18 shows R_g variants for the unfolded systems. For CI-2 and CspTm, both dye types and positions impact $R_{g,gmx}$, indicating a subtle influence of the dyes on the chain dynamics. Again, $R^{+dyes}_{g,gmx}$ is consistently larger than $R_{g,gmx}$. The observed shift increases with the dyes' distance in the protein sequence. The more peripheral the labeling positions in a sequence, the more the dyes and linkers extend a system's dimensions as reflected by R_g, especially for fully elongated unfolded conformations. While SAXS-derived R_g of the smaller CI-2, CspTm, and ^{10}FNIII are consistently larger than their respective $R_{g,gmx}$ references as for the folded case, $R_{g,saxs}$ and $R^{+dyes}_{g,saxs}$ are nearly identical to $R_{g,gmx}$ and $R^{+dyes}_{g,gmx}$ for the larger ClyA.

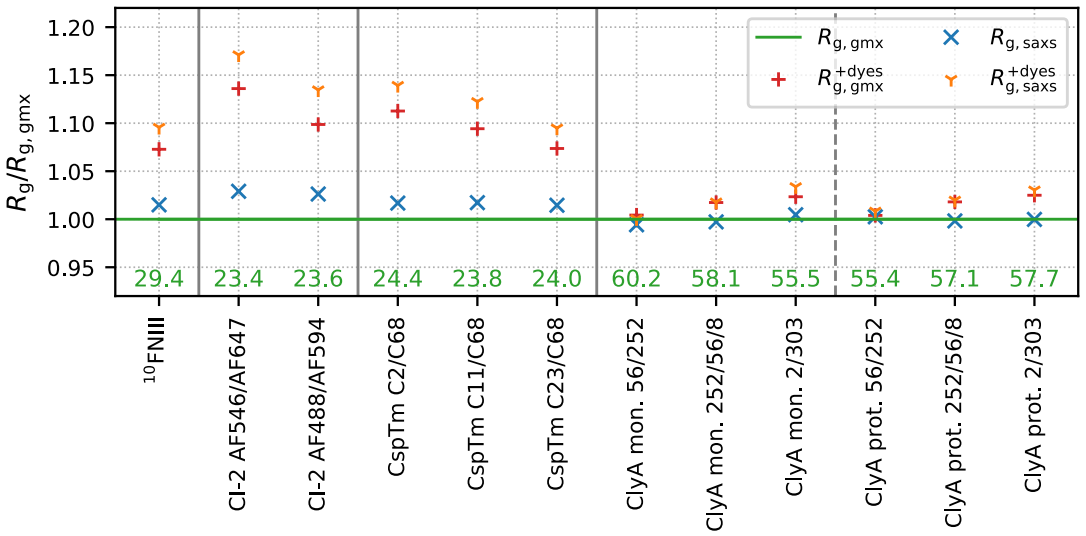


Figure 4.18. R_g variants for the unfolded systems with respect to $R_{g,gmx}$ (green line) given at the bottom in Å. We studied ^{10}FNIII, CI-2 with two different AF dye pairs, CspTm with AF dyes at three different labeling positions, and ClyA monomer and protomer with three dyes at different labeling sites (see Ref.[29] for more details). R_g calculated from atomic structures with dyes ($R^{+dyes}_{g,gmx}$, red) and SAXS-derived R_g without ($R_{g,saxs}$, blue) and with dyes ($R^{+dyes}_{g,saxs}$, orange) are shown. R_g errors from the Guinier linear regression can be found in the Supplementary Information of Ref.[29]. Reproduced from Ref.[29] under CC BY 4.0.

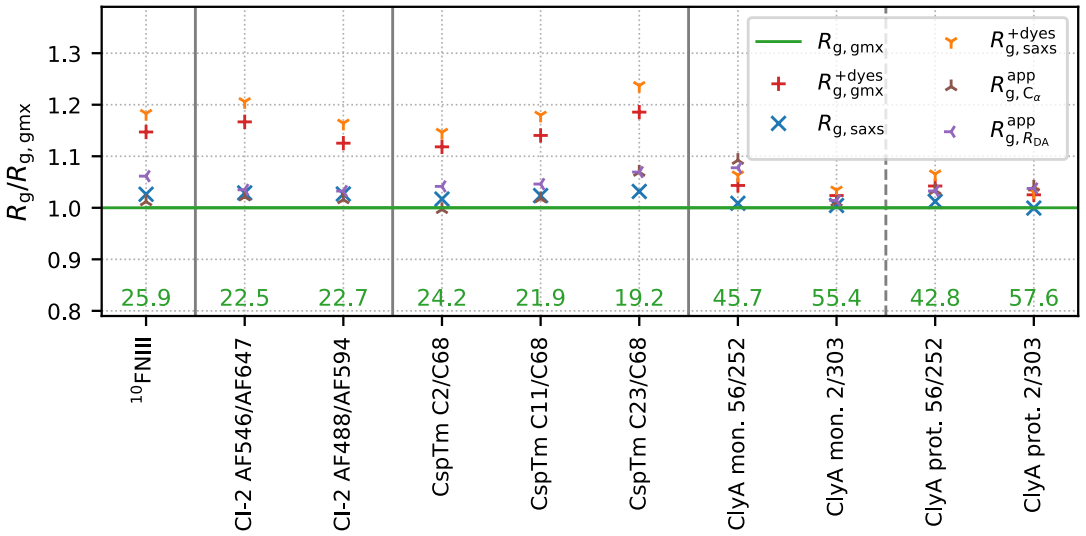


Figure 4.19. R_g variants for different truncated systems in the unfolded states with respect to $R_{g,gmx}$ (green line) given at the bottom in Å. We studied ^{10}FNIII, CI-2 with two different AF dye pairs, CspTm with AF dyes at three different labeling positions, and ClyA monomer and protomer with AF dyes at different labeling sites. More details on the systems can be found in Ref.[11]. R_g values calculated from atomic structures with dyes ($R^{+dyes}_{g,gmx}$, red), those derived from SAXS curves without ($R_{g,saxs}$, blue) and with dyes ($R^{+dyes}_{g,saxs}$, orange), and apparent values calculated from C_α end-to-end distance (R^{app}_{g,C_α}, brown) and inter-dye distance (R^{app}_{DA}, purple) are shown. R_g errors derived from the Guinier linear regression are listed in Ref.[29]. Reproduced from Ref.[29] under CC BY 4.0.

Finally, we analyzed R_g variants obtained from end-to-end distances R_e between dye-labeled residues in the truncated systems (see Fig. 4.19). As before, the ratios of $R_{g,gmx}^{+dyes}$ and $R_{g,saxs}^{+dyes}$ with respect to $R_{g,gmx}$ are in good agreement and shifted to higher values (also see Fig. 4.18). $R_{g,saxs}$, R_{g,C_α}^{app}, and $R_{g,R_{DA}}^{app}$ are all very similar to $R_{g,gmx}$. $R_{g,R_{DA}}^{app}$ is consistently larger than $R_{g,saxs}$, suggesting a subtle systematic difference in the quantities accessible to FRET and SAXS.

Discussion

Attaching FRET dyes to a protein affects concomitant SAXS measurements considerably as the dyes change both its size and shape. This effect is particularly pronounced for small systems. In the unfolded case, only a small difference is observable, which becomes almost insignificant for larger systems. As evident from their larger R_g, dye-labeled systems appear to be larger than the plain proteins, especially for small systems. Dye types also influence R_g. For the unfolded ensembles, we observe a subtle influence of the dyes on the chain dynamics, where their positions also affect derived values. This has to be taken into account in the data analysis when combining FRET and SAXS measurements. SAXS-derived variants overestimate R_g for small systems while underestimating it slightly for some of the larger systems. To conclude, we find FRET- and SAXS-derived R_g values to agree well. This is consistent with prior work suggesting the observed discrepancies to be primarily founded in the analysis methods[127,128]. However, the FRET-derived R_g appear to be generally larger than the SAXS-derived ones, suggesting a subtle yet fundamental difference in the quantities accessible in FRET and SAXS.

"Never, I said never, compare with experiment."

Magnus Bergh

5

Data-Assisted Protein Simulations

This chapter introduces the concept of data-assisted protein simulations, along with sophisticated state-of-the-art techniques for interpreting SAXS data. Understanding protein function on the molecular level requires understanding their conformational ensembles, which can be observed using SAXS. Data-assisted MD is the most powerful tool for accessing the low-resolution information in SAXS intensities to obtain atomic models of the encoded structure. Various frameworks to integrate scattering data into compute-intensive explicit-solvent simulations exist. Here, I present my `XSBM` *method for rapid interpretation of SAXS data to derive structural models using minimal computational resources and time. I combine the data with all-atom structure-based models, which are computationally efficient yet realistically describe the systems' dynamics.*

Nanoscale proteins can only be observed indirectly. As a consequence, experimental data are often ambiguous, incomplete, or of such a low resolution that they require interpretation to access their limited information content. A practical example is SAXS, an increasingly accurate method for obtaining information on biomolecular structures, ensembles, and dynamics at quasi physiological conditions (see Sec. 3.1). Experimentally, the X-ray scattering intensity of dissolved proteins is recorded for small scattering angles. The molecular electron density, however, is the Fourier transform of the inaccessible complex-valued scattering amplitude. Recovering structural models from such an intensity, i.e., the absolute amplitude squared, is an ill-posed inverse problem. As the sparse information in the experimental data is insufficient to determine all degrees of freedom, the reconstruction of 3D protein structures heavily depends on combining experimental results and computational methods. In order to best possibly access the information in the data, it must be complemented by physical, stereochemical, or structural knowledge to avoid overfitting during structure determination and refinement. Physical models of different complexity, accuracy, predictive power, and computational cost have been used,

ranging from rigid-body models[146,147] to flexible all-atom MD force fields[12,18]. Choosing an appropriate physical model mostly depends on the information content of the data. The less information in the data, the more predictive physical models are needed[148]. Approaches to complementing experimental data with physical models subdivide into three groups[148]:

i. Following a sequential "sample and select" strategy, experimental data can be compared with back-calculated data from unbiased biomolecular simulations. This requires sufficiently accurate physical models to generate a proper structural ensemble.

ii. Unbiased Boltzmann sampling can be used to propose an approximate ensemble of candidate structures, which is reweighted a posteriori in a statistically meaningful way such that it reproduces the experimental data. As the unbiased ensemble must already contain all relevant states, ensemble reweighting requires exhaustive sampling, and computationally efficient coarse-grained models are typically used.

iii. The most effective approach for interpreting low-information experimental data is to complement the data with molecular dynamics[12–14]. The data are incorporated into a computational description of the protein's physical motions, whereby the simulations are restrained to conformations that comply with both the data and the physical model.

Data-assisted MD has become very popular and many recent studies highlight the potential of combining experimental and simulation expertise[16,17,27,149]. Typically, a biasing energetic restraint on the target data, V_{data}, is added to the force field, V_{MD}, to favor conformations consistent with the data,

$$V = V_{\text{MD}} + V_{\text{data}} \text{ with } V_{\text{data}} \propto \chi^2 \,. \tag{5.1}$$

The bias potential is proportional to the least-squares deviation of theoretical data back-calculated from simulated structures and the experimental data, χ^2[12,18,27,150,151]. Derived forces are assumed to drive the molecular system towards conformations reproducing the target data. As molecular systems seek to minimize their free energy within the force field, the bias effectively determines a cost for disregarding the data in the simulation. The better the simulated structures align with the data, the smaller the energetic penalty, or bias, is. The data can be interpreted while retaining the physicochemical knowledge and sampling power of the force field. In return, the simulation can overcome energetic barriers associated with large-scale conformational changes. Thermal ensembles of dissolved proteins can be sampled while simultaneously having regard to the structural information in the data, which may even help to compensate possible shortcomings of the physical model[148].

5.1 State of the Art: Interpreting Solution X-Ray Scattering of Proteins in Explicit-Solvent MD

In the following, I describe a state-of-the-art method for interpreting solution X-ray scattering data within biomolecular simulations. This method was developed in the Computational Biophysics group of Jochen Hub, Saarland University. Many similar approaches exist. Here, I exemplarily present the Hub method as one of the most consistent and accurate approaches to refining protein structures towards solution scattering data.

Calculating Solution X-Ray Scattering from Explicit-Solvent MD

This section is based on the Biophysical Journal article "Validating Solution Ensembles from Molecular Dynamics Simulation by Wide-Angle X-ray Scattering Data" (2014) by Po-chia Chen and Jochen Hub[152]. It introduces a method for calculating X-ray scattering profiles from explicit-solvent MD simulations, which the authors use in their refinement protocol for computing scattering patterns of simulated structures at runtime.

From a computational simulation perspective, calculating accurate scattering patterns from atomic positions is a key factor for interpreting SAXS data. One drawback of the computationally efficient and widely used implicit-solvent methods, such as `CRYSOL`[41] (see Supplementary Sec. A.2), is their dependence on at least two free parameters, associated with the solvation shell and the excluded volume. As these parameters are experimentally inaccessible, they have to be adjusted by fitting computed spectra to the experimental data. Reducing the available information in this way increases the risk of overfitting. Chen and Hub developed a method for calculating scattering profiles from explicit-solvent MD with only one fitting parameter related to experimental uncertainties[152]. They use a pure-water simulation to account for excluded-solvent effects[152]. In order to evaluate profiles from heterogeneous ensembles, they define the molecular solvation shell as a shaped envelope enclosing all conformational states of the protein as well as the solvation layer itself.

The central quantity in solution X-ray scattering is the excess scattering of the solution with respect to the solvent,

$$I(q) = \underbrace{I_A(q)}_{\text{solution}} - \underbrace{I_B(q)}_{\text{pure solvent}} . \tag{5.2}$$

I_A is the intensity of the solution, I_B that of the pure solvent. q is the absolute value of the directed momentum transfer $\mathbf{q}$. In the low-dilution regime, a scattering experiment can be modeled by a single solute molecule in a water droplet, referred to as system A[152]. System B is the droplet without solute. Their respective intensities read (also see Supplementary Sec. A.1)

$$I_A(q) = \left\langle |A_A(\mathbf{q})|^2 \right\rangle' \quad \text{and} \quad I_B(q) = \left\langle |A_B(\mathbf{q})|^2 \right\rangle' . \tag{5.3}$$

$A_A(\mathbf{q})$ and $A_B(\mathbf{q})$ are the Fourier transforms of the electron densities $\rho_A(\mathbf{r})$ and $\rho_B(\mathbf{r})$ with position vector $\mathbf{r}$, respectively. $\langle \cdots \rangle'$ is the ensemble average over possible solute rotations and conformational fluctuations in each system[152],

$$\langle \cdots \rangle' = \left\langle \langle \cdots \rangle^{(\omega)} \right\rangle_\Omega . \tag{5.4}$$

$\langle \cdots \rangle_\Omega$ is the spatial average over all solute orientations, $\langle \cdots \rangle^\omega$ is the average over solute and solvent fluctuations at fixed solute orientation ω. Thus:

$$I(q) = \langle D(\mathbf{q}) \rangle_\Omega \quad \text{with} \quad D(\mathbf{q}) := \left\langle |A_A(\mathbf{q})|^2 \right\rangle^{(\omega)} - \left\langle |A_B(\mathbf{q})|^2 \right\rangle^{(\omega)} \tag{5.5}$$

To compute $D(\mathbf{q})$ from a simulation, Chen and Hub construct an auxiliary envelope surrounding the solute from an icosphere[152]. It embeds all conformational states of the solute with sufficient distance and remains constant while averaging over the system's fluctuations at fixed ω. The electron densities

thus can be split into contributions inside and outside the envelope, ρ^{i} and ρ^{o}, respectively[152]:

$$\begin{aligned}\rho_A(\mathbf{r}) &= \rho_A^{\mathrm{i}}(\mathbf{r}) + \rho_A^{\mathrm{o}}(\mathbf{r}) \\ \rho_B(\mathbf{r}) &= \rho_B^{\mathrm{i}}(\mathbf{r}) + \rho_B^{\mathrm{o}}(\mathbf{r})\end{aligned} \tag{5.6}$$

Furthermore, the authors assume that (i) the average electron densities outside the envelope, ρ_A^{o} and ρ_B^{o}, are equal, (ii) the density in the pure-solvent system is homogeneous, and (iii) density correlations between inside and outside are only relevant near the envelope's surface[152]. This allows them to express $D(\mathbf{q})$ via only densities inside the envelope and thus to calculate it from a finite MD simulation[152]:

$$\begin{aligned}D(\mathbf{q}) = &\left\langle \left| A_A^{\mathrm{i}}(\mathbf{q}) \right|^2 \right\rangle^{(\omega)} - \left\langle \left| A_B^{\mathrm{i}}(\mathbf{q}) \right|^2 \right\rangle^{(\omega)} \\ &+ 2\mathrm{Re}\left[- \left\langle \left(A_B^{\mathrm{i}}(\mathbf{q}) \right)^* \right\rangle^{(\omega)} \left\langle A_A^{\mathrm{i}}(\mathbf{q}) - A_B^{\mathrm{i}}(\mathbf{q}) \right\rangle^{(\omega)} \right]\end{aligned} \tag{5.7}$$

The first and second term are the intensities from atoms inside the envelope in systems A and B, respectively. The third term is the correlation between the bulk water outside the envelope and the density contrast between systems A and B inside the envelope[152]. With $N_{A/B}$ atoms inside the envelope, a simulated structure's scattering amplitude is calculated from its atomic coordinates via

$$A_{A/B}^{\mathrm{i}}(\mathbf{q}) = \sum_{j=1}^{N_{A/B}} f_j(q) \exp\left(-i\mathbf{q}\cdot\mathbf{r}_j\right). \tag{5.8}$$

$f_j(q)$ is the Cromer-Mann parameterized form factor and $\mathbf{r}_j$ the position vector of atom j. Spherical averaging to calculate $I(q)$ from $D(\mathbf{q})$ is done numerically. Most of the computational effort is required for calculating the scattering amplitudes, where the cost per frame scales quadratically with N_A. As an example, averaging over 1000 frames of a lysozyme simulation required approximately 12 minutes on a 16-core server node[152]. Reasonably converged intensities can be obtained by averaging over $\mathcal{O}(100)$ frames, resulting in $\mathcal{O}(\mathrm{min})$ intensity calculations on a modern desktop computer. Chen and Hub validated their method by computing small- and wide-angle X-ray scattering (SWAXS) profiles for five proteins[152]. Implemented in a modified version of `GROMACS 4.6`, it serves as a key ingredient of their SWAXS refinement protocol presented below[12].

Integrating Solution X-Ray Scattering into Explicit-Solvent MD

This section is based on the Biophysical Journal article "Interpretation of Solution X-Ray Scattering by Explicit-Solvent Molecular Dynamics" (2015) by Po-chia Chen and Jochen Hub[12]. It introduces their SWAXS-driven MD protocol for interpreting solution scattering data within explicit-solvent MD, which builds on their method for calculating SWAXS profiles described above.

Solution X-ray scattering and MD are complementary approaches to studying the conformational dynamics of dissolved biomolecules. To take advantage of both methods, Chen and Hub include SWAXS data into explicit-solvent MD via a differentiable energy restraint that guides the system into conformations consistent with the data[12]. Accurate calculation of scattering patterns from simulated frames at runtime is essential for such a hybrid approach. Interpreting solution scattering data within computational simulations is not only complicated by the low information content of the data but also by scattering contributions from the solvation shell and unknown systematic errors[148]. Compared to the various and widely used implicit-solvation methods, SWAXS-driven MD uses explicit water and thus provides a more detailed description of solvation, however at significantly higher computational costs[12,152]. The simulations are coupled to a target intensity, I_{exp}, via a hybrid potential, $V = V_{\mathrm{MD}} + V_{\mathrm{SWAXS}}$[12].

V_{MD} is the physico-chemical MD potential, or force field, and V_{SWAXS} an energetic penalty for simulated structures that are inconsistent with the target data. Chen and Hub quantify the discrepancy between the simulated and the target curves by either a non-weighted set of harmonic potentials on a logarithmic scale, $V_{\mathrm{SWAXS}}^{(\log)}$, or a weighted linear form, $V_{\mathrm{SWAXS}}^{(\mathrm{w})}$ [12]:

$$V_{\mathrm{SWAXS}}^{(\log)} = \alpha\left(t\right) k_c \frac{k_{\mathrm{B}}T}{n_q} \sum_{i=1}^{n_q} \left[\log\, I_{\mathrm{calc}}\left(q_i, t\right) - \log\, I_{\mathrm{exp}}\left(q_i\right)\right]^2 \tag{5.9}$$

$$V_{\mathrm{SWAXS}}^{(\mathrm{w})} = \alpha\left(t\right) k_c \frac{k_{\mathrm{B}}T}{n_q} \sum_{i=1}^{n_q} \frac{\left[I_{\mathrm{calc}}\left(q_i, t\right) - I_{\mathrm{exp}}\left(q_i\right)\right]^2}{\sigma^2\left(q_i\right)} \tag{5.10}$$

I_{calc} is the scattering intensity computed at runtime from recent simulation frames, n_q is the number of considered q points, and k_c specifies an empirically chosen weight of the experimental data relative to the force field. The time-dependent function $\alpha\left(t\right)$ allows to gradually ramp up the SWAXS-derived potential in the beginning of a refinement. With errors $\sigma\left(q_i\right)$, only $V_{\mathrm{SWAXS}}^{(\mathrm{w})}$ can account for experimental, statistical, and systematic uncertainties. As described before, the difference in scattering intensities between the dissolved solute (system A) and an independent pure-buffer simulation (system B) is calculated as

$$I_{\mathrm{calc}}\left(q, t\right) = \left\langle \left\langle \left|A_A\left(\mathbf{q}\right)\right|^2 \right\rangle_{t;\tau}^{(\omega)} - \left\langle \left|A_B\left(\mathbf{q}\right)\right|^2 \right\rangle^{(\omega)} \right\rangle_{\Omega} . \tag{5.11}$$

Relying on explicit water, solvent fluctuations must be temporally averaged over multiple frames to calculate a converged I_{calc} [12]. The time average is computed from previous frames at fixed solute orientation ω using exponentially decaying weights,

$$\left\langle \left|A_A\left(\mathbf{q}\right)\right|^2 \right\rangle_{t;\tau}^{(\omega)} \propto \int_0^t \left|A_A\left(\mathbf{q}, t'\right)\right|^2 \exp\left(\frac{t'-t}{\tau}\right) dt' , \tag{5.12}$$

and thus is dominated by the most recent fluctuations [12]. The memory time τ determines the extent of fluctuations represented by I_{calc}. In this way, computed SWAXS intensities can be updated in response to structural rearrangements during the refinement.

SWAXS-derived forces are calculated analytically as the spatial derivative of V_{SWAXS} (see Eqs. 5.9 and 5.10) with respect to the solute's atomic positions [12]. As these forces also depend on previous states of the system, the energy is not strictly conserved, where significant drifts can be avoided by a tight stochastic-dynamics temperature coupling [12]. A statistical uncertainty of I_{calc} is derived from the standard deviation of the set of $D\left(\mathbf{q}_i\right)$ that is used to calculate the orientational average $\left\langle \cdots \right\rangle_{\Omega}$ (see Eqs. 5.5 and 5.11). A systematic error is propagated into the scattering amplitudes from a small uncertainty in the buffer density [12].

Chen and Hub implemented SWAXS-driven MD into their in-house version of `GROMACS 4.6`. The authors validated their method by refining atomic structures of different proteins towards artificial and experimental SWAXS data without a priori knowledge of reaction paths [12]. A major drawback is the method's heavy dependence on heuristic parameters, such as the coupling strength k_c, whose optimal choice hinges on the system's a priori unknown energy landscape [12]. One possible approach to resolving this issue is to reformulate the refinement protocol within a statistically founded Bayesian inference framework.

Bayesian-Inference Based Refinement of Protein Structures Towards Solution X-Ray Scattering

This section is based on the PLOS Computational Biology article "Bayesian refinement of protein structures and ensembles against SAXS data using molecular dynamics" (2017) by Roman Shevchuk and Jochen Hub[150]. It presents one approach to resolving the problem of balancing experimental information with respect to the underlying physical model in data-assisted MD. In the context of my work, it shall serve as an example of a sophisticated state-of-the-art method for interpreting solution scattering data within biomolecular simulations.

Deriving protein structures from low-information scattering data requires complementing the data by a physical model. Such data-assisted methods typically rely on empirical parameters, whose choice is nontrivial and crucial to simulation performance. The key challenge is weighting the experimental information with respect to the physical model, which translates into choosing an adequate k_c in SWAXS-driven MD. One possible way is to embed the refinement into a statistically founded framework by combining Bayesian inference with all-atom MD simulations and explicit-solvent SAXS calculations[150]. This approach was proposed by Shevchuk and Hub who extended the concept of "inferential structure determination"[153] towards ensembles. The authors derive posterior distributions of protein structures in light of the SAXS data and the physical model, i.e., explicit-solvent MD, where the force field provides an accurate prior of protein structures[150]. Unknown parameters are not chosen ad hoc but estimated simultaneously with the protein structures and their relative weights in the ensemble[150]. The authors claim that their Bayesian formulation automatically balances the experimental data versus the prior physical knowledge[150]. Confidence intervals can be derived for both the refined structures and their weights as a quantitative precision measure[150]. The structural weights' posterior distribution provides a probabilistic criterion for identifying the number of conformational states required in the ensemble to explain the experimental data best[150].

Concept. The refinement problem is formulated by considering a protein adopting a small ensemble of N distinct states, such as a mixture of active and inactive or holo and apo conformations. The aim is to derive these states' coordinates, $\mathbf{R} = (\mathbf{R}_1, ..., \mathbf{R}_N)$, together with their relative weights, $\mathbf{w} = (w_1, ..., w_N)$, from given SAXS data[150]. In this context, an "ensemble" thus is not the usual thermodynamic ensemble but a specific set $(\mathbf{R}, \mathbf{w})$. With the number of independent data points in a scattering curve being much smaller than a protein conformation's number of degrees of freedom, many ensembles are compatible with the data. Shevchuk and Hub formulate a conditional probability, $p(\mathbf{R}, \mathbf{w}, \theta|D, K)$, that quantifies the plausibility of a specific ensemble $(\mathbf{R}, \mathbf{w})$ in light of the data D and the prior physical knowledge K. θ is the set of less interesting nuisance parameters that are required for evaluating the posterior distribution using Bayes' theorem[150]:

$$p(\mathbf{R}, \mathbf{w}, \theta|D, K) \propto L(D|\mathbf{R}, \mathbf{w}, \theta, K)\, \pi(\mathbf{R}|K)\, \pi(\mathbf{w}|K)\, \pi(\theta|K) \tag{5.13}$$

While $\pi(\mathbf{R}|K)$, $\pi(\mathbf{w}|K)$, and $\pi(\theta|K)$ are the prior distributions of possible protein conformations, weights, and nuisance parameters, respectively, $L(D|\mathbf{R}, \mathbf{w}, \theta, K)$ is the likelihood for measuring the data D given an ensemble $(\mathbf{R}, \mathbf{w})$ and nuisance parameters θ. As the data contain only limited information, L is a wide function of $\mathbf{R}$, and a tight prior on the protein conformations must be imposed. This translates to applying an accurate physical model, such as an unbiased MD simulation where K comprises the physical laws and the force field. $\pi(\mathbf{R}_j|K)$ thus corresponds to a Boltzmann factor of the MD potential[150]. Assuming no prior information on the weights, $\pi(\mathbf{w}|K)$ is chosen as a flat Dirichlet

distribution[150]. L is set up as[150]

$$L\left(D|\mathbf{R},\mathbf{w},\theta,K\right) \propto \exp\left[-\frac{N_{\text{indep}}}{2N_q}\sum_{i=1}^{N_q}\frac{\left[I_{\text{calc}}\left(q_i,\mathbf{R},\mathbf{w}\right)-\left(fI_{\text{exp}}\left(q_i\right)+c\right)\right]^2}{f^2\sigma_{\text{exp}}^2\left(q_i\right)+\sigma_{\text{calc}}^2\left(q_i\right)+\sigma_{\text{buf}}^2\left(q_i;\delta\rho_{\text{buf}}\right)}\right]. \tag{5.14}$$

f and c are fitting parameters and σ_{exp} and σ_{calc} are the statistical errors of experimental data and calculated intensities, respectively. The calculated intensity is a weighted average over the intensities of the N states in an ensemble, $I_{\text{calc}}\left(q_i,\mathbf{R},\mathbf{w}\right)=\sum_{j=1}^{N}w_jI\left(q_i,\mathbf{R}_j\right)$[150]. N_q and N_{indep} are the numbers of total and independent data points in the SAXS curve, respectively. Shevchuk and Hub apply a flat prior for the fitting parameters, i.e., $\pi\left(f\right)=\pi\left(c\right)=1$. A Gaussian prior is used for the buffer density mismatch, $\delta\rho_{\text{buf}}$, which is a common source of systematic error, σ_{buf}, in SAXS[12].

Taking the posterior's negative logarithm yields an expression for a hybrid energy that is typically used in structure refinement yet corrected by contributions from the priors[150],

$$\begin{aligned}E_{\text{hybrid}} &= -\beta^{-1}\ln p\left(\mathbf{R},\mathbf{w},\theta|D,K\right)\\ &= V_{\text{ff}}\left(\mathbf{R},K\right)+E_{\text{exp}}\left(\mathbf{R},\mathbf{w},\theta,D,K\right)-\beta^{-1}\ln\left[\pi\left(\mathbf{w}|K\right)\pi\left(\theta|K\right)\right],\end{aligned} \tag{5.15}$$

where β is the inverse temperature. Shevchuk and Hub take the prior of protein structures from an MD potential as $V_{\text{ff}}\left(\mathbf{R},K\right)=-\beta^{-1}\ln\pi\left(\mathbf{R}|K\right)$[150]. The experiment-related energy is derived from the likelihood as $E_{\text{exp}}=-\beta^{-1}\ln L$ and introduces an energetic penalty for SAXS curves calculated from ensembles incompatible with the data. After translating the probabilities into energies, the authors sample protein structures with Newtonian dynamics, where forces are calculated as the hybrid energy's gradient with respect to the atomic positions at fixed $\mathbf{w}$ and θ. Scattering intensities and SAXS-derived forces are calculated as described before[152]. Practically, only the ensemble $(\mathbf{R},\mathbf{w})$ is of interest which is why nuisance parameters are marginalized out[150]:

$$p\left(\mathbf{R},\mathbf{w}|D,K\right)=\int d\theta\, p\left(\mathbf{R},\mathbf{w},\theta|D,K\right) \tag{5.16}$$

While the fitting parameters f and c (see Eq. 5.14) are marginalized out analytically in the likelihood, the buffer density mismatch $\delta\rho_{\text{buf}}$ and the structural weights $\mathbf{w}$ are sampled numerically via Monte Carlo moves at fixed protein structure[150]. As the weights are normalized and have non-negative elements, the relevant space corresponds to the $(N-1)$ simplex. It is explored using accelerated Umbrella sampling[150]. Since weight vectors with elements equal to zero specify ensembles with less states, the posterior of an N-state ensemble includes all smaller ensembles automatically. It provides a probabilistic criterion for choosing the number of states required to explain the experimental data, where a smaller ensemble is plausible if the posterior peaks at the simplex' edge.

Shevchuk and Hub validated their method by refining ensembles of the periplasmic binding protein against calculated data and deriving solution ensembles of heat shock protein 90 against experimental data[150]. Even though Bayesian inference provides a rigorous route for combining data and physical models, the authors emphasize that, applied to protein structure determination, it becomes computationally expensive and technically challenging[150]. In a subsequent study, Jochen Hub and Markus Hermann developed a data-assisted method based on the maximum-entropy principle to interpret ensemble-averaged SAXS intensities of disordered systems. As explained in Supplementary Sec. D.1, they chose the bias weight as an empirical parameter[16]. The fact that the authors themselves did not use their Bayesian inference framework in further work shows how complicated this really is. Practically, such sophisticated approaches are inapplicable for scientists with a primarily experimental background. To date, SAXS data are still conveniently interpreted by obsolete ab initio reconstruction of low-resolution envelopes[154–156].

In particular with large-scale conformational changes being involved, such methods do not yield reliable results. Other outdated approaches include rigid body refinement[146,147], simulated annealing of dummy atom collections[37,155], and selection of suitable structures from biomolecular simulations[157–159]. In the following, I present the outcome of my first main project, an accessible and easy-to-use simulation method for rapidly interpreting SAXS data in terms of protein structures[27], as a promising alternative. I explore how simplistic structure-based models can be used as a framework for computationally efficient interpretation of solution scattering data within biomolecular simulations.

5.2 PROJECT: SAXS-Guided Protein Simulations Using Structure-Based Models

This section largely builds on my PLOS Computational Biology article "Rapid interpretation of small-angle X-ray scattering data" (2019)[27]. I integrate the sparse information from SAXS into all-atom structure-based simulations to rapidly produce physically reasonable protein models in agreement with the data. I demonstrate my method's performance using the example of three protein systems. Running on common workstations instead of supercomputers, my simulations are faster than regular MD approaches by more than two orders of magnitude and achieve comparable accuracy. I find that minimalist SBMs provide a suitable framework for initial refinement of protein structures towards SAXS data with special focus on computational efficiency.

5.2.1 Starting Point: Solution-Scattering Guided Molecular Dynamics

This section is based on the Journal of Chemical Theory and Computation article "Deciphering Solution Scattering Data with Experimentally Guided Molecular Dynamics Simulations" (2015) by Alexander Björling et al.[18]. It introduces their XS-guided MD method for structural interpretation of difference data from time-resolved X-ray scattering experiments. A practical advantage of processing scattering intensities in the form of differences is that the solvation shell's contribution largely cancels out and thus is negligible. That is why XS-guided MD provides a strong basis for interpreting SAXS data within structure-based simulations, which use stochastic dynamics to implicitly model solvent effects.

Time-resolved solution X-ray scattering is suited for studying the structural dynamics of molecular reactions[146,157–160]. A conformational change can be triggered by, e.g., ligand binding and the elastically scattered intensity is recorded in the small-angle regime as a function of time. To observe the molecular response on the structural level, the measured data are processed in the form of differences, where the intensity of a reference state is subtracted from that of each time point during the reaction. These curves represent differences in the protein's intramolecular distance distribution of different conformational states and thus encode structural changes. X-ray scattering-guided molecular dynamics (XS-guided MD) combines the information contained in such difference curves with the physico-chemical knowledge of molecular force fields and the sampling power of MD[18]. Compared to conventional SAXS, the solvation shell's contribution to the difference scattering signal is negligible, facilitating the computational treatment of solvent effects[18]. However, a static initial state is required for interpretation.

Concept. Similarly to the Hub method (see Sec. 5.1), a biasing pseudo energy, V_{XS}, is added to the MD potential, V_{MD}, to favor conformations reproducing the target data. This harmonic restraint is based on a direct fit of 1D intensities[18],

$$\begin{aligned} V &= V_{\mathrm{MD}} + V_{\mathrm{XS}} = V_{\mathrm{MD}} + \frac{k_\chi}{2}\chi^2 \\ \text{with } \chi^2 &= \sum_q \left[\frac{\Delta I_{\mathrm{exp}}(q) - \alpha\{I_{\mathrm{calc}}(q) - I_{\mathrm{ref}}(q)\}}{\sigma_q}\right]^2, \end{aligned} \tag{5.17}$$

where the least-squares deviation χ^2 is a dissimilarity measure of the current conformation's scattering in the simulation with respect to the target data. $\Delta I_{\mathrm{exp}}(q)$ is the target difference curve, $I_{\mathrm{ref}}(q)$ the initial structure's reference intensity, and $I_{\mathrm{calc}}(q)$ the theoretical Debye scattering intensity back-calculated from the simulation's current conformation (see Eq. 3.1). α is the fraction of the observed sample undergoing conformational change given by the relative yield of the difference experiment. k_χ and σ_q are the weighting factors for the scattering bias and each q value, respectively. The latter are calculated from experimental errors as[18]

$$\sigma_q = \frac{\sigma_\Delta(q)}{\Delta I_{\mathrm{exp}}(q)} + 1 \qquad \text{or} \qquad \sigma_q = \frac{\sigma_{\mathrm{ref}}(q)}{I_{\mathrm{ref}}(q)} + 1\,. \tag{5.18}$$

Errors in the difference data, $\sigma_\Delta(q)$, are preferred over errors in the reference curve, $\sigma_{\mathrm{ref}}(q)$. Not given any errors, all σ_q are set to 1. As the data are more and more affected by the errors with increasing scattering angle, wide-angle data are naturally given less weight compared to small-angle data.

In the simulations, the total force on the k^{th} particle is calculated as the potential's negative gradient with respect to the particle's position[18],

$$\mathbf{F}_k = -\nabla_k V = \underbrace{-\nabla_k V_{\mathrm{MD}}}_{\text{standard force field}} \underbrace{-\frac{k_\chi}{2}\nabla_k \chi^2}_{\text{scattering force}}\,. \tag{5.19}$$

The first term on the right-hand side is the standard force field. The problem thus reduces to calculating $\nabla_k\chi^2$. With constant target and reference data, $\nabla_k \Delta I_{\mathrm{exp}}(q)$ and $\nabla_k I_{\mathrm{ref}}(q)$ vanish[18]:

$$\begin{aligned} \nabla_k \chi^2 &= \sum_q \frac{1}{\sigma_q^2}\nabla_k \left[\Delta I_{\mathrm{exp}}(q) - \alpha\{I_{\mathrm{calc}}(q) - I_{\mathrm{ref}}(q)\}\right]^2 \\ &= -2\alpha \sum_q \frac{1}{\sigma_q^2}\left[\Delta I_{\mathrm{exp}}(q) - \alpha\{I_{\mathrm{calc}}(q) - I_{\mathrm{ref}}(q)\}\right]\nabla_k I_{\mathrm{calc}}(q) \end{aligned} \tag{5.20}$$

Taking the appropriate derivatives in the Debye formula (see Eq. 3.1) leads to[18]

$$\nabla_k I_{\mathrm{calc}}(q) = 2\sum_j f_k(q)\, f_j(q) \left[\cos(q r_{kj}) - \frac{\sin(q r_{kj})}{q r_{kj}}\right]\frac{\mathbf{r}_{kj}}{r_{kj}^2}. \tag{5.21}$$

f_k is the k^{th} scatterer's form factor and $r_{kj} = |\mathbf{r}_{kj}| = |\mathbf{r}_k - \mathbf{r}_j|$ with the k^{th} scatterer's position $\mathbf{r}_k$. Implementations of atomic form factors with and without displaced-solvent corrections[161,162] and library-averaged and coarse-grained scattering factors for amino acids[163] and MARTINI beads[164] are available[18]. The scattering force on the k^{th} atom is evaluated by combining Eqs. 5.19, 5.20, and 5.21[18]:

$$\begin{aligned} \mathbf{F}_k^{\mathrm{scat}} = 2k_\chi\alpha &\sum_q \frac{1}{\sigma_q^2}\left[\Delta I_{\mathrm{exp}}(q) - \alpha\{I_{\mathrm{calc}}(q) - I_{\mathrm{ref}}(q)\}\right] \\ &\cdot \sum_j f_k(q)\, f_j(q)\left[\cos(q r_{kj}) - \frac{\sin(q r_{kj})}{q r_{kj}}\right]\frac{\mathbf{r}_{kj}}{r_{kj}^2} \end{aligned} \tag{5.22}$$

The Debye scattering terms are represented explicitly by a special type of bonded interactions in the topology, where any (virtual) interaction site can be defined as a scatterer[18]. Björling et al. implemented their method as an extension of `GROMACS 5`. They validated it using three test cases and applied it to the photoconversion of a phytochrome to obtain a view of the dissolved protein from its crystal structure[18].

Even though a bias towards experimental data may accelerate conformational transitions, studying large-scale structural changes on biologically relevant time scales is still technically challenging in regular MD. A workable approach to overcome this is reducing the system's effective degrees of freedom by coarse-graining its structural representation, simplifying the force field, or both[165]. In addition, the $\mathcal{O}\left(N^2\right)$ Debye summation (see Eq. 3.1) becomes increasingly compute-intensive for large biomolecules, suggesting to also coarsen the scattering calculation.

5.2.2 PROJECT: Solution-Scattering Guided Structure-Based Simulations

In my first main project, I systematically researched how difference scattering data can be incorporated into robust SBMs, which probe the dynamics arising from a protein's native geometry[21–23,25]. While SBMs use the same bonded interactions as regular MD, the crucial difference lies in the non-bonded interactions categorized as either native, and thus favorable, or non-native, and thus unfavorable (see Sec. 4.4). Decreasing force field complexity in this way improves the sampling power without the loss of essential information on the system's characteristics. In contrast to regular MD, SBMs provide information about complex processes on biologically relevant time and length scales. They support full molecular flexibility and provide an easily extendable framework to efficiently interpret biological solution scattering data.

Fast back-calculation of SAXS intensities from structural models is a key factor for successful refinement. To reduce the computational costs to a minimum and take advantage of the intrinsic low-resolution nature of solution scattering data, I combined SBMs with Debye-based on-the-fly calculations of SAXS intensities using residue-based form factors corrected for displaced solvent[162–164]. These calculations thus do not account for the fact that the solvent density in the solvation shell differs from its bulk value. However, the solvation shell's scattering is typically several orders of magnitude smaller than the scattering of solute and excluded solvent[166]. Systematic errors and the solvation shell's contributions effectively cancel upon taking differences, and a less sophisticated solvation treatment can be used to reliably model difference data[18,164]. This is why the XS-guided MD method by Björling et al. provides a suitable basis for interpreting SAXS data within structure-based simulations, which only partially model solvent effects via stochastic dynamics (see Sec. 4.2).

Concept. To bias a simulation towards conformations reproducing a certain difference scattering curve, I choose the force field in Eq. 5.17 as the structure-based potential V_{SB}, that is,

$$V = V_{\mathrm{SB}} + V_{\mathrm{XS}} = V_{\mathrm{SB}} + \frac{k_\chi}{2}\chi^2 \,. \tag{5.23}$$

My scattering-guided structure-based "`XSBM`" simulations thus use the common SBM potential (see Eq. 4.13) extended by a scattering-derived Debye-based term (see Fig. 5.1). Note that the relative weighting factor of V_{XS} with respect to V_{SB}, k_χ, is to be specified in the structure-based reduced energy unit ε.

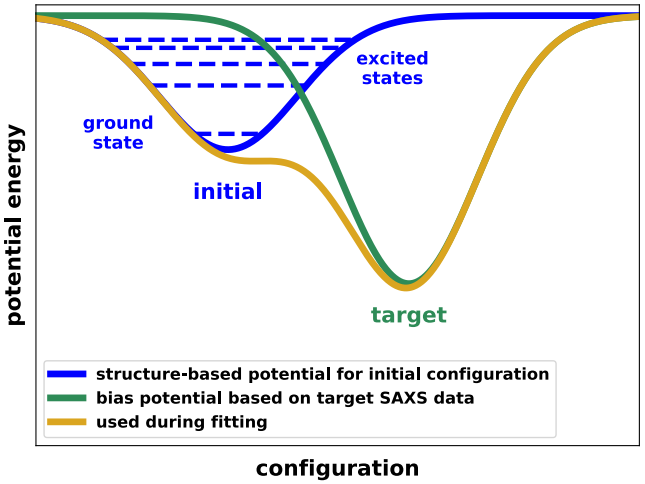


Figure 5.1. Schematic of the XSBM energy landscape. In the structure-based potential, excited and transiently populated configurations are accessible. The excited states of interest can be identified by simply capturing those in accordance with the target restraints. In order to populate excited states of the initial structure-based potential, V_{SB} (blue), an additional potential, V_{XS} (green), is introduced in due consideration of a priori experimental knowledge of the target complex. A structural model of the system's intermediate configurations is provided by the sum of the initial SBM's potential built from the native structure and the energetic bias based on the scattering data (yellow). Adapted from Ref.[13].

5.2.3 Results

To validate my XSBM method, I investigated structural transitions in three well-characterized two-state protein systems, whose experimentally determined structures can be retrieved from the PDB[6].

Villin headpiece. A basic approach to elucidating the mechanisms of protein folding and conformational transitions is to study short sequences with fast folding kinetics. Such sequences contain the complete information for folding and help to identify the basic requirements for stable folded structure. They can be obtained by reducing polypeptide length and removing complex stabilization mechanisms, such as oligomerization or ligand binding. The actin-binding protein villin consists of multiple domains capped by a C-terminal headpiece, which forms a fast and independently folding three-helix bundle stabilized by hydrophobic interactions[167]. Villin headpiece (VHP) retains full binding activity and is one of the shortest autonomously folding proteinogenic amino-acid sequences. Based on its solution-NMR structure (PDB code 1VII[167]), I set up a proof-of-principle system from the 21-amino-acid subregion between residues 54 and 74 (VHP_{54}^{74}). VHP_{54}^{74} consists of two short helices connected by a loop confining an angle of approx. 60 °, referred to as the bent conformation (see Fig. 5.2 a, gray). I constructed a corresponding elongated conformation as a perfect alpha helix with identical sequence in PyMOL[1] (see Fig. 5.2 a, colored). Bent and elongated structures have a GDT of 32.14 and a C_α RMSD of 5.5 Å. Scattering-guided simulations started from the elongated state and aimed at the bent state, and reversed. Elongated-to-bent and bent-to-elongated transitions are referred to as e → b and b → e, respectively.

Lysine-, arginine-, ornithine-binding protein. Many small ligands such as sugars, amino acids, and vitamins are actively transported into bacteria across cell membranes[168]. Dedicated transport systems comprise a receptor, that is the binding protein, and a membrane-bound protein complex. Interactions of the ligated binding protein with the membrane components induce conformational changes in the latter, thus forming an entry pathway for the ligand. I studied lysine-, arginine-, ornithine-binding (LAO) protein. This amino-acid binding protein consists of two lobes connected by short strands (see

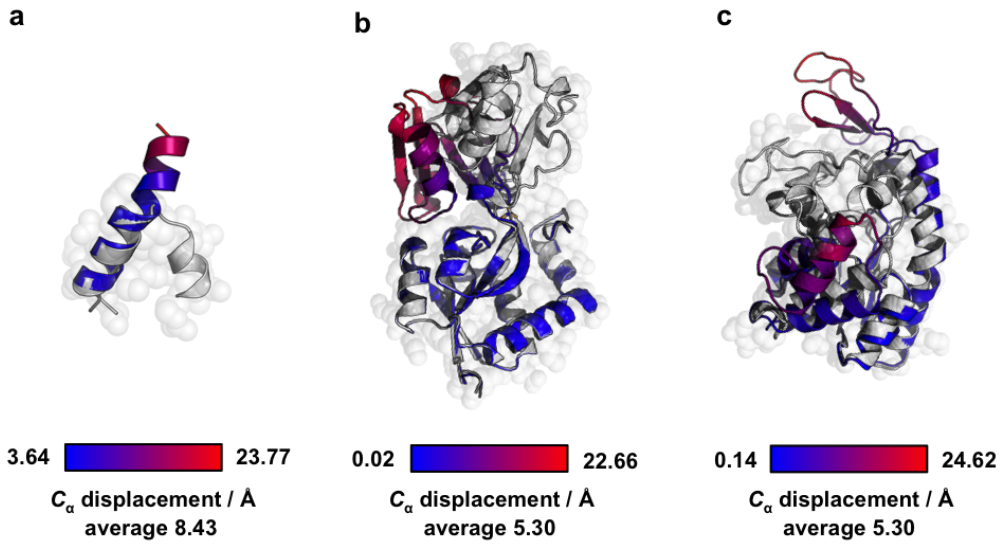


Figure 5.2. Two-state test protein systems. The coloring indicates the displacement of each C_α atom in the colored state with respect to the gray state. **a**. VHP^{74}_{54}. Elongated (colored) and bent (gray) conformations have inter-terminal distances of 10.7 Å and 26.8 Å, respectively, a GDT of 32.14, and a C_α RMSD of 5.5 Å. **b**. LAO protein. Holo (colored) and apo (gray) conformation have a GDT of 39.39 and C_α RMSD of 4.7 Å. **c**. ADK. Open (colored) and closed (gray) conformation are depicted with aligned CORE domain and have a GDT of 33.06 and a C_α RMSD of 7.1 Å. Visualized in PyMOL[1].

Fig. 5.2 b). Upon ligand binding, it undergoes a conformational change from an apo (unligated; PDB code 2LAO[81]) to a holo (ligated; PDB code 1LST[81]) state. One domain rotates, and whereas the lobes are clearly separated in the unligated state, they are in contact with a ligand bound. Both structures were determined using X-ray diffraction. Neglecting the ligand in the holo state, they have a GDT of 39.39 and a C_α RMSD of 4.7 Å. Scattering-guided simulations started from the unligated holo state and aimed at the apo state, and reversed. Holo-to-apo and apo-to-holo transitions are referred to as $\mathrm{h} \to \mathrm{a}$ and $\mathrm{a} \to \mathrm{h}$, respectively.

Adenylate kinase. Adenosine triphosphate (ATP) is the universal energy source in living cells and drives many vital processes, e.g., muscle contraction and nerve impulse propagation. By continuously checking ATP levels, adenylate kinase (ADK) provides the cell with a mechanism to dynamically measure energetic levels and monitor metabolic processes. It catalyzes the interconversion of adenine nucleotides and plays a key role in cellular energy homeostasis. The reversible transition between an open (PDB code 4AKE[169]) and closed (PDB code 1AKE[170]) state (see Fig. 5.2 c) is quintessential to its catalytic function and directly related to the competing native interactions of these states[171]. Open and closed X-ray diffraction structures have a GDT of 33.06 and a C_α RMSD of 7.1 Å. Scattering-guided simulations started from the open state and aimed at the closed state, and reversed. Open-to-closed and closed-to-open transitions are referred to as $\mathrm{o} \to \mathrm{c}$ and $\mathrm{c} \to \mathrm{o}$, respectively.

General Analysis

For each conformational transition, I derived artificial difference data from absolute SAXS intensities by subtracting the initial structure's back-calculated scattering from the target structure's back-calculated scattering. I used the GDT and the C_α RMSD to quantify the structural similarity between different conformations in a simulation (see Sec. 4.5). To evaluate a simulated ensemble's structural convergence to the target state, I calculated the trajectory's GDT and C_α RMSD with respect to the initial and

target structure as a function of simulated time, referred to as initial and target GDT and RMSD, respectively. I considered each intersection of the initial and target GDT (RMSD) a structural transition. Furthermore, I extracted the fraction of simulated time related to the trajectory approaching a target GDT equal to or greater than 75, τ_{sim}^{75}, along with the corresponding absolute core time, T_{comp}^{75}, as an efficiency estimate. To study a simulation's convergence on the data level, I considered the bias energy V_{XS} (see Eq. 5.17) versus simulated time. I calculated its Pearson correlation ρ with the target GDT and RMSD to assess its suitability as a reliable identifier of physical structures matching the target data. For each simulated ensemble, I determined the median target GDT, the median target RMSD, the maximum GDT and its corresponding RMSD, the minimum target RMSD and its corresponding GDT, and both the GDT and RMSD at minimum bias energy. While median quantities are calculated from the entire conformational ensemble and can be considered "macroscopic", minimum and maximum quantities reflect single structural snapshots and thus are "microscopic". As proteins are intrinsically dynamic, I am interested in conformational ensembles rather than in single static structures. That is why I focused on median quantities as a simulation's key performance indicators. I used the median target GDT as the main metric since it more accurately accounts for local misalignments than the more common RMSD.

For each protein system, I present results for parameter combinations (T, k_χ) of temperature and bias weight determined via grid-search variational studies, where I considered the median target GDT, target RMSD, and χ^2 deviation versus k_χ in the range of $10^{-11}\,\varepsilon$ and $10^{-7}\,\varepsilon$ at $T = 50$, 70, 90, and 110.

I conducted analogous scattering-guided explicit-solvent MD simulations for a comparative performance check. Results are presented in Appendix D. A detailed description of the simulation setups is provided in Supplementary Sec. D.2. To ensure structural conformity between the ensembles generated by SBM and explicit-solvent MD, I calculated the radius of gyration and the asphericity as functions of simulated time (see Supplementary Sec. D.3).

Villin Headpiece

I used VHP_{54}^{74} to probe how a small peptide's conformational distribution can be influenced by biasing structure-based simulations towards reproducing an artificial difference scattering curve. I calculated the target scattering data via the Debye formula using per-residue form factors corrected for displaced solvent[163] (see Eq. 3.1). To begin with, I analyzed the backbone's elongatedness by extracting the distances between N-terminal and C-terminal C_α atoms for unbiased and scattering-guided structure-based simulations. The distance distributions of the scattering-guided simulations show a clear shift towards each target structure's end-to-end distance (see Fig. 5.3), i.e., conformations not matching the target data are avoided as anticipated. The compute time scaled as 1.4 to 1 for scattering-guided and unbiased simulations. With the $\mathcal{O}\left(N^2\right)$ Debye summation, this ratio will increase with a system's number of atoms, N. This renders a rapid evaluation of SAXS profiles from structural models as employed in `XSBM` even more important.

Structural analysis. Time-dependent GDT, RMSD, and χ^2 deviation are depicted in Fig. 5.4 and Supplementary Fig. D.10 for the e → b and the b → e transition, respectively. Corresponding performance indicators are summarized in Table 5.1, along with the values from analogous explicit-solvent simulations. Biasing an SBM towards conformations reproducing the target data caused the e → b transition to readily occur (see Fig. 5.4). For the b → e direction, only one distinct transition took place and the system mostly proceeded closer to the initial state than to the target state (see Supplementary Fig. D.10). The best structures in terms of maximum target GDT are shown in Fig. 5.5 a and d. While the best median target

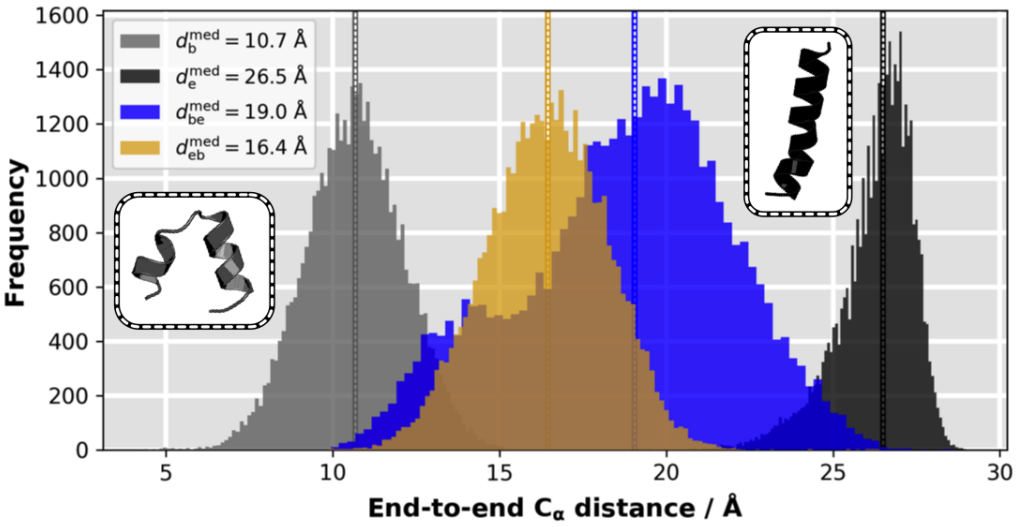


Figure 5.3. VHP$_{54}^{74}$ end-to-end C$_\alpha$ distance distributions. d_b^{med}, d_e^{med}, d_{eb}^{med}, and d_{be}^{med} are the median end-to-end distances for free bent (T = 50, grey), free elongated (T = 50, black), scattering-guided e → b (T = 50, $k_\chi = 1 \cdot 10^{-7}$, yellow), and scattering-guided b → e (T = 50, $k_\chi = 7 \cdot 10^{-8}$, blue) structure-based simulations.

GDTs (RMSDs) are 65.58 (2.55 Å) and 44.05 (4.33 Å) for the e → b and b → e transition, the maximum target GDTs (minimum target RMSDs) are 85.71 (1.59 Å) and 64.29 (2.72 Å), respectively.

These numbers suggest that the scattering-derived forces (see Eqs. 5.19 and 5.22) have a compacting effect, resulting in compacting structural changes from extended to globular states to be easier to sample than structure-opening changes. Unbiased SBM simulations of the bent and elongated state have median target GDTs (RMSDs) of 94.05 (0.86 Å) and 90.48 (0.99 Å), respectively. As geometry-derived SBMs are strongly biased towards their respective native state, VHP$_{54}^{74}$ with its globally changing tertiary structure is a difficult test case. Even though the `XSBM` simulations did not reproduce each target structure with the method's inherent accuracy for this system, they could sample the conformational transitions reversibly and generate physically reasonable structures near each target conformation. All tendencies described for `XSBM` could also be observed in analogous explicit-solvent runs (see Supplementary Figs. D.13 and D.15 for e → b and b → e transition, respectively). I find my `XSBM` method to be faster than the regular MD approach by more than two orders of magnitude in terms of compute times T_{comp}^{75} (see Table 5.1). This underlines the excellent price-performance ratio of simplistic knowledge-based SBMs with regard to structural accuracy and computational efficiency.

Correlation analysis. I analyzed the mutual correlations among target GDT, target RMSD, and V_{XS}. Desired target-like structures matching the data have a large target GDT (low target RMSD) and a small V_{XS}. The scalar V_{XS}, or equally, the χ^2 deviation of simulated and target data (see Eq. 5.17), provides a reduced representation of a structure's agreement with the low-information scattering data and thus the target state. Minimizing χ^2 against the backdrop of a biased force field is assumed to produce physically reasonable structures in accordance with the data. That is why I certainly expect a negative (positive) correlation of V_{XS} with the target GDT (RMSD). It is important to note that the Pearson correlation only reflects the degree of linearity in the relationship of two quantities, and I anticipate the inherent ambiguity in the SAXS data to hinder perfectly linear relations. As can be seen in Fig. 5.6, a low bias potential is indeed associated with a high target GDT and low target RMSD. A Pearson correlation of -0.23 (0.47) indicates that they are in fact suitably correlated. As expected, I find a wide spread of different structures at equal V_{XS} levels, suggesting the bias potential by itself

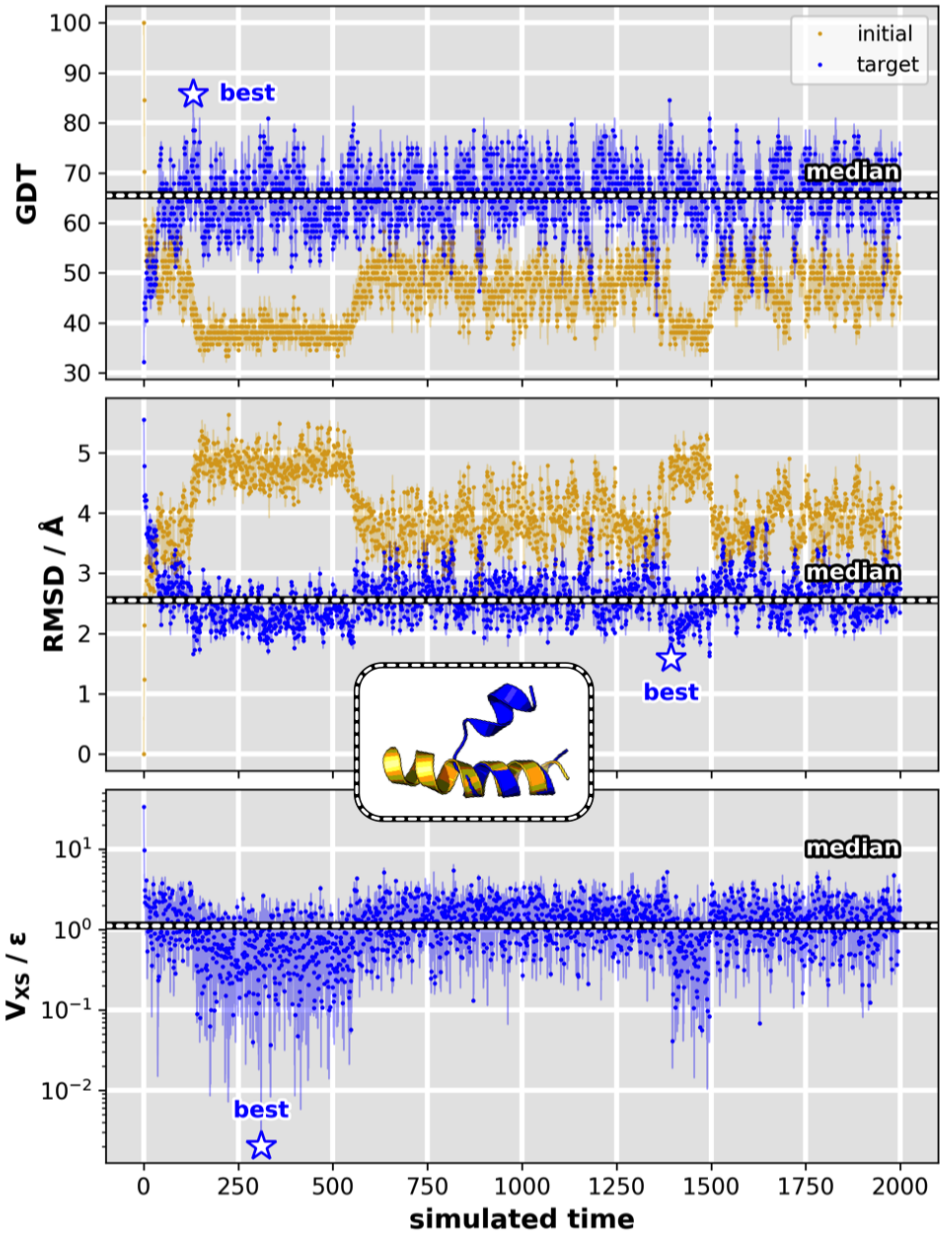


Figure 5.4. `XSBM` **simulation results for VHP$^{74}_{54}$ e → b transition.** Results are shown for parameters $(T,\ k_\chi) = \left(50,\ 1 \cdot 10^{-7}\,\varepsilon\right)$. Initial and target GDT (top), initial and target RMSD (middle), and bias energy (bottom) versus simulated time. Maximum GDT, minimum RMSD, and minimum V_{XS} are marked by a star.

to be insufficient to identify physically reasonable structures. This once more highlights how crucially the interpretation of ambiguous experimental data depends on the complementation with prior physical knowledge. Analogous explicit-solvent MD results can be found in Supplementary Figs. D.14 and D.16.

Variational grid search. To examine the influence of temperature and bias weight, I conducted grid-search variational studies where I considered the median target GDT, the median target RMSD, and the median χ^2 deviation of simulated and target data versus k_χ at $T = 50$, 70, 90, and 110. Results are shown in Fig. 5.7 and Supplementary Fig. D.12 for the e → b and the b → e transition, respectively. As soon as χ^2 (bottom) dropped down, the median target GDT (top) and RMSD (middle) started to improve steadily towards their respective bests (see Table 5.1). Near this major drop, I assume the

Table 5.1. Results for scattering-guided simulations of $\mathbf{VHP_{54}^{74}}$. k_χ, bias weight, T, temperature, $\mathrm{GDT}^{\mathrm{med}}$, median target GDT, $\mathrm{RMSD}^{\mathrm{med}}$, median target RMSD, $\mathrm{GDT}^{\mathrm{max}}$, maximum target GDT, $\mathrm{RMSD}\left(\mathrm{GDT}^{\mathrm{max}}\right)$, corresponding target RMSD, $\mathrm{RMSD}^{\mathrm{min}}$, minimum target RMSD, $\mathrm{GDT}\left(\mathrm{RMSD}^{\mathrm{min}}\right)$, corresponding target GDT, $\mathrm{GDT}_{\mathrm{free}}^{\mathrm{med}}$, median GDT of an unbiased simulation of each target, $\mathrm{RMSD}_{\mathrm{free}}^{\mathrm{med}}$, median RMSD of an unbiased simulation of each target, $\mathrm{GDT}\left(V_{\mathrm{XS}}^{\mathrm{min}}\right)$, target GDT associated with minimum bias potential, $\mathrm{RMSD}\left(V_{\mathrm{XS}}^{\mathrm{min}}\right)$, target RMSD associated with minimum bias potential, τ_{sim}^{75}, fraction of simulated time associated with a target GDT greater than 75, T_{comp}^{75}, corresponding absolute core time.

method	**SBM**		**MD**	
transition	**e → b**	**b → e**	**e → b**	**b → e**
k_χ	$1 \cdot 10^{-7}\,\varepsilon$	$7 \cdot 10^{-8}\,\varepsilon$	$5 \cdot 10^{-9}$ kJ/mol	
T	50	50	330 K	
$\mathrm{GDT}^{\mathrm{med}}$	65.48	44.05	79.76	40.48
$\mathrm{RMSD}^{\mathrm{med}}$ / Å	2.55	4.33	1.88	4.98
$\mathrm{GDT}^{\mathrm{max}}$	85.71	64.29	97.62	71.43
$\mathrm{RMSD}\left(\mathrm{GDT}^{\mathrm{max}}\right)$ / Å	1.66	2.72	0.95	2.36
$\mathrm{RMSD}^{\mathrm{min}}$ / Å	1.59	2.72	0.71	2.31
$\mathrm{GDT}\left(\mathrm{RMSD}^{\mathrm{min}}\right)$	80.96	64.29	95.24	70.24
$\mathrm{GDT}_{\mathrm{free}}^{\mathrm{med}}$	94.05	90.48	72.62	65.48
$\mathrm{RMSD}_{\mathrm{free}}^{\mathrm{med}}$ / Å	0.86	0.99	2.82	2.36
$\mathrm{GDT}\left(V_{\mathrm{XS}}^{\mathrm{min}}\right)$	67.86	48.81	84.52	48.81
$\mathrm{RMSD}\left(V_{\mathrm{XS}}^{\mathrm{min}}\right)$ / Å	2.10	4.00	1.87	3.95
τ_{sim}^{75} (fraction)	0.02	-	0.34	-
T_{comp}^{75} (absolute)	12 s	-	21 h 29 min	-

Table 5.2. Maximum target GDT structures from best median target GDT simulations.

system	$\mathbf{VHP_{54}^{74}}$		**LAO protein**		**ADK**	
transition	**e → b**	**b → e**	**h → a**	**a → h**	**o → c**	**c → o**
target GDT	85.71	64.29	93.38	92.96	83.30	84.35
target RMSD / Å	1.66	2.72	0.94	0.91	1.65	1.30

structure-based potential and the scattering bias to be thoroughly balanced. To facilitate rapid conformational transitions according to the target data in an SBM, the scattering bias must be weighted so as to introduce a competing minimum to the single-basin energy funnel. At the same time, choosing k_χ as the minimum bias weight that yields satisfactory χ^2 reduces the risk of overfitting. This allows a thorough sampling of the conformational transition while modifying the underlying geometry-derived potential as little as possible. The global change in orientation of secondary structure affects VHP_{54}^{74}'s overall shape significantly. Thus, comparably large k_χ were required to attach sufficient importance to the information from SAXS and enable sampling of conformations near the target state within the biased energy landscape. The course of χ^2 versus k_χ is very similar for both directions of the conformational transition. In the e → b case, χ^2 proceeded on a slightly lower level, suggesting this simulation to reproduce the data more accurately. This confirms my previous finding of compacting structural transitions to be favored over opening ones.

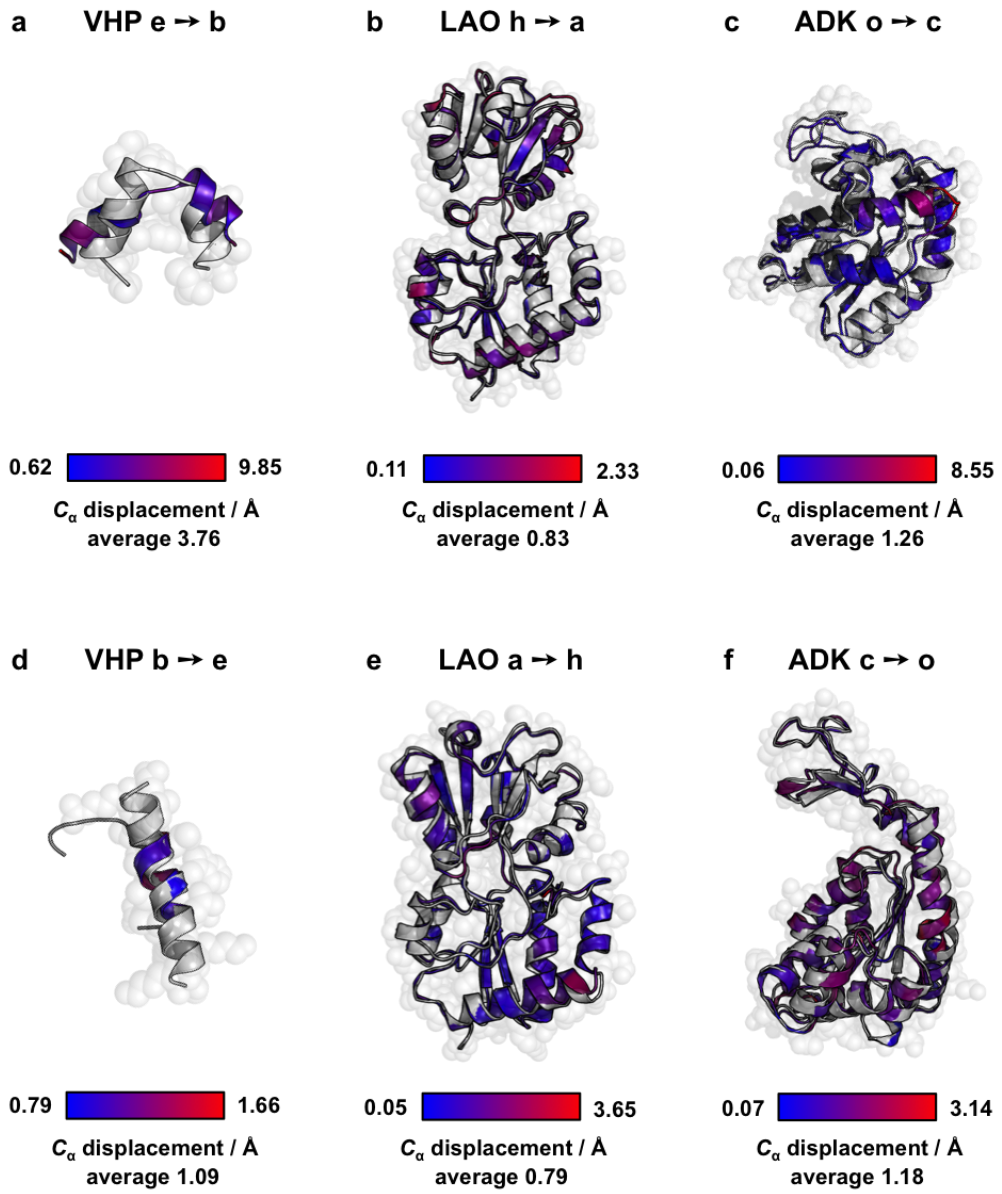


Figure 5.5. Representative structures from best median target GDT simulations. Maximum target GDT structures from each `XSBM` simulation with best median target GDT (see Tables 5.1, 5.3, and 5.4) are colored according to the C_α displacement with respect to the target state (gray). For each system, the maximum GDT structure's target GDT and target RMSD are listed in Table 5.2.

For both transitions, I find the effect of gradually increasing k_χ to be less distinct at higher temperatures. The increased thermal energy allows the protein to overcome barriers in the energy landscape more easily, resulting in greater structural flexibility and sampling power. Thermal fluctuations are isotropic in the conformational space and not directed towards any particular conformation as is the case with the scattering bias. With increasing temperature, the data thus could be reproduced more accurately already at smaller bias weights at the cost of lower structural similarity to the target state at larger k_χ. Because of the dynamic nature of proteins, this may even be useful in some cases.

Even though a basic trend should be preserved, these findings do not directly translate to other proteins. A suitable choice of T and k_χ depends on the system and should be determined by a systematic optimization method. In SBMs, the overall contact and dihedral energy is set to the system's number of

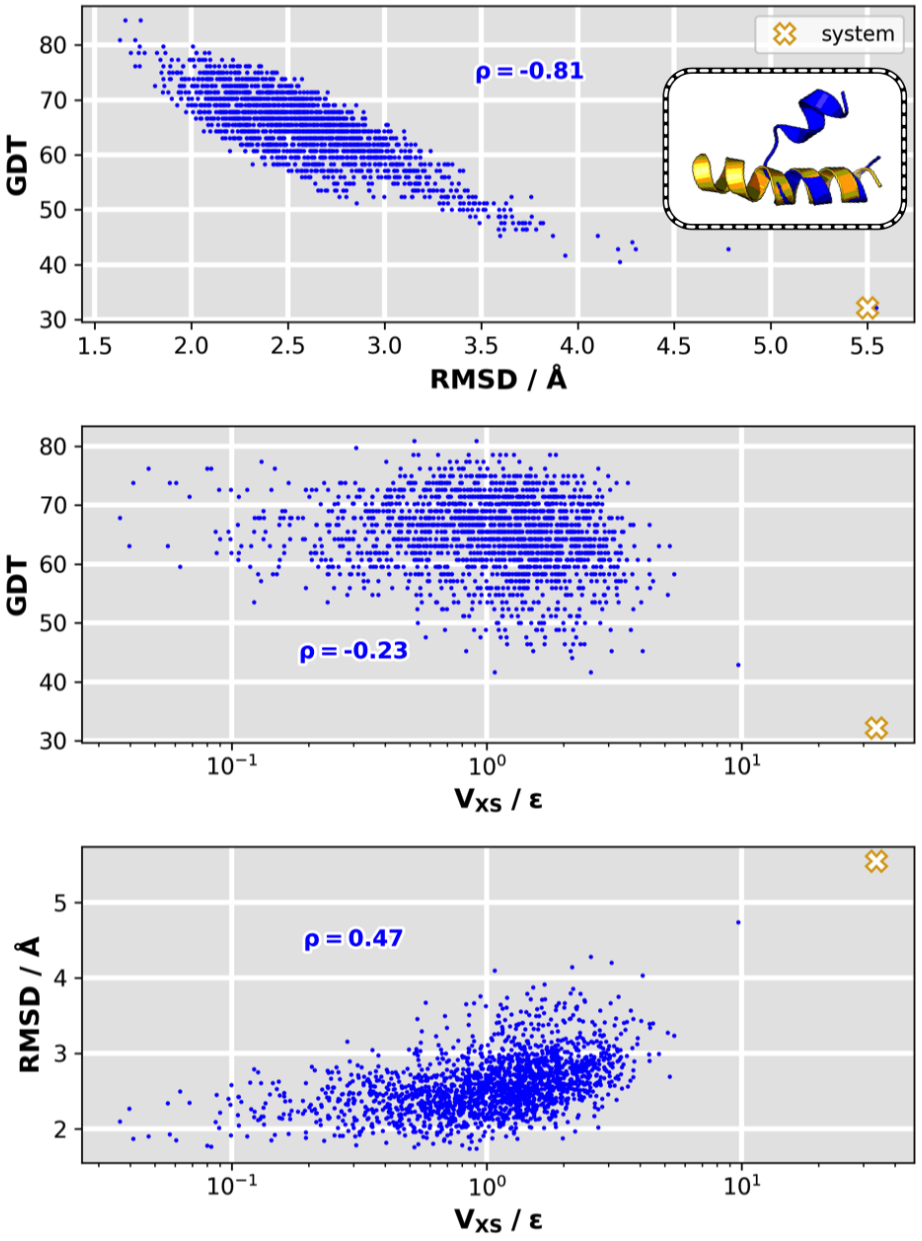


Figure 5.6. VHP_{54}^{74} e → b transition: Mutual correlations among GDT, RMSD, and bias potential. Target GDT versus target RMSD (top), target GDT versus V_{XS} (middle), and target RMSD versus V_{XS} (bottom). ρ is the Pearson correlation of each two plotted quantities.

atoms (see Sec. 4.4)[80]. This ensures a consistent parameterization and yields folding temperatures near 1 in the SBM's reduced units. As a consequence, model-inherent absolute energies are system-specific and not comparable among different proteins. Not only differ biomolecular systems in general and thus their respective absolute energies but also the nature of their conformational transitions each associated with an energy barrier of typically unknown height. Owing to their high diversity, different proteins require different bias weights and temperatures to suitably reshape the structure-based potential and provide sufficient thermal energy to induce or accelerate the conformational transition of interest. The parameters are not transferable and have to be determined separately for each system.

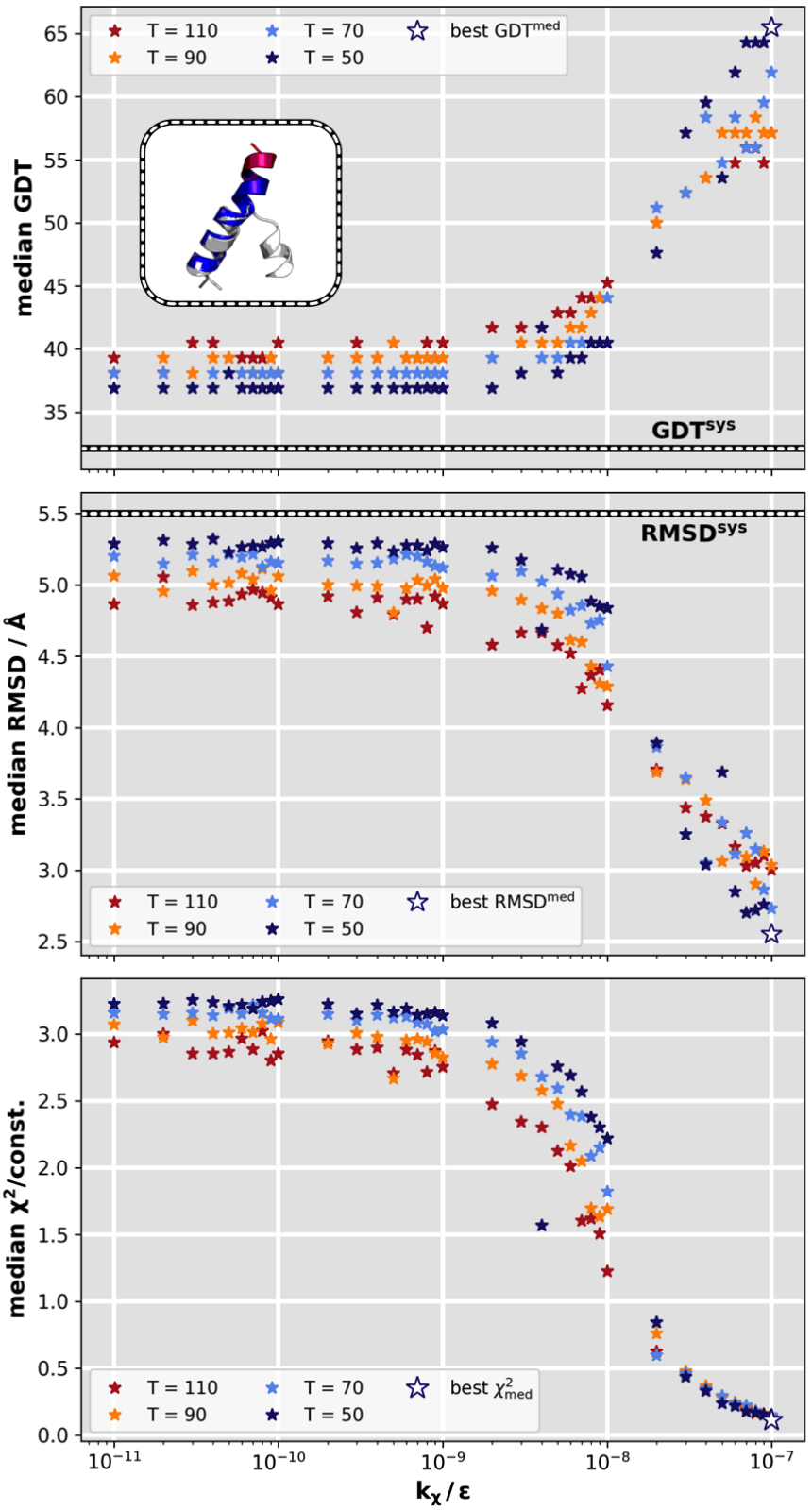


Figure 5.7. Variational study for VHP_{54}^{74} e → b transition. Median target GDT, median target RMSD, and median χ^2 deviation versus bias weight k_χ at different temperatures T. The variational series comprised 148 simulations. Best (maximum) GDT, best (minimum) RMSD, and best (minimum) V_{XS} are marked by a white star, each outlined in the color of the related temperature.

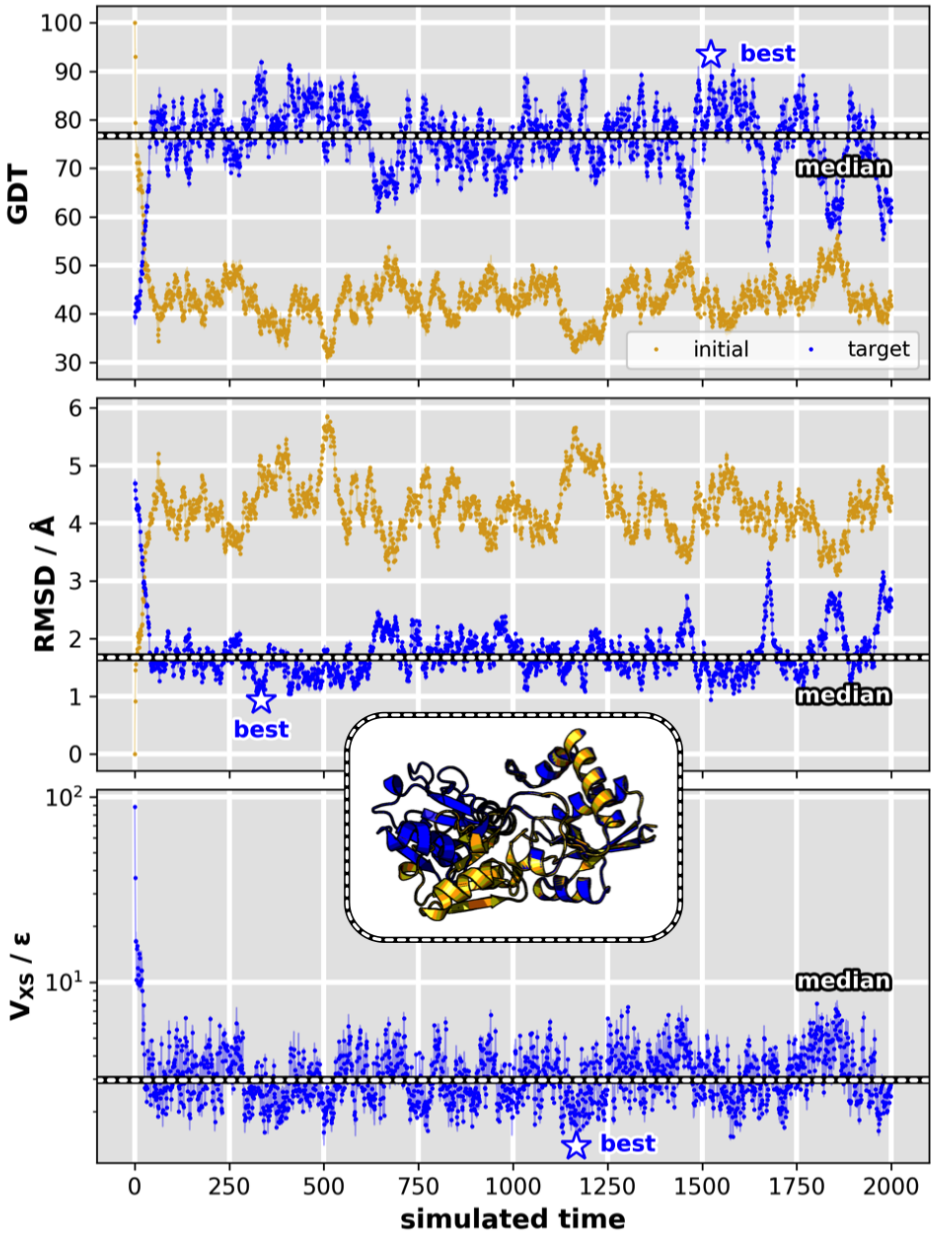


Figure 5.8. `XSBM` simulation results for LAO protein h → a transition. Results are shown for parameters $(T,\ k_\chi) = (50,\ 1 \cdot 10^{-10}\,\varepsilon)$. Initial and target GDT (top), initial and target RMSD (middle), and bias energy (bottom) versus simulated time. Maximum GDT, minimum RMSD, and minimum V_{XS} are marked by a star.

Lysine-, Arginine-, Ornithine-Binding Protein

Upon binding lysine, LAO protein undergoes a structural change from an apo to a holo state (see Fig. 5.2 b)[81]. Modeling this domain motion based on artificial difference data gives a theoretically constructed test example of a real protein movement. I back-calculated reference and target scattering curves from the crystal structures with `CRYSOL`[41], which thus include solvation-shell contributions (see Supplementary Sec. A.2).

Structural analysis. Time-dependent initial and target GDT, RMSD, and bias potential are shown in Fig. 5.8 and Supplementary Fig. D.17 for the h → a and the a → h transition, respectively. I find that

Table 5.3. Results for scattering-guided simulations of LAO protein. k_χ, bias weight, T, temperature, $\mathrm{GDT}^{\mathrm{med}}$, median target GDT, $\mathrm{RMSD}^{\mathrm{med}}$, median target RMSD, $\mathrm{GDT}^{\mathrm{max}}$, maximum target GDT, $\mathrm{RMSD}\,(\mathrm{GDT}^{\mathrm{max}})$, corresponding target RMSD, $\mathrm{RMSD}^{\mathrm{min}}$, minimum target RMSD, $\mathrm{GDT}\,(\mathrm{RMSD}^{\mathrm{min}})$, corresponding target GDT, $\mathrm{GDT}^{\mathrm{med}}_{\mathrm{free}}$, median GDT of an unbiased simulation of each target, $\mathrm{RMSD}^{\mathrm{med}}_{\mathrm{free}}$, median RMSD of an unbiased simulation of each target, $\mathrm{GDT}\,(V^{\mathrm{min}}_{\mathrm{XS}})$, target GDT associated with minimum bias potential, $\mathrm{RMSD}\,(V^{\mathrm{min}}_{\mathrm{XS}})$, target RMSD associated with minimum bias potential, τ^{75}_{sim}, fraction of simulated time associated with a GDT greater than 75, T^{75}_{comp}, corresponding absolute core time.

method	**SBM**		**MD**	
transition	**h → a**	**a → h**	**h → a**	**a → h**
k_χ	$1 \cdot 10^{-10}\,\varepsilon$	$2 \cdot 10^{-10}\,\varepsilon$	$1 \cdot 10^{-9}$ kJ/mol	
T	50	50	300 K	
$\mathrm{GDT}^{\mathrm{med}}$	76.79	82.46	79.30	90.86
$\mathrm{RMSD}^{\mathrm{med}}$ / Å	1.67	1.39	1.59	1.05
$\mathrm{GDT}^{\mathrm{max}}$	93.38	92.96	94.54	97.38
$\mathrm{RMSD}\,(\mathrm{GDT}^{\mathrm{max}})$ / Å	0.94	0.91	0.90	0.83
$\mathrm{RMSD}^{\mathrm{min}}$ / Å	0.93	0.91	0.90	0.74
$\mathrm{GDT}\,(\mathrm{RMSD}^{\mathrm{min}})$	91.91	92.86	94.54	96.64
$\mathrm{GDT}^{\mathrm{med}}_{\mathrm{free}}$	87.08	95.17	74.27	38.13
$\mathrm{RMSD}^{\mathrm{med}}_{\mathrm{free}}$ / Å	1.17	0.82	1.80	5.05
$\mathrm{GDT}\,(V^{\mathrm{min}}_{\mathrm{XS}})$	73.11	87.08	78.57	86.03
$\mathrm{RMSD}\,(V^{\mathrm{min}}_{\mathrm{XS}})$ / Å	1.89	1.31	1.57	1.19
τ^{75}_{sim} (fraction)	0.02	0.01	0.02	0.01
T^{75}_{comp} (absolute)	5 min 24 s	4 min 11 s	8 min 39 s	7 min 6 s

biasing the simulations towards artificial difference data resulted in both transitions to occur instantaneously. The simulations showed one clear transition from initial to target conformation at the beginning, where V_{XS} minimized accordingly. Target GDT and RMSD proceeded near their respective median values from unbiased simulations afterwards. Table 5.3 gives a summary of all performance indicators. The best structures as measured by the maximum target GDT are shown in Fig. 5.5 b and e. While the best median target GDTs (RMSDs) are 76.79 (1.67 Å) and 82.46 (1.39 Å) for the h → a and a → h transition, the maximum target GDTs (minimum C_α RMSDs) are 93.38 (0.93 Å) and 92.96 (0.91 Å), respectively. These numbers are consistent with corresponding unbiased simulations. My scattering-guided simulations thus could reproduce each target structure with the SBM's inherent accuracy. I find my `XSBM` method to be capable of refining structures towards full agreement with the target state, at least for this test system. The compute times of unbiased to scattering-guided simulations scaled as 1 to 2.7 (holo to h → a) and 1 to 4.2 (apo to a → h). Analogous explicit-solvent results are shown in Supplementary Figs. D.20 and D.22. Provided equal computing resources, the `XSBM` refinements required only a half the compute time T^{75}_{comp} by comparison with scattering-guided explicit-solvent MD.

Correlation analysis. As evident from Fig. 5.9 and Supplementary Fig. D.18, the bias energy is suitably correlated with the target GDT and RMSD for both directions of the conformational transition. With the transitions happening almost instantaneously, various structures with small V_{XS} and high target GDT, or equally, low target RMSD, exist. The resulting cluster of target-like structures that reproduce the scattering data disrupts a linear relationship between V_{XS} and the structural similarity measures. This leads to relatively small but certainly negative (positive) Pearson correlations of V_{XS} with the target

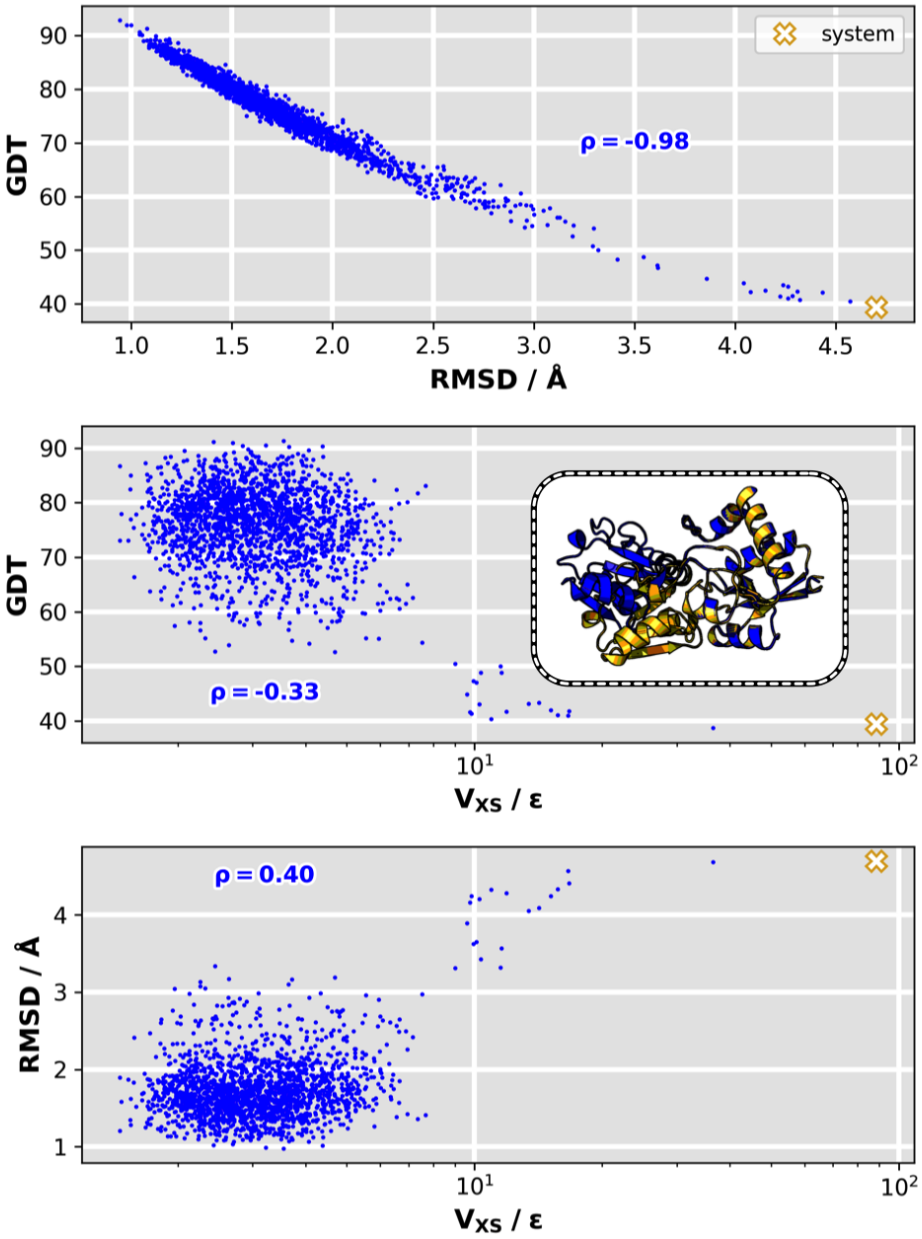


Figure 5.9. LAO protein h → a transition: Mutual correlations among GDT, RMSD, and bias potential. Target GDT versus target RMSD (top), target GDT versus V_{XS} (middle), and target RMSD versus V_{XS} (bottom). ρ is the Pearson correlation of each two plotted quantities.

GDT (RMSD). Analogous explicit-solvent results are shown in Supplementary Figs. D.21 and D.23 for the h → a and the a → h transition, respectively. According to Fig. 5.9 and Supplementary Fig. D.21, the structural diversity at equal bias potential levels is comparable for `XSBM` and explicit-solvent refinements. As expected, the best structures cannot be identified on the basis of V_{XS} alone due to the low information content of the SAXS data. It however can serve as an indicator for a refinement's current state and eventual success or failure.

Variational grid search. As before, I conducted grid-search variational studies on bias weight and temperature. My results for the h → a and the a → h transition are shown in Fig. 5.10 and Supplementary Fig. D.19, respectively. In both cases, the median χ^2 deviation dropped down significantly near a bias weight of $10^{-10}\,\varepsilon$. While the best median target GDT and RMSD were reached at this point,

both quantities started to become worse again for larger k_χ. As indicated by the shaded regions of undefined bias potential, `XSBM` simulations using a k_χ greater than $6 \cdot 10^{-8}\,\varepsilon$ blew up. Depending on χ^2, a disproportionate bias weight may produce a very large bias potential. This generates excessive scattering forces which eventually lead to a failure of the integrator. I observed an overall similar temperature dependency as for the VHP_{54}^{74} test system. Since the hinge-like subdomain motion in LAO protein does not change the molecular shape drastically, low temperatures were sufficient to sample structures near the target state with high accuracy.

Adenylate Kinase

Modeling ADK's large-scale structural transition between an open and closed conformation based on artificial difference data gives another test case of a real protein movement (see Fig. 5.2 c). I computed artificial target difference data from the crystal structures using the Debye equation on amino-acid level corrected for displaced solvent [18,163].

Structural analysis. Initial and target GDT, RMSD, and the bias energy versus simulated time are shown in Fig. 5.11 and Supplementary Fig. D.24 for the o → c and the c → o transition, respectively. Similar to LAO protein, I observed one clear instantaneous transition from the initial to the target state in both cases. Maximum target GDT structures are shown in Fig. 5.5 c and f. While the best median target GDTs (RMSDs) are 69.16 (2.37 Å) and 70.79 (2.03 Å) for the o → c and c → o transition, the maximum target GDTs (minimum target RMSDs) are 83.30 (1.45 Å) and 84.35 (1.30 Å), respectively. All performance indicators are summarized in Table 5.4. The compute times of unbiased and scattering-guided structure-based simulations scaled as 1 to 7.3. Analogous explicit-solvent results for the o → c and the c → o transition can be found in Supplementary Figs. D.27 and D.29, respectively. Considering compute times T_{comp}^{75}, my `XSBM` refinements turned out to be faster by more than two orders of magnitude than the explicit-solvent approach while yielding comparably or even more accurate structures in terms of target GDT and RMSD.

Correlation analysis. I analyzed the mutual correlations of target GDT, target RMSD, and bias energy (see Fig. 5.12 and Supplementary Fig. D.25), where I observed the same tendencies as described for LAO protein. Analogous explicit-solvent results are shown in Supplementary Figs. D.28 and D.30.

Variational grid search. Grid-search variational studies on bias weight and temperature also revealed a similar behavior as for LAO protein (see Fig. 5.13 and Supplementary Fig. D.26 for the o → c and c → o transition, respectively). I find the χ^2 deviation to minimize for k_χ slightly below $10^{-10}\,\varepsilon$. In contrast to LAO protein, the target GDT and RMSD continued to improve steadily with increasing k_χ until the simulations finally failed due to excessively large scattering forces at $k_\chi \geq 10^{-8}\,\varepsilon$. Again, the evolution of median target GDT and RMSD indicates that lower temperatures were suited to stably reach the target conformation.

5.2.4 Discussion

Even though solution SAXS has significantly gained in importance and popularity for structural analyses of biomolecules, reconstructing 3D atomistic models from 1D scattering intensities is still a challenge. In my first main project, I developed the `XSBM` method for interpreting difference scattering data within structure-based simulations to obtain structural models [27]. Reaching median and maximum GDTs up to 80 and 90 with respect to the desired target state, I have shown my method to be capable of probing

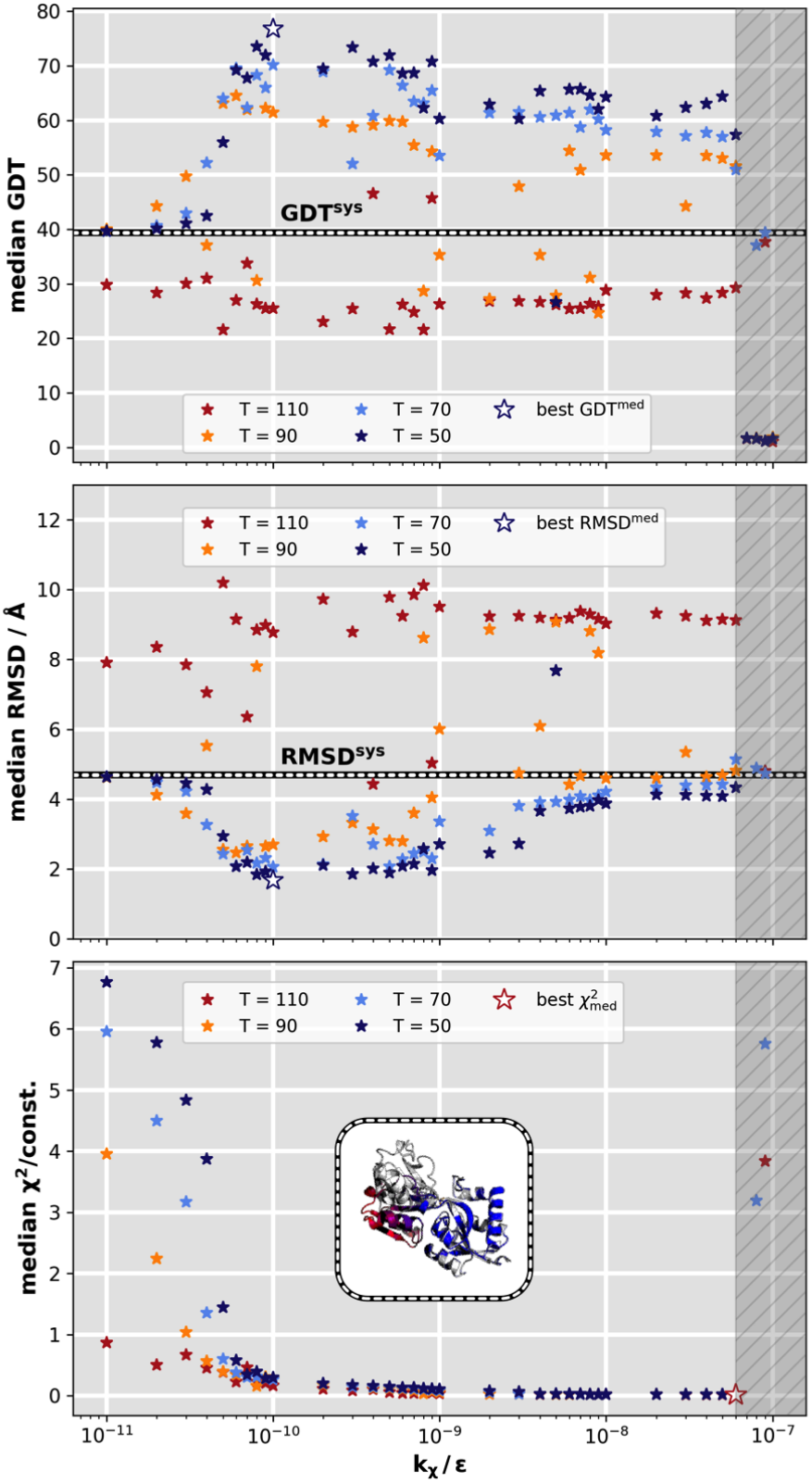


Figure 5.10. Variational grid search for LAO protein h → a transition. Median GDT, median target RMSD, and median χ^2 deviation versus bias weight k_χ at different temperatures T. The variational series comprised 148 simulations. Best (maximum) GDT, best (minimum) RMSD, and best (minimum) V_{XS} are marked by a white star, each outlined in the color of the related temperature.

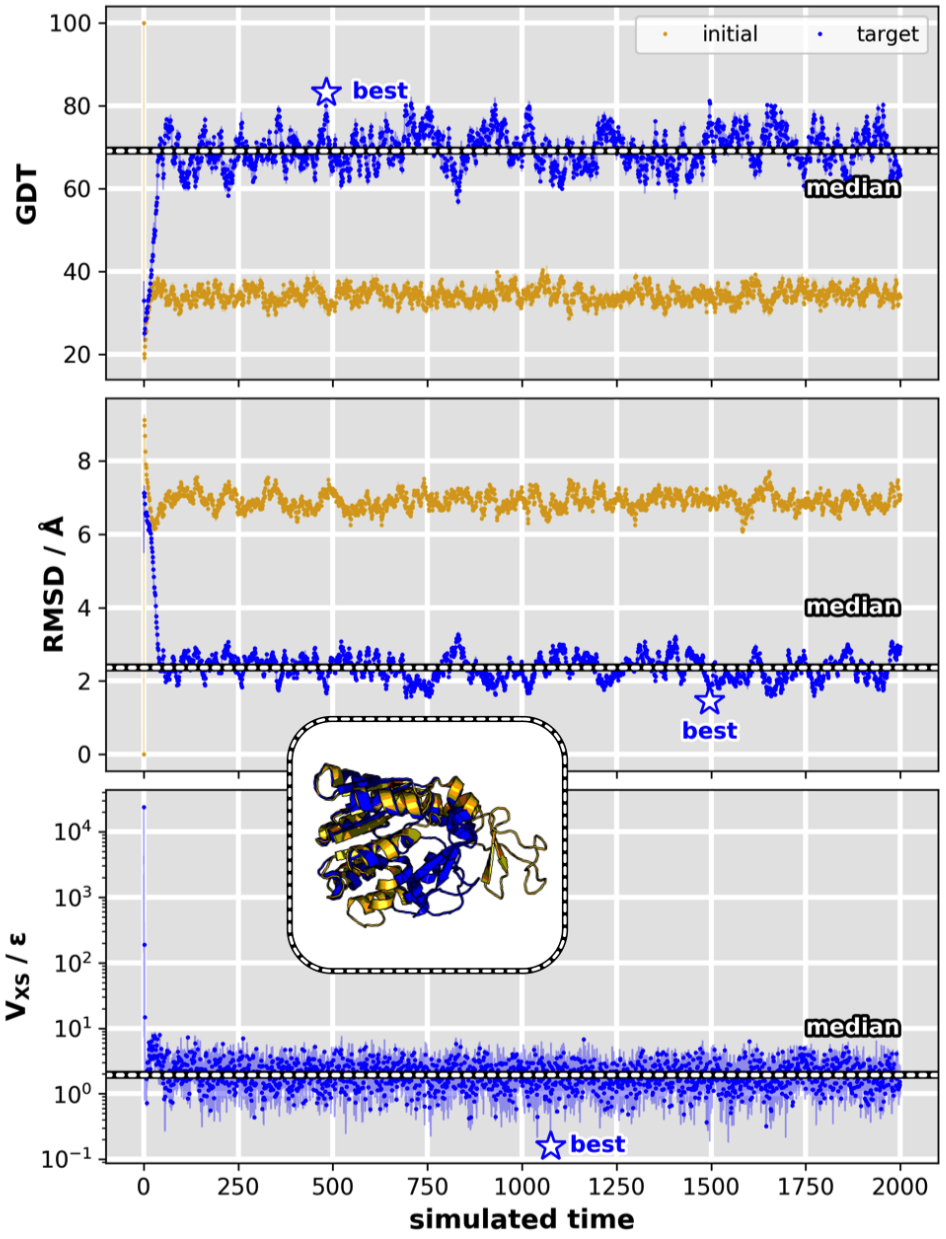


Figure 5.11. `XSBM` **simulation results for ADK o → c transition.** Results are shown for parameters $(T,\, k_\chi) = (50,\, 1 \cdot 10^{-8}\,\varepsilon)$. Initial and target GDT (top), initial and target RMSD (middle), and bias energy (bottom) versus simulated time. Maximum GDT, minimum RMSD, and minimum V_{XS} are marked by a star.

real protein transitions based only on low-resolution scattering data. Incorporating the experimental information on the level of the SBM's force field guides the system from one conformation to another in due consideration of the target data. `XSBM` benefits from the extensive sampling and intrinsically accelerated dynamics of structure-based simulations. The computational efforts reduce up to two orders of magnitude compared to equivalent explicit-solvent simulations while achieving equal-quality results[27]. This suggest that explicit solvation is not a mandatory requirement for successful protein refinement towards scattering data up to a momentum transfer of $0.5\,\text{Å}^{-1}$. The fact that structure-based simulations coupled to difference SAXS data could reproduce each target state with high accuracy indicates that such curves hold sufficient information to guide a simulation towards the correct conformations, at least

Table 5.4. Results for scattering-guided simulations of ADK. k_χ, bias weight, T, temperature, $\mathrm{GDT}^{\mathrm{med}}$, median target GDT, $\mathrm{RMSD}^{\mathrm{med}}$, median target RMSD, $\mathrm{GDT}^{\mathrm{max}}$, maximum target GDT, $\mathrm{RMSD}\left(\mathrm{GDT}^{\mathrm{max}}\right)$, corresponding target RMSD, $\mathrm{RMSD}^{\mathrm{min}}$, minimum target RMSD, $\mathrm{GDT}\left(\mathrm{RMSD}^{\mathrm{min}}\right)$, corresponding target GDT, $\mathrm{GDT}^{\mathrm{med}}_{\mathrm{free}}$, median GDT of an unbiased simulation of each target, $\mathrm{RMSD}^{\mathrm{med}}_{\mathrm{free}}$, median RMSD of an unbiased simulation of each target, $\mathrm{GDT}\left(V^{\mathrm{min}}_{\mathrm{XS}}\right)$, target GDT associated with minimum bias potential, $\mathrm{RMSD}\left(V^{\mathrm{min}}_{\mathrm{XS}}\right)$, target RMSD associated with minimum bias potential, τ^{75}_{sim}, fraction of simulated time associated with a target GDT greater than 75, T^{75}_{comp}, corresponding absolute core time.

method	**SBM**		**MD**	
transition	**o → c**	**c → o**	**o → c**	**c → o**
k_χ	$1 \cdot 10^{-8}\,\varepsilon$	$7 \cdot 10^{-9}\,\varepsilon$	$5 \cdot 10^{-10}$ kJ/mol	$5 \cdot 10^{-11}$ kJ/mol
T	50	50	300 K	300 K
$\mathrm{GDT}^{\mathrm{med}}$	69.16	70.79	68.69	71.61
$\mathrm{RMSD}^{\mathrm{med}}$ / Å	2.37	2.03	2.27	2.21
$\mathrm{GDT}^{\mathrm{max}}$	83.30	84.35	76.75	85.75
$\mathrm{RMSD}\left(\mathrm{GDT}^{\mathrm{max}}\right)$ / Å	1.65	1.30	1.95	1.32
$\mathrm{RMSD}^{\mathrm{min}}$ / Å	1.45	1.30	1.87	1.31
$\mathrm{GDT}\left(\mathrm{RMSD}^{\mathrm{min}}\right)$	81.54	84.35	74.18	85.05
$\mathrm{GDT}^{\mathrm{med}}_{\mathrm{free}}$	93.10	80.84	36.92	68.11
$\mathrm{RMSD}^{\mathrm{med}}_{\mathrm{free}}$ / Å	0.91	1.57	6.35	2.33
$\mathrm{GDT}\left(V^{\mathrm{min}}_{\mathrm{XS}}\right)$	64.84	72.55	70.33	71.73
$\mathrm{RMSD}\left(V^{\mathrm{min}}_{\mathrm{XS}}\right)$ / Å	2.62	1.88	2.14	2.23
τ^{75}_{sim} (fraction)	0.03	0.07	0.45	0.05
T^{75}_{comp} (absolute)	19 min 14 s	46 min 47 s	100 h 2 min 50 s	11 h 6 min 59 s

for the studied systems. As illustrated in Fig. 5.1,the efficiency of my structure-based approach is a direct consequence of the underlying energy landscape. In explicit-solvent models, the lowest-energy excited states are not necessarily related to functionally relevant conformational changes. Only few predefined fluctuations occur, which, in contrast, are inherent to SBMs.

As the simulations conveniently run on commodity hardware, `XSBM` is particularly suited for initial high-throughput analyses of scattering data. Technical advances in experimental X-ray sources and detectors have made the wide-angle regime encoding local molecular structure increasingly accessible. To level up with the experimental resolution, the resulting `XSBM` structures may be given a subsequent polish using a more fine-grained force field at the cost of increased computational demands [12,150].

It is important to note that some protein systems cannot be analyzed straightforwardly in SAXS. The structural transitions in my test systems can be viewed as a collective movement along one effective degree of freedom. This influences the protein's shape and thus the difference curve at $q \lessapprox 0.2\,\text{Å}^{-1}$ crucially, facilitating unambiguous fitting of molecular structures. As another example, I studied the structural change in the cytoplasmic portion of a sensor histidine kinase (PDB code 2C2A [172]). This change can be described as a rotation of one subdomain around a helix bundle. It induces a C_α RMSD shift of 12.5 Å but influences the overall molecular shape only marginally. Despite a significant decrease in the χ^2 deviation of simulated and target data, `XSBM` refinements towards artificial difference data did not converge to the target structure. As already explained in Ref. [18], multiple candidate structures can generate interfering features in the difference profiles at higher q values. This implies that structures exist

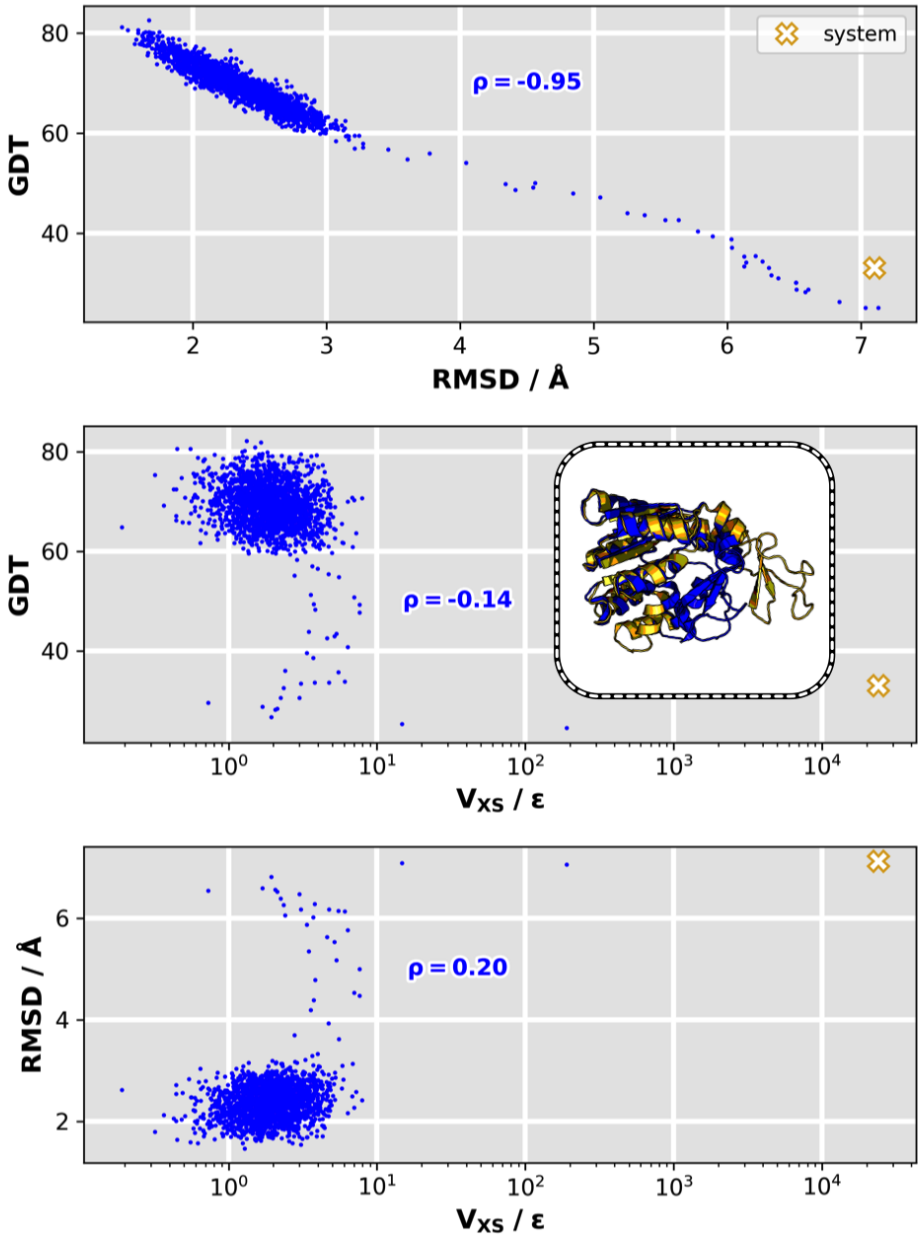


Figure 5.12. ADK o → c transition: Mutual correlations among GDT, RMSD, and bias potential. Target GDT versus target RMSD (top), target GDT versus V_{XS} (middle), target RMSD versus V_{XS}.

which reproduce the difference data adequately but are incompatible with crystallographic structural models.

Built around an interaction potential derived from the known native state, minimalist SBMs can reproduce a protein's geometrical folding properties successfully. Exceptions arise for highly symmetrical systems, where asymmetrical intermediates cannot be inferred from the native topology[65]. These findings confirm that folding indeed is dominated by a protein's native interactions. The very same interactions are also present during the folded protein's functional dynamics, suggesting the structure-based energy landscape to be applicable beyond folding. However, studying function-related conformational dynamics in SBMs may require more flexible formulations, such as the integration of conformations other than the

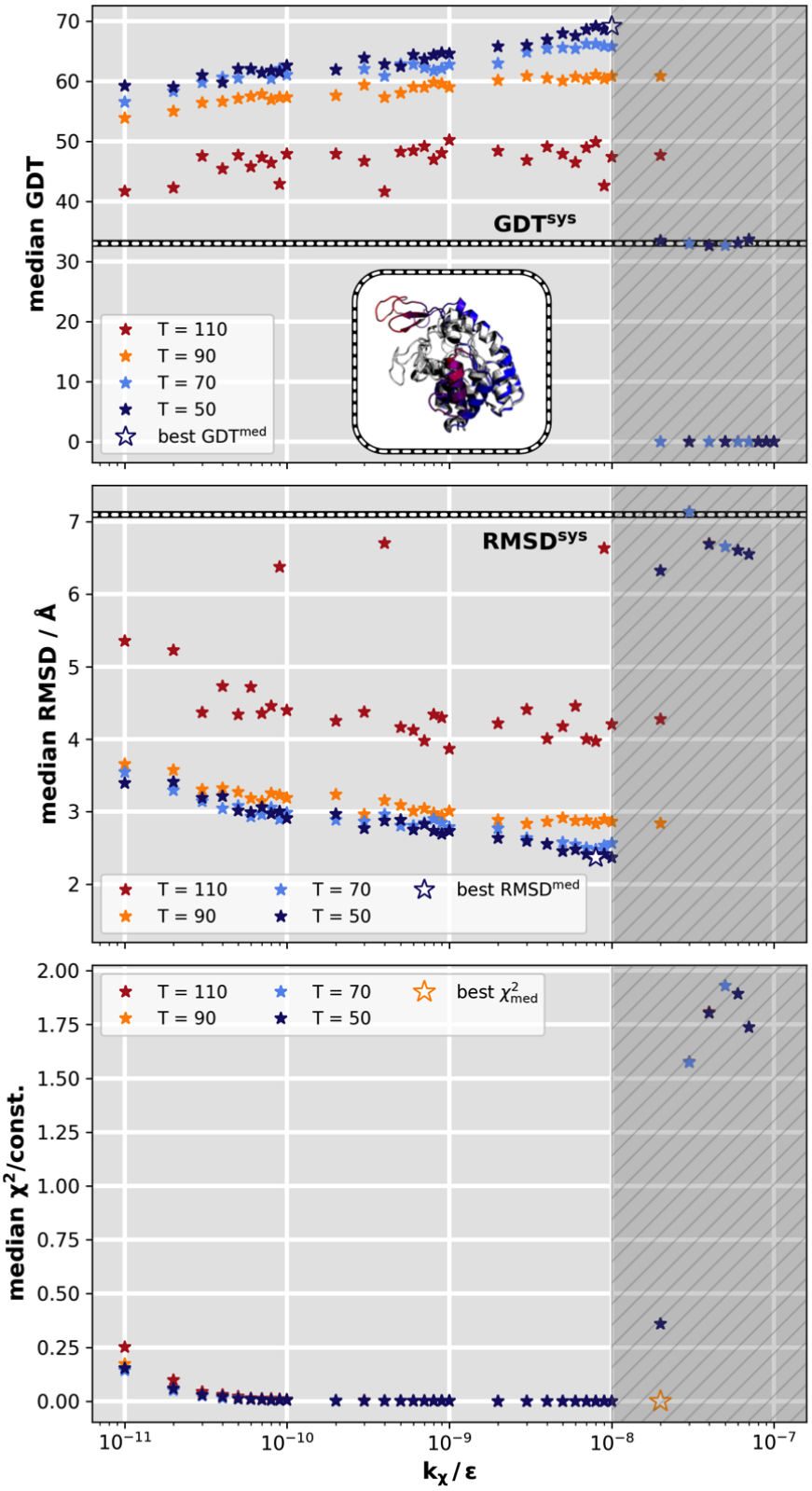


Figure 5.13. Variational study for ADK o → c transition. Median GDT, median target RMSD, and median χ^2 deviation versus bias weight k_χ at different temperatures T. The variational series comprised 148 simulations. Best (maximum) GDT, best (minimum) RMSD, and best (minimum) $V_{\chi S}$ are marked by a white star, each outlined in the color of the related temperature.

native state[23,65]. Even though `XSBM` simulations are complemented by additional structural information from experimental sources, this may limit the application of my data-assisted structure-based approach.

`XSBM` is a solid starting point for further developments towards refining protein structures in data-assisted molecular simulations. These include expanding single-basin SBMs to multi-Gō models with several minima[65], testing other forms of the bias potential[12], and enhancing `XSBM` towards a multireplica framework for accurate ensemble refinement[16]. In addition, I see several possibilities to extend `XSBM` towards including information from other experimental sources. While co-evolutionary contact information from biomolecular sequence data[74] can be integrated via additional contact potentials, distance and angle information from NMR can be included as spatial restraints. Based on a spatial overlap, another energetic term can be introduced to bias a simulation towards a cryo-EM electron density map[13]. In such a multi-data assisted approach, the system is assumed to relax into configurations that are consistent with all these contributions from various sources.

Performing simulations using a combined structure-based/biased/restraint force field raises the important question of how to mutually balance the different contributions, both in relation to each other and in relation to the underlying physical model. Approaches to choosing adequate weights include sophisticated but technically infeasible Bayesian inference[150] (see Sec. 5.1) and easy-to-apply but inefficient grid search[18,27] (see Sec. 5.2). Adopting concepts from computational intelligence, I developed an alternative method for optimizing biomolecular simulation parameters. In the following chapter, I present a fully integrated and automated protocol to resolve the MD parameter selection problem in data-assisted biomolecular simulations.

Beautiful is better than ugly.
Explicit is better than implicit.
Simple is better than complex.
Complex is better than complicated.
Flat is better than nested.
Sparse is better than dense.
Readability counts.
Special cases aren't special enough
to break the rules.
Although practicality beats purity.

Zen of Python

6

PROJECT: Optimizing Biomolecular Simulation Parameters with Computational Intelligence

This chapter focuses on metaheuristic optimization of MD parameters for data-assisted protein simulations, which are a powerful tool to interpret ambiguous experimental data for structural models. The performance of such simulations crucially depends on the non-trivial choice of MD parameters, where the key challenge is balancing experimental information and the physical model. In my second main project, I explore how computational intelligence can be harnessed to overcome this parameter selection challenge. Computational intelligence comprises a class of nature-inspired methodologies to address complex real-world problems which cannot be handled in traditional mathematical modeling approaches. I introduce `FLAPS`*, a self-adapting variant of dynamic particle swarm optimization.* `FLAPS` *is suited for optimizing composite objective functions that depend on both the optimization parameters of interest and additional, a priori unknown weighting parameters which significantly influence the search-space topology. These weighting parameters are learned at runtime, yielding a dynamically evolving and iteratively refined search-space topology. As a practical example, I show how* `FLAPS` *can be applied to find functional parameters for my* `XSBM` *simulations (see Sec. 5.2). The following chapter is reproduced from my Nature Machine Intelligence article "Dynamic particle swarm optimization of biomolecular simulation parameters with flexible objective functions" (2021)*[173].

Since nanoscale proteins can only be observed indirectly, protein structure determination stands and falls with combining experimental data and computational methods. A powerful approach is to complement the data with a physical description of molecular motion within computational simulations[12–14,16,17,27,149] by adding a biasing restraint on the data to the force field (see Chap. 5). An

inherent issue with such data-assisted simulations is balancing experimental and theoretical information. This translates into determining the bias potential's weight, an empirical MD parameter expressing the degree of confidence in the data versus the force field. Selecting the bias weight adequately is nontrivial and crucial for simulation performance. In Bayesian methods, the right weighting arises out of a statistical treatment, along with probabilistically founded values of other parameters[150,151] (see Sec. 5.1). However, such sophisticated approaches are practically inapplicable for users with a primarily experimental background. It still is common practice to determine an "optimal" bias weight by manually searching a fixed grid in the parameter space[17,18,27]. I propose a fundamentally different approach to resolve this parameter selection problem. Adopting concepts from computational intelligence, I introduce `FLAPS` (**Fl**exible self-**A**dapting **P**article **S**warm optimization), a self-learning metaheuristic based on particle swarms. My contributions include:

1. A new type of flexible objective function (OF) to assess a data-assisted simulation's plausibility in terms of simulated structures and thus suitability of the MD parameters used.
2. A self-adapting particle swarm optimizer for dynamically evolving environments resulting from multiple quality features of different scales in the flexible OF.
3. Fully integrated and automated parameter selection for data-assisted biomolecular simulations.

As a proof of concept, I apply `FLAPS` to the selection of relevant MD parameters in my `XSBM` simulations (see Sec. 5.2)[27].

6.1 Particle Swarm Optimization

This section introduces particle swarm optimization as the bedrock of `FLAPS`. Founded on the fact that sharing information among individuals is evolutionarily advantageous, PSO was originally intended for simulating social behavior via a stylized representation of collectively moving organisms, such as bird flocks and fish schools. A more abstract objective was developing a paradigm for modeling the social behavior of humans and their ability to process knowledge as a society in the sense of that we tend to adjust our beliefs and attitudes towards conformity with our fellow humans. Interestingly, the algorithm was found to be performing optimization as it was simplified. PSO is a very accessible and efficient optimizer. It can be implemented in few lines of code, rendering it particularly suited for my MD parameter selection problem.

Particle swarm optimization (PSO) is a nature-inspired technique for optimizing continuous nonlinear functions and solving computationally hard problems[174]. Based on swarm intelligence, it exploits the collective behavior emerging in decentralized, self-organized systems of cooperating individuals. Even though there is no supervising control dictating the individuals' local behavior, their interactions unconsciously lead to an intelligent global behavior of the swarm as a whole (see Fig. 6.1). Each individual is termed a "particle" which, in fact, represents a potential solution to the considered problem.

As outlined in the pseudocode in Algorithm 1, the optimizer aims at iteratively improving a candidate solution with respect to a quality-gauging OF. The problem is approached by considering a swarm of particles, each corresponding to a specific position in the searched parameter space. Across multiple search rounds (generations), the particles move about in the search space according to their positions and velocities. Due to a discretized motion, they do not follow a smooth trajectory but jump around in a discontinuous particle flight, where each particle adjusts its flight according to both its own and its peers' flying experience. Particles remember their personal best position visited so far, p^p_{best}, and

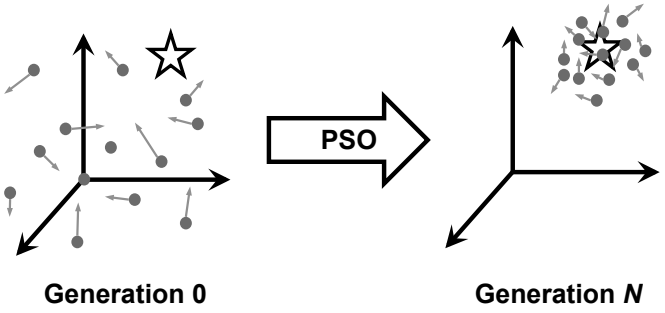


Figure 6.1. PSO operating principle. PSO is a bio-inspired optimization technique based on the emergent behavior of swarms. The exploration of a problem space is modeled by a cooperating particle swarm, where the individuals' successes influence their searches and those of their peers. The particles are evolved by mutual cooperation and competition from one generation to another. Although there is no supervising control dictating the individuals' local behavior, their interactions lead to the swarm behaving intelligently as a whole.

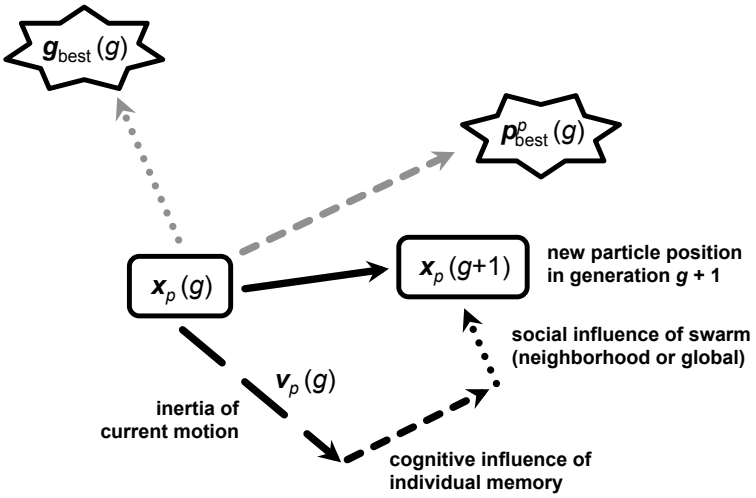


Figure 6.2. PSO particle update. The velocity that is added to a particle's current position to propagate it to the next generation is the vectorial sum of three contributions, i.e., the inertia of its current direction of motion, the cognitive influence of its individual memory, and the social influence of the swarm.

the current global best position in their social communication network, g_{best}. These locations act as attractors in the search space and are updated as they improve at each particle refresh. To propagate a particle to the next generation, a velocity is added to its current position (see Fig. 6.2). This velocity has a cognitive component towards p^p_{best} and a social component towards g_{best}, each weighted by a random number in the range $[0, \phi_i]$, where the acceleration coefficients ϕ_i balance exploration versus exploitation. To eventually yield convergence, the particle flight must progressively contract, which is ensured by mechanisms such as velocity clamping[175], inertia weight[176], or constriction[177]. Altogether, this is assumed to move the swarm towards the best parameter combinations.

PSO uses only basic arithmetic operations and is computationally cheap in terms of both memory and speed[174]. The algorithm turned out to work well for various static problems[178]. Real-world problems, however, are often dynamic. The global optimum can shift with time and the optimizer has to track this change. The difficulty consists in the twofold problem of outdated memory due to environment dynamism and diversity loss due to premature convergence[179]. Application to such problems revealed

Algorithm 1: PSO pseudocode.

Initialize population *pop* with swarm size S particles at random positions $\boldsymbol{x}_p$, $p = 1, ..., S$, between upper and lower bounds of the search space $\boldsymbol{b}_{\mathrm{up}}$ and $\boldsymbol{b}_{\mathrm{lo}}$, respectively.
for $g \leftarrow 1$ **to** *maximum generations* G **do**
  for *particle* **in** *pop* **do**
    Evaluate objective function f at $particle.position = \boldsymbol{x}_p$: $particle.fitness = f(\boldsymbol{x}_p)$
    Update personal best p^p_{best}:
    if $particle.fitness$ **is better than** $f(p^p_{\mathrm{best}})$ **then**
      $p^p_{\mathrm{best}} = particle.position$
    end
  end
  Update global best g_{best} accordingly: $g_{\mathrm{best}} = \mathrm{best}(p^p_{\mathrm{best}})$
  for *particle* **in** *pop* **do**
    Update velocity and position:
    $particle.speed \mathrel{+}=$
    $\mathrm{rand}(0, \phi_1)(p^p_{\mathrm{best}} - particle.position) + \mathrm{rand}(0, \phi_2)(g_{\mathrm{best}} - particle.position)$
    Regulate velocity via $\boldsymbol{s}_{\mathrm{max}}$:
    if $particle.speed > \boldsymbol{s}_{\mathrm{max}}$ **then**
      $particle.speed = \boldsymbol{s}_{\mathrm{max}}$
    end
    if $particle.speed < -\boldsymbol{s}_{\mathrm{max}}$ **then**
      $particle.speed = -\boldsymbol{s}_{\mathrm{max}}$
    end
    $particle.position \mathrel{+}= particle.speed$
  end
end
Result: g_{best}

that the algorithm needs to be enhanced by concepts such as repulsion, dynamic network topologies, or multi-swarms [179,180].

Maintaining a population of diverse solutions enables enormous exploration and paves the way to large-scale parallelization. Swarm-based algorithms can be easily scaled to exploit the full potential of modern supercomputers[181]. PSO has several hyperparameters affecting its behavior and efficiency, and selecting them has been researched extensively[182,183]. Strategies include using higher-level meta-optimizers[184,185] or refining them at runtime[186]. Ref.[187] provides an overview of practical PSO applications. Ref.[179] and Ref.[188] give comprehensive reviews of PSO with particular focus on dynamic environments and hybridization perspectives, respectively.

Other swarm-based optimizers[189] include genetic algorithms[190], differential evolution[191], ant colony optimization[192,193], and the artificial bee colony algorithm[194,195]. Among these, PSO is one of the simplest and most intuitively understandable approaches. Its two major ingredients, that is, particle dynamics and the information network, make it a practical and widely applicable optimizer[179] that is known to routinely deliver useful results, irrespective of whether it is guaranteed to give the absolutely best performance on a problem or not. It can be easily adapted to various domains and conveniently hybridized with other techniques. Compared with, e.g., the genetic algorithm, PSO is a more accessible concept, has fewer parameters, and is easier to implement, which renders it particularly well-suited for my MD parameter optimization problem.

6.2 Multi-Response Problems

Real-life optimization problems often involve multiple incomparable or even conflicting quality features (responses). To obtain compatible solutions, these responses must be taken into account simultaneously, which is often accomplished by combining them within one composite OF. A trade-off solution for which no contribution can be improved without worsening another is called Pareto-optimal. PSO has been applied to multi-response optimization in many fields, e.g., industrial manufacturing and processing[196,197] or electric-motor design[198,199]. Commonly, the set of responses is boiled down via multiplication by manually chosen weights, henceforth referred to as OF parameters, followed by summation and uni-objective optimization. Choosing OF parameters is nontrivial yet strongly impacts global optimization performance by skewing the OF. I present `FLAPS`, a new highly scalable and adaptive type of canonical PSO. `FLAPS` builds on a flexible OF which automatically and interdependently balances different responses. OF parameters are learnt on the fly through iterative refinement, yielding a dynamically evolving OF landscape. In this way, `FLAPS` can cope with various responses of different scales.

6.3 A Flexible Self-Adapting Objective Function

Typically, the set of responses is mapped to a scalar score by calculating the scalar product with a priori fixed weights. These OF parameters supposedly reflect relative importance and thus implicitly encode arbitrary prior beliefs. I set up a "maximum-entropy" OF with the fewest possible assumptions instead:

$$f\left(\boldsymbol{x};\, \boldsymbol{z} = \left(\{\mu,\, \sigma\}_j\right)\right) = \sum_j \frac{R_j\left(\boldsymbol{x}\right) - \mu_j}{\sigma_j} \stackrel{\text{def}}{=} \sum_j \left[R_j\left(\boldsymbol{x}\right)\right]_{\text{std}} \tag{6.1}$$

$\boldsymbol{z}$ is the set of OF parameters, μ_j the mean, and σ_j the standard deviation of response R_j for a particle at position $\boldsymbol{x}$. All responses are considered equally important but can have different orders of magnitude and units. To make them comparable on a shared scale, I standardize each response's set of values gathered over previous generations. This strategy effectively imitates the concept of rolling batch normalization[200]. Each layer's inputs are recentered and rescaled with the aim to improve an artificial neural network's speed, performance, and stability. Batch normalization is believed to introduce a regularizing and smoothing effect and promote robustness with respect to different initialization schemes. The OF in Eq. 6.1 depends not only on the parameters of interest, $\boldsymbol{x}$, in my case the MD parameters, but also on a priori unknown, context-providing OF parameters, $\boldsymbol{z} = \left(\{\mu,\, \sigma\}_j\right)$, from the standardization. Their values cannot be deduced from individual OF evaluations, yet fundamentally control OF performance and hence the optimization process.

My self-adapting PSO variant `FLAPS` solves this problem. Its pseudocode is shown in Algorithm 2. Provided a comprehensive history of all previous particles and their responses, OF parameters are learned on the fly. They are continuously refined according to the current state of the optimization, yielding a dynamically evolving and increasingly distinct OF topology. This environmental dynamism may cause convergence problems if the OF fails to approach a stable topology. As more particles are evaluated, the individual responses' ranges and distributions become better understood. Consequently, OF parameters become more accurate, improving OF performance in assessing suitability of the actual parameters of interest, $\boldsymbol{x}$. After each generation, the values $\boldsymbol{z} = \left(\{\mu, \sigma\}_j\right)$ are used to reevaluate the OF for all particles in the history. Personal best positions, p^p_{best}, and the swarm's global best position, g_{best}, are updated accordingly for propagating particles to the next generation. `FLAPS` uses a traditional PSO velocity formulation[174]. Several strategies to prevent diverging velocities include introducing an inertia

Algorithm 2: `FLAPS` **pseudocode.**

Initialize population *pop* with swarm size S particles at random positions $\boldsymbol{x}_p$, $p = 1, ..., S$, between upper and lower bounds of the search space $\boldsymbol{b}_{\text{up}}$ and $\boldsymbol{b}_{\text{lo}}$, respectively.
for $g \leftarrow 1$ **to** *maximum generations* G **do**
  for *particle* **in** *pop* **do**
    Evaluate responses at $particle.position = \boldsymbol{x}_p$: $particle.fargs = \left[\text{response}_j\left(\boldsymbol{x}_p\right)\right]_j$
  end
  Append current generation *pop* to history *histp*: *histp.append(pop)*
  Update OF parameters $\boldsymbol{z}_g$ based on current knowledge state of responses in *histp*: $\boldsymbol{z}_g =$ *updateParams(histp)*
  for *particle* **in** *histp* **do**
    (Re-)evaluate objective function using most recent $\boldsymbol{z}_g$:
    $particle.fitness = f\left(\boldsymbol{x}_p; \boldsymbol{z}_g\right)$
  end
  for *generation* **in** *histp* **do**
    for *particle* **in** *generation* **do**
      Determine personal best p^p_{best} and update global best g_{best} accordingly.
    end
  end
  for *particle* **in** *pop* **do**
    Update velocity and position:
    $particle.speed \mathrel{+}=$
    $\text{rand}\left(0, \phi_1\right)\left(p^p_{\text{best}} - particle.position\right) + \text{rand}\left(0, \phi_2\right)\left(g_{\text{best}} - particle.position\right)$
    Regulate velocity via $\boldsymbol{s}_{\max} = 0.7\, G^{-1}\left(\boldsymbol{b}_{\text{up}} - \boldsymbol{b}_{\text{lo}}\right)$:
    if $particle.speed > \boldsymbol{s}_{\max}$ **then**
      $particle.speed = \boldsymbol{s}_{\max}$
    end
    if $particle.speed < -\boldsymbol{s}_{\max}$ **then**
      $particle.speed = -\boldsymbol{s}_{\max}$
    end
    $particle.position \mathrel{+}= particle.speed$
  end
end
Result: g_{best}

weight[176] or a constriction factor[177]. I regulate the velocities at each particle update via a maximum value that also determines the granularity of the search[175]. Inspired by the "simplifying PSO" paradigm, `FLAPS` builds on a slim standard PSO core and can easily be complemented by concepts such as inertia weight[176] and swarm constriction[177] or diversity increasing mechanisms[179]. Its time complexity in big $\mathcal{O}$ notation is similar to that of a standard PSO with $\mathcal{O}\left(\frac{S}{P} \cdot G \cdot Sim + S \cdot Opt\right) = \mathcal{O}\left(\frac{S}{P} \cdot G \cdot Sim + S^2 \cdot G\right)$ where P is the number of simulation processors, Sim the maximum simulation time, and all other variables as defined in Algorithm 2.

6.4 Application to Data-Assisted Protein Simulations

I used `FLAPS` to optimize the MD parameters of my `XSBM` simulations for interpreting SAXS data within computationally efficient SBMs[27] (see Sec. 5.2). To assess the utility of different MD parameter sets, I need a metric for simulation quality in terms of physically reasonable structures matching the data. Designing such an OF ad hoc is nontrivial and includes two major aspects: (i) physical plausibility of a simulated ensemble, i.e., how reasonable the simulated structures are from a physics point of view, and (ii) agreement with the target data, i.e., how well the data are reproduced by the simulated structures.

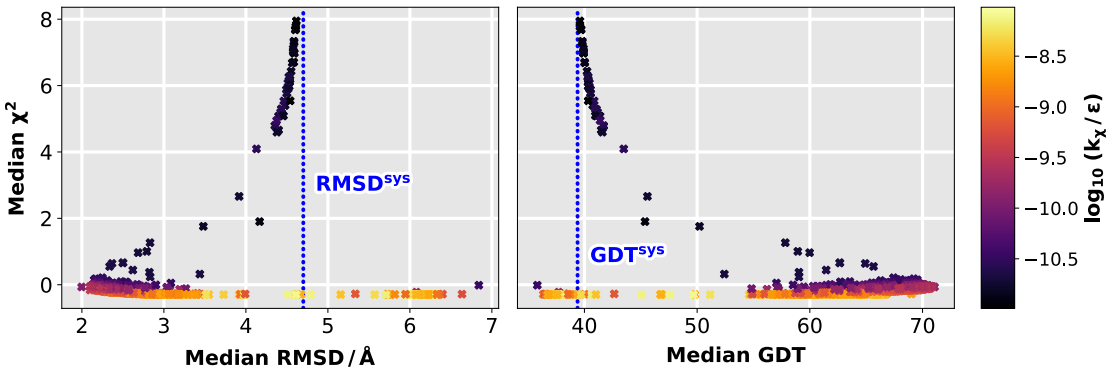


Figure 6.3. Ambiguous χ^2. The ill-posedness of the SAXS inverse problem results in a pronounced ambiguity in the χ^2 deviation of simulated data from the target data. This manifests in a non-injective "two-branch" behavior of the second response as a function of similarity of simulated structures to the desired target state, here quantified by the RMSD and the GDT. RMSD^{sys}, RMSD between initial and target structure, GDT^{sys}, GDT between initial and target structure, k_χ, bias weight. Data from all LAO protein h→a runs combined. Derived from Ref.[173] (extended from original by additional data), used under CC BY 4.0.

To represent these aspects, I use the Rosetta energy function 2015, REF15[201,202], and the least-squares deviation of simulated data from the target data, χ^2[27] (see Eq. 5.17).

Protein structure determination relies on quick and reliable scoring of many models to select the ones closest to the native state. Structures are rated by energetic scores associated with their conformational state. REF15 is a weighted sum of energy terms efficiently approximating the energy of a biomolecular conformation as a function of geometric degrees of freedom and chemical identities[201]. Most terms are physics-based, representing forces like electrostatics and van der Waals interactions. As nature widely conserves protein folds, statistical terms, such as the probability of finding certain torsion angles in the protein backbone, favor structures highly similar to known ones. With a protein's native fold corresponding to the state with minimal free energy, a lower scoring structure is expected to be more native-like. Since the scores cannot be converted to physical energy units directly, REF15 and structural stability are not correlated across different proteins. Similarly, χ^2 values without context are inconclusive and must be compared for each protein system at a time. Both REF15 and χ^2 are available from a simulation, which yields a molecular system's atomic positions over time.

Two MD parameters are particularly important for my `XSBM` simulations, i.e., the bias weight k_χ and the temperature T. As explained in Sec. 5.2, k_χ balances the information in the SAXS data with the physical model. T is a measure of the available thermal energy and controls the system's conformational flexibility. Thus, a particle corresponds to a simulation using a particular set $\boldsymbol{x} = (k_\chi, T)$ of MD parameters to be optimized. I set up the OF as

$$f\left(\boldsymbol{x} = (k_\chi, T)\,;\, \boldsymbol{z}\right) = \left[\text{REF15}_{\text{av}}\right]_{\text{std}} + \left[\chi^2_{\text{med}}\right]_{\text{std}} + \left[\chi^{-2}_{\text{av}}\right]_{\text{std}} . \tag{6.2}$$

While the first response evaluates the simulated structures' average physical stability, the second is the median χ^2 deviation of the simulated data from the target data. Due to the ill-posed nature of the SAXS inverse problem (see Sec. 3.1), globally distinct protein structures can possess the same scattering intensity. As shown in Fig. 6.3, this may lead to a pronounced ambiguity in χ^2. To resolve the resulting non-injectivity in the OF, I introduce a third response, the inverse average χ^2 deviation. This term acts as a regularizer, rewarding deviations from the target data and thus preventing possible overfitting. While each response by itself cannot reflect a simulated ensemble's similarity to the target structure accurately, combining them yields a robust surrogate model of the latter. This means the smaller

the OF, the more physical, data-consistent, and thus (likely) similar to the target state the simulated structures are. It is important to note that different combinations of bias weight and temperature can equally yield useful results. There is no MD parameter ground truth for this type of data-assisted SBM simulations so that I applied a purely evidence-based evaluation according to the structural similarity to the target state. I used the global distance test[82–84] to quantify differences between two conformations of a protein and to validate the OF as a surrogate model of an ensemble's similarity to the target structure. The GDT is a more suitable similarity measure than the widely used RMSD, which is very sensitive to small numbers of locally displaced atoms in an otherwise accurate structure (see Sec. 4.5). Higher GDTs mean a stronger similarity between two models, and I considered structures with a target GDT greater than 50 topologically accurate[85].

6.5 Results

I optimized k_χ and T for my `XSBM` simulations of the two well-characterized protein systems considered before, that is LAO protein and ADK (see Sec. 5.2)[27].

Setup

My simulation setups[203] are available on `GitHub`[1] (also see Supplementary Sec. D.2). For both proteins, I used `CRYSOL`[41] in its default mode to back-calculate artificial SAXS intensities from each two states (see Supplementary Sec. A.2). Each intensity contained 700 equidistant data points up to a momentum transfer of $q = 0.35\,\text{Å}^{-1}$. As described before, residue-based Debye intensities of simulated structures were calculated at runtime in `GROMACS`[27]. I rescaled the `CRYSOL` intensities such that the extrapolation of the forward scattering at $q = 0\,\text{Å}^{-1}$ matched the `GROMACS` value. I calculated the target SAXS data as the difference of the initial and target structures' rescaled `CRYSOL` intensities, including realistic uncertainties[204]. The rescaled initial `CRYSOL` intensity, including uncertainties, served as the absolute reference scattering[27]. I included 17 data points as the difference curve's local extrema and interjacent points centered between each two extrema. For both proteins, this corresponds well to the number of independent Shannon channels and thus to the number of independent data points in the SAXS curve[12].

I performed seven `FLAPS` runs with different initial conditions for each protein transition. It is important to note that swarm-based metaheuristics such as `FLAPS` have hyperparameters influencing their optimization behavior. Their efficacy can only be demonstrated empirically by a finite number of computational experiments. For the application of `FLAPS` to `XSBM`, I used a swarm of ten particles and 15 generations as a workable trade-off between optimization performance and compute time, which I found to be sufficient for convergence in preceding trial runs. I performed the calculations on 1 000 cores of a supercomputer, where one run cost approximately 40 000 core hours.

General Analysis

For each simulation in a `FLAPS` run, I calculated the median GDT with respect to the target state from all structures in the trajectory. I compared the g_{best} median GDT with the actual best median GDT in each run. For the sake of completeness, I also considered the more common RMSD as a structural similarity measure. To validate the OF as a surrogate model of a simulated ensemble's similarity to the target

[1] https://github.com/FLAPS-NMI/FLAPS-sim_setups/releases/tag/v1.0

System	LAO protein (h → a)			Adenylate kinase (o → c)		
Seed	**1790954**	**1791104**	**1791106**	**1795691**	**1798723**	**1810891**
ρ	-0.94	-0.87	-0.87	-0.85	-0.81	-0.74
min (f)	-2.34	-1.79	-1.99	-1.42	-1.57	-1.62
max (f)	8.32	5.86	4.41	6.92	9.05	7.47
Best simulation in terms of OF						
$k_\chi \,/\, \varepsilon$	2.170 e-10	3.339 e-11	5.081 e-11	1.969 e-09	1.970 e-09	1.819 e-09
T	13.19	28.82	11.06	16.90	10.56	10.09
$\mathrm{GDT}^{\mathrm{med}}$	70.59	69.22	69.44	63.20	63.78	63.55
Best simulation in terms of $\mathrm{GDT}^{\mathrm{med}}$						
$\mathrm{GDT}^{\mathrm{med}}$	70.69	69.54	70.69	63.78	63.78	63.90
$f\left(\mathrm{GDT}^{\mathrm{med}}\right)$	-2.03	-1.73	-1.55	-1.28	-1.57	-1.59
$k_\chi \,/\, \varepsilon$	3.001 e-10	3.422 e-11	4.190 e-10	2.030 e-09	1.970 e-09	2.170 e-09
T	11.98	29.63	10.03	10.84	10.56	10.23

Table 6.1. `FLAPS` optimization results[173]. For each protein transition, the best three runs are listed. OF, objective function f, ρ, Pearson correlation of OF and median GDT, k_χ, bias weight, T, temperature, $\mathrm{GDT}^{\mathrm{med}}$, median GDT.

structure, I considered its Pearson correlation with the median GDT, ρ, as a measure of linear correlation. Since minimizing the OF should be equivalent to maximizing the GDT, I expect negative correlations, ideally -1. As the Pearson correlation only reflects linearity, I also studied the exact relations of the OF and its individual responses with the median GDT. Results of each three best runs for LAO protein's h → a transition and ADK's o → c transition are listed in Table 6.1. Complete results for all systems can be found in Supplementary Tables E.1 to E.4. A detailed discussion of the swarm's convergence behavior is presented in Supplementary Sec. E.2. I did comparative grid-search optimizations to evaluate `FLAPS`'s efficiency where I found superior performance of `FLAPS` for all considered protein systems. Related results are given in Supplementary Sec. E.3.

Results

As shown exemplarily for LAO protein's h → a transition in Fig. 6.4, the OF consistently converged to a stable topology. The OF and its separate responses are shown in Fig. 6.5 as functions of the median GDT. ρ values up to -0.94 and -0.85 confirmed the OF's suitability for both LAO protein's h → a transition and ADK's o → c transition, respectively. To identify the best MD parameter combinations, the OF must have low values for large GDTs, irrespective of how complex the actual relationship may be. I find this to be the case for both systems.

For LAO protein's h → a transition (see Fig. 6.5 left), no response by itself is suited to reliably identify physically reasonable structures matching the target data. The two minima of $\mathrm{REF15}_{\mathrm{av}}$ at $\mathrm{GDT}^{\mathrm{sys}}$ and the maximum GDT correspond to the stable holo (initial) and apo (target) state, respectively. χ^2_{med} with its distinct two-branch behavior is a perfect example of how the ambiguity in the inverse problem of SAXS hinders accurate structural modeling based on the data alone. Trivially, this also reflects in χ^{-2}_{av}. It is only the combination of all three standardized responses that resolves this ambiguity and gives an accurate surrogate of simulation quality in terms of physical structures matching the data.

For ADK's o → c transition (see Fig. 6.5 right), the $\mathrm{REF15}_{\mathrm{av}}$ minimum near $\mathrm{GDT}^{\mathrm{sys}}$ is significantly less pronounced than for LAO protein. This is due to the fact that the immediate structural change from initial open to target closed state already occurs at smaller bias weights compared to LAO protein (see

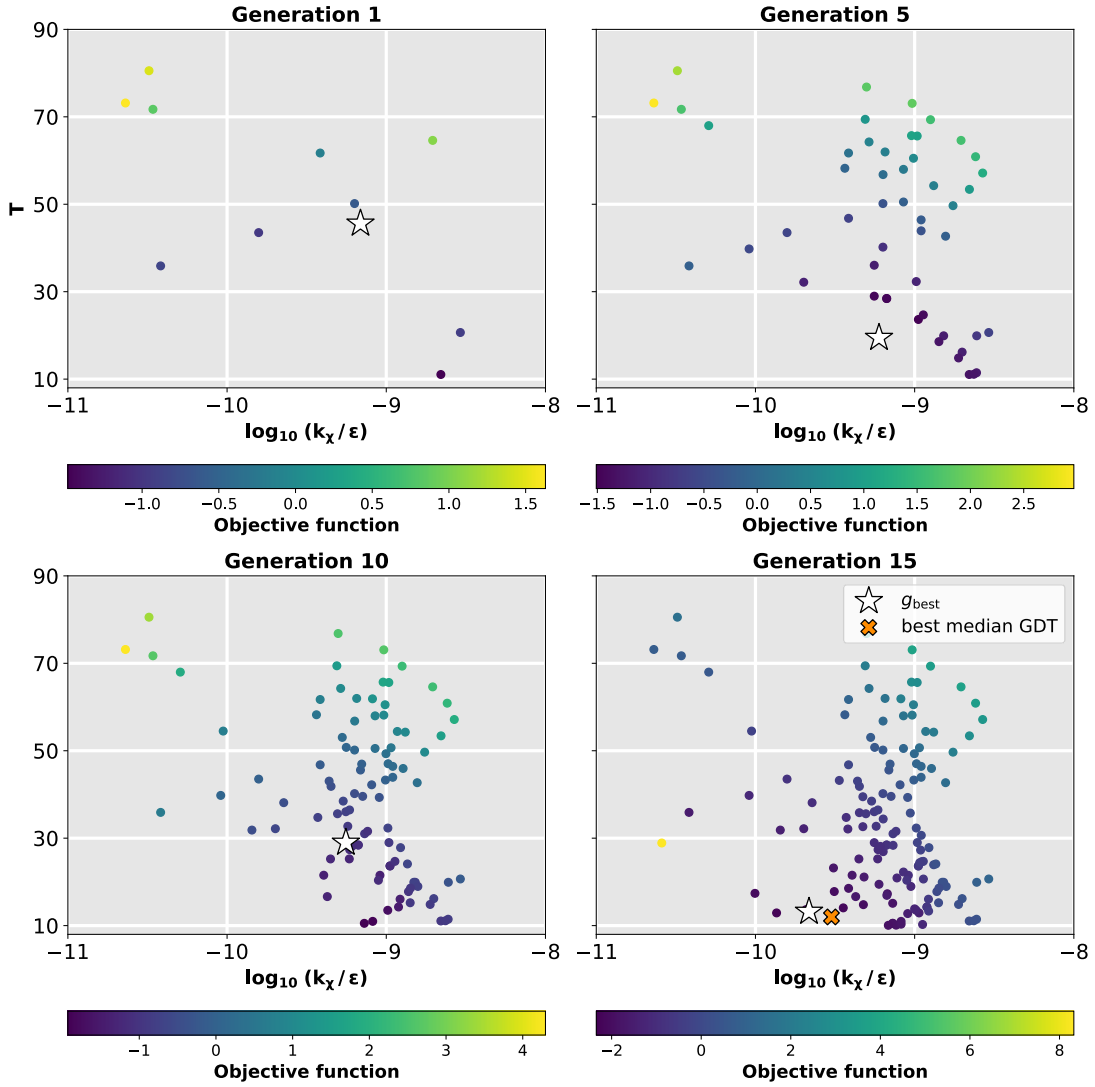


Figure 6.4. Dynamically evolving OF topology. Results are shown for LAO protein's h → a transition (seed 1790954). The current global best position is marked. k_χ, bias weight, T, temperature, g_{best}, current global best, GDT, global distance test. Reproduced from Ref.[173] under CC BY 4.0.

variational study in Fig. 5.13). It is conspicuous that χ^2_{med} by itself has the best Pearson correlation with the median GDT. Its two-branch behavior is only slightly pronounced. In contrast to LAO protein, the structural transition of ADK changes its overall molecular shape considerably. This leads to a reduced ambiguity in the scattering data, thus making the χ^{-2}_{av} regularizer less important for this system. I nevertheless find my OF to be a better surrogate in the interesting high-GDT region than the plain χ^2_{med} as it could resolve the remaining ambiguity in the latter entirely. As shown in Supplementary Fig. E.2, it performed also better than an OF without the regularizer, comprising only the standardized REF15_{av} and χ^2_{med}. In light of these findings, my OF proved to be a universal and robust surrogate of an `XSBM` simulation's quality.

I find global best positions, g_{best}, to return functional MD parameter combinations throughout. For LAO protein's h → a transition, g_{best} median GDTs consistently were in the order of 70 and correspond well to the best values reached. This means that for half of the simulated structures at least 70% of all C_α atoms lie within a small radius from their positions in the target state. These results indicate structural accuracy of the simulated ensemble for g_{best} and thus convergence to the target state and

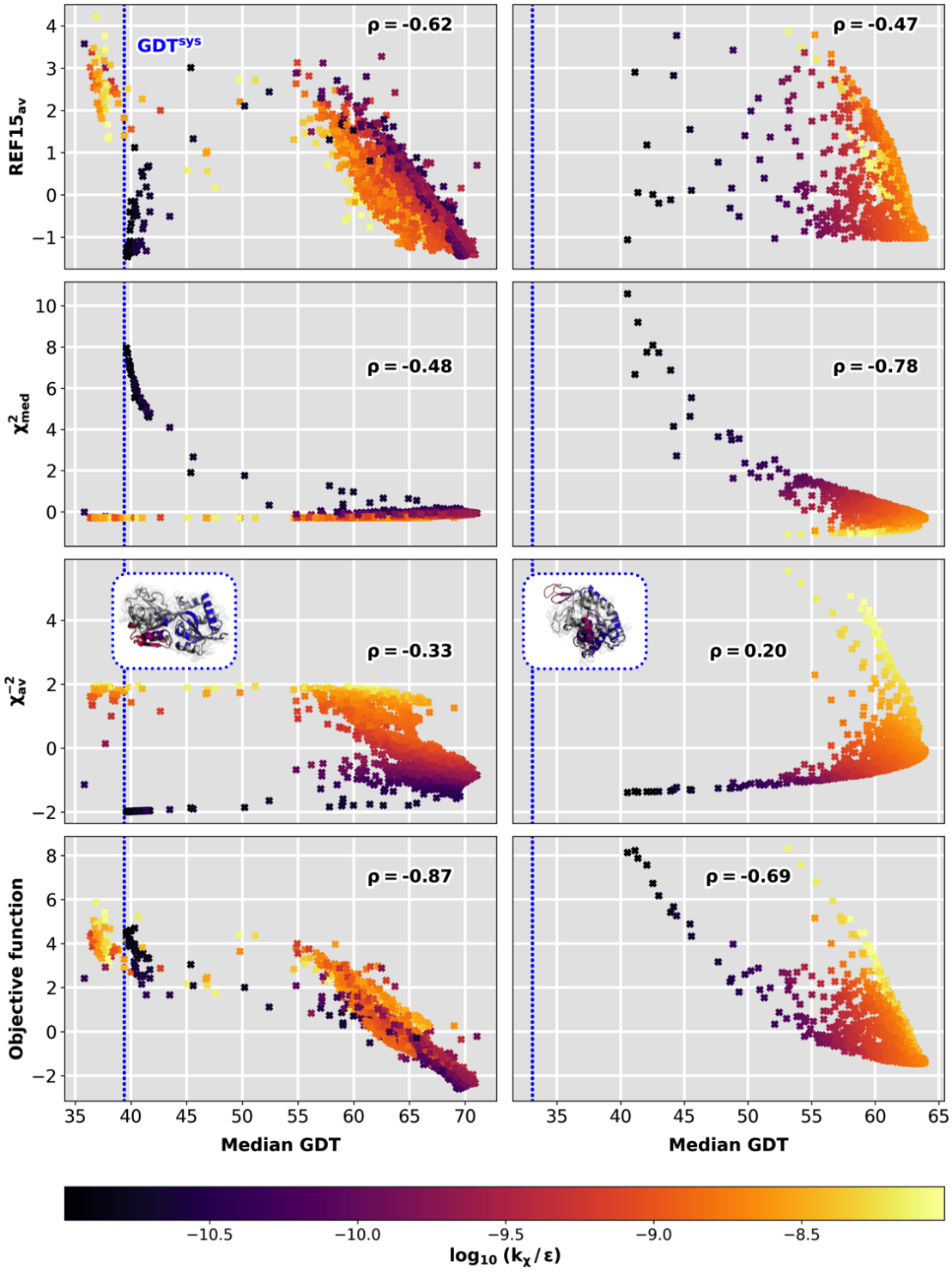


Figure 6.5. OF and its standardized responses versus median GDT. LAO protein's h → a transition (left) and ADK's o → c transition (right). GDT^{sys}, GDT between initial and target structure of each test system, ρ, Pearson correlation of each quantity and median GDT, k_χ, bias weight. Derived from Ref.[173] (extended from original by additional data), used under CC BY 4.0.

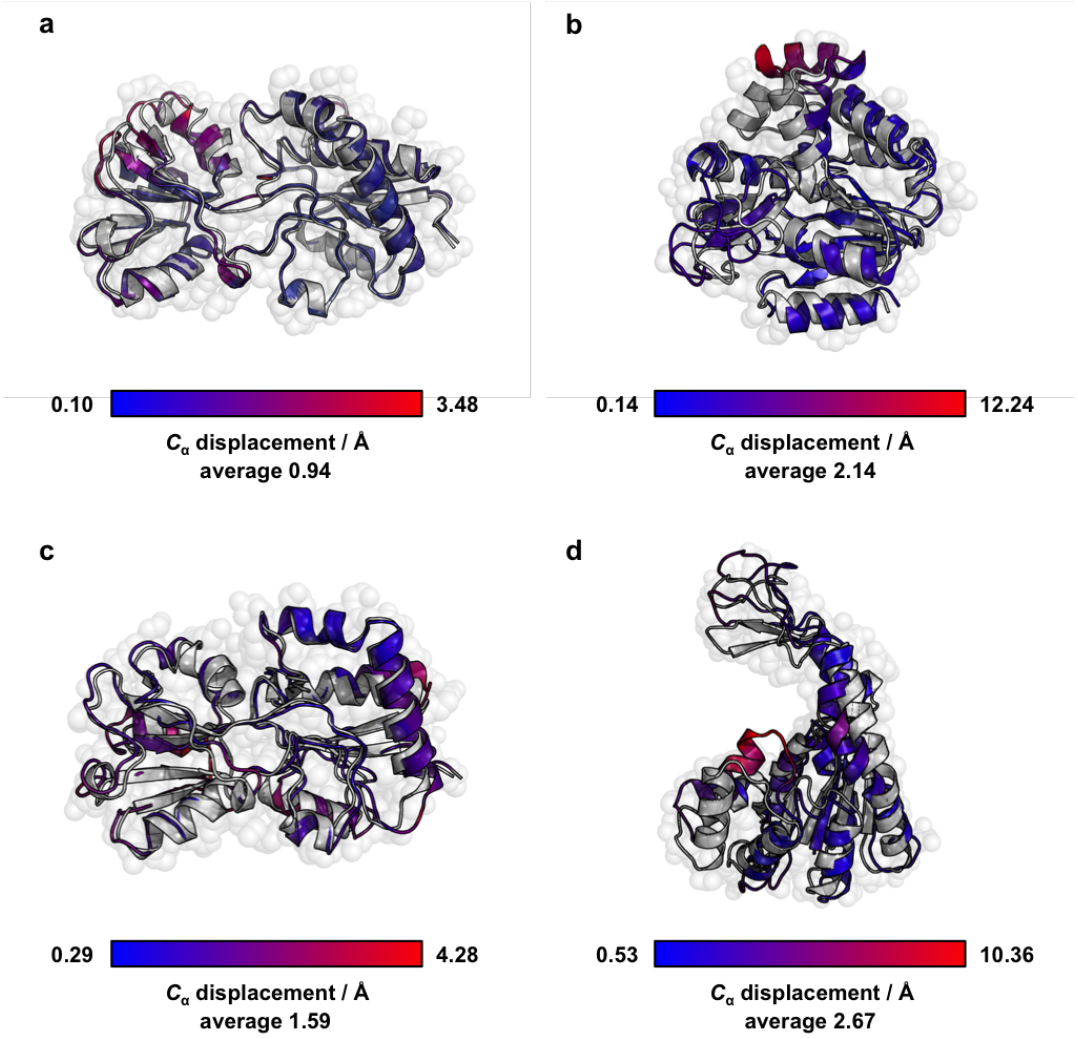


Figure 6.6. Representative structures from global best simulations. **a**. LAO protein's h → a transition (seed 1790954), **b**. ADK's o → c transition (seed 1795691), **c**. LAO protein's a → h transition (seed 1800994), and **d**. ADK's c → o transition (seed 1801030). Structures with maximum GDT are shown (colored) and almost identical to respective target states (gray). The coloring indicates the displacement of each alpha carbon in the simulated structure with respect to the target state. Structures visualized in PyMOL[1]. Derived from Ref.[173] (extended from original by additional information), used under CC BY 4.0.

successful refinement against the data. The same is true for ADK's o → c transition, where g_{best} and the best median GDTs are around 63 and more similar than those of LAO protein. Exemplary structures from g_{best} simulations shown in Fig. 6.6 a and b are in nearly perfect accordance with the respective target states.

In addition, I studied the reversed conformational transitions. Results of LAO protein's a → h transition and ADK's c → o transition can be found in Supplementary Tables E.3 and E.4, respectively. Representative structures are shown in Fig. 6.6 c and d. With ρ between -0.56 and -0.88, my OF could also identify functional MD parameters for `XSBM` simulations of LAO protein's a → h transition in `FLAPS`. The OF and its standardized responses as functions of the median GDT reveal a similar behavior as observed for the h → a transition (see Supplementary Fig. E.1). g_{best} and the best median GDTs were consistently slightly below 70, indicating high similarity of simulated structures and the desired target state for the MD parameter combinations found. For ADK's c → o transition, Pearson correlations up

to -0.53 and g_{best} median GDTs of approx. 45 indicate only average structural accuracy. The overall best median GDTs were around 50. I observed a cluster of `XSBM` simulations with low bias weights that produced structurally accurate ensembles despite having a comparably high χ^2_{med} (see Supplementary Fig. E.1 right). These outliers induce an upward shift in parts of the OF as a sum of the three standardized responses, thus hindering the desired relationship with the median GDT. Dropping the regularizing χ^{-2}_{av} in the OF leads only to a minor improvement, and the ambiguity still cannot be fully resolved (see Supplementary Fig. E.2 right). With only half of all C_α atoms lying within a small radius from their target positions, I assume that the observed behavior is not a problem of `FLAPS` itself but due to limits of the `XSBM` method[27]. Using `CRYSOL` intensities including realistic errors, the information in the simplistic SBM and the low-resolution scattering data seems to be insufficient to determine the molecular structure with the same accuracy as for the other test cases. The parameters from the variational study presented in Fig. D.26 yielded even worse results with the `CRYSOL` data. However, even under these circumstances, `FLAPS` was capable of determining the best possible parameters. With a maximum GDT of 58.41 and a minimum RMSD of 3.16 Å, the resulting structures nevertheless were topologically accurate and the parameters found thus quite useful.

Discussion

The inverse problem of reconstructing molecular structures from low-resolution SAXS data is still unsolved. Biomolecular simulations are among the most powerful tools to eliminate the arising ambiguity and access the valuable structural information content of such data. However, data-assisted simulations rely on MD parameters, where above all experimental information must be weighted accurately with respect to the physical model. Here, I showed how computational intelligence can be used to systematically explore MD parameter spaces and optimize the performance of complex physics-based simulation techniques. I introduced `FLAPS`, a data-driven solution for fully automatic and reproducible parameter search based on particle swarms.

To identify the best MD parameters for SAXS-guided protein simulations, I designed an OF as an accurate surrogate of simulation quality in terms of physical structures matching the target data. A suitable OF will depend on multiple quality features of different scales to equally reflect a data-assisted simulation's physical plausibility and its agreement with the data. To handle multiple responses in classical PSO, they need to be mapped to a scalar score via multiplication by fixed weights. These additional OF parameters must be either chosen manually (and likely suboptimally) in advance or absorbed into the search space, resulting in a massive increase of its dimensionality. `FLAPS` solves this problem by intelligently learning OF parameters in the optimization process, preventing the need to set them as "magic numbers" while reducing the search-space dimensionality to a minimum. Various responses are automatically balanced with respect to each other to enable a meaningful and unbiased comparison on a shared scale. Implemented in `FLAPS`, my conceptual OF reliably identified useful MD parameters for two different protein systems, where I observed convergence of simulated structures to the target state. Due to a dedicated in situ processing, the algorithm leverages nowadays available computing resources and transparently scales from a laptop to current supercomputers.

In previous studies, the bias weight is conveniently chosen as the smallest value yielding satisfactory χ^2 [16,27]. This criterion is purely data-based and neglects the physical information provided by the molecular simulations, risking the selection of physically dysfunctional values due to the ill-posedness of the SAXS inverse problem. The flexible composite OF in `FLAPS` allows to include multiple selection criteria, yielding a direct surrogate model of simulation quality with respect to not only data conformity

but also physical plausibility of simulated structures. In contrast to grid search, the optimization does not rely on a predefined list of manually chosen and a priori fixed candidates. This circumvents missing an optimum, spanning large search spaces due to fine-grained grids, or both. Instead, a directional search led by the best solutions found so far provides guidance in the selection process and a valid context for meaningful interpretation of multi-layer criteria. In addition, decreasing the dimensionality reduces the compute time required to find useful parameters.

`FLAPS` can easily be transferred to optimize other biophysical applications, e.g., MDfit[13], a simulation method incorporating data from X-ray crystallography, cryogenic electron microscopy, and biochemical studies. In a more general context, `FLAPS` solves the problem of weighting different contributions in all kinds of composite OFs encountered in multi-response optimization. Such problems frequently occur in industrial manufacturing, processing, and design, and manually chosen weights are commonly used[197,198]. In `FLAPS`, OF parameters can be learned in a self-improving manner according to any desired scheme, e.g., standardization, rescaling, mean normalization, or a relative weighting after one of the aforementioned steps. Taking the example of computational biophysics and structural biology, `FLAPS` shows how computational-intelligence concepts can successfully be harnessed for practical optimization problems at the forefront of life sciences.

PART III

CONCLUSIONS

"No matter where you go, there you are."

Yogi Berra

7

Summary

Time to wrap up! This thesis presented my work in deriving realistic protein structures in silico by integrating experimental data into biomolecular simulations. In this chapter, I will recapitulate and evaluate the final outcomes of my work. To better understand and contextualize the significance of the computational methods I developed, I will give a summary of my projects with a clear focus on major strengths and limitations. I will close this chapter by discussing prospects for future directions of research and applications.

Proteins are the cellular nanomachines in our bodies and support nearly all vital functions on the molecular level. As their biological function is largely determined by their malleable shape, a comprehensive understanding requires us to capture both their static structure and their function-related dynamics. When proteins lose their normal structure, they can become dysfunctional or even toxic, thus disrupting the healthy function of our cells, tissues, and organs. Profound insights into protein structure and dynamics are invaluable for retracing how abnormal molecular interactions cause severe diseases, such as Alzheimer's and Parkinson's, and eventually developing effective treatments for them.

The various methods for protein structure analysis include both experimental and computational approaches. However, each individual method typically provides only partial information. Experimentally determined crystal structures and computationally predicted structural models fail to capture the intrinsic dynamism of proteins. Biological small-angle X-ray scattering can yield information on protein dynamics, however only in the form of separate ambiguous snapshots averaged over the conformational ensemble. Despite massive advances in molecular imaging and structure prediction, we still lack a general understanding of protein function on the molecular level. Here, physics-based modeling and simulation of molecular dynamics can help fill the gap. Employing traditional concepts from classical mechanics, static

protein structures can be enriched with an atomically resolved view of their dynamics. Data-assisted MD has emerged as a new standard to unite the structural puzzle pieces from various sources and obtain a complete picture of protein structure and function on the molecular level[16–18]. Such simulations incorporate the experimental data as an integral component into a biased physical model. They are among the most powerful tools to access and interpret the limited information in the data and eliminate ambiguities with their complementary knowledge. In return, the data can compensate for shortcomings of the physical model and accelerate conformational changes.

Based on empirical physical laws, classical molecular mechanics provides a descriptive model of atomic interactions. This richness of detail makes the simulations infeasible for studying large-scale motions, even with an experimentally derived bias. A more efficient alternative is reducing a protein's degrees of freedom by simplifying its interactions in the force field. This minimalist approach is implemented in structure-based models. While SBMs use the same bonded interactions as classical molecular mechanics, the crucial difference lies in the non-bonded interactions categorized as either native and favorable or non-native and unfavorable. With the native structure encoding a protein's functional dynamics, SBMs provide a simple model of the behavior emerging from its evolutionarily optimized topology. Built around a minimal set of initial assumptions, SBMs are particularly useful to discover mechanistic insights and gain intuition on biomolecular systems. It is important to note that models are always an approximation of the exact situation, and choosing a suitable model requires the specification and iterated revision of the input assumptions of the studied system as well as a clear definition of the research goal.

In this thesis, I explored the capabilities and limits of SBMs as a tool to access and complement the fragmentary information in structural data on proteins. I worked towards the development and application of a self-contained pipeline for systematically deriving realistic protein structures in accordance with the data from SAXS. I presented `XSBM`, a structure-based simulation framework for rapidly interpreting SAXS data in the form of physically reasonable protein models[27]. I demonstrated my method's efficacy for three proteins using minimal computational resources and time[27]. Integrating the data on the level of the force field, `XSBM` could probe real protein transitions and guide the systems between different conformations. Note that studying other proteins may require more flexible formulations, such as the inclusion of additional conformations into the SBM[23,65]. `XSBM` produced equally accurate results as classical molecular-mechanics approaches while reducing compute demands by up to two orders of magnitude[27]. Its efficiency is a direct result of the underlying energy landscape where, unlike in classical molecular mechanics, the lowest-energy excited states reflect function-related conformational changes.

As a data-assisted method, `XSBM`'s performance crucially depends on the highly non-trivial choice of MD parameters, where the key challenge is balancing experimental information and theoretical knowledge in the combined force field. I showed how computational intelligence can be used to explore such parameter spaces efficiently and optimize the performance of physics-based simulation techniques. I introduced `FLAPS`, a data-driven algorithm for fully automated and reproducible parameter search for biomolecular simulations. Inspired by the natural behavior of bird flocks and fish schools, `FLAPS` is a self-adapting particle-swarm based optimizer that solves the problem of weighting various quality features in multi-response optimization. `FLAPS` balances the multiple responses in composite objective functions by refining their relative weights at runtime, thus enabling a meaningful and unbiased comparison on a shared scale. Note that the resulting dynamism in the OF can cause convergence problems if no stable topology is reached. To identify functional `XSBM` parameters, I designed a flexible OF that equally reflects both the agreement with the target data and the physical plausibility of a simulation using a particular set of parameters to be optimized. Previously, the latter were selected in a grid-search or purely data-based manner, neglecting the information provided by the physical model[16,27]. In contrast, a directional search led by the current best solutions guides the selection process in `FLAPS` and provides a meaningful

context for multilayer criteria. As a showcase example, I applied `FLAPS` to determine functional `XSBM` parameters for two protein systems. I validated the OF as a robust and reliable surrogate of simulation quality in terms of physical structures matching the data. `FLAPS` transparently and fully automatically identified functional parameters for both systems, where I observed convergence of simulated structures to the target state. Using the example of `FLAPS`-optimized `XSBM` simulations, I have illustrated how advanced computational-intelligence concepts can successfully be harnessed for practical optimization problems in biophysics.

The latest successes in molecular imaging and computational protein modeling have raised various starting points for further research. Most urgently, there is a need for piecing together the growing variety of available structural information. The combined framework of `FLAPS`-optimized `XSBM` simulations provides a suitable basis for this. `XSBM` can easily be extended towards including data from various sources. The question of how to balance the different contributions in such a hybrid force field can be solved efficiently with `FLAPS`. Further developments towards refining protein structures in data-assisted structure-based simulations include expanding single-basin SBMs to multi-Gō models[65], testing other forms of the bias potential[12], and enhancing the framework towards multireplica ensemble refinement[16]. The rapid growth of the available experimental data promises a prosperous future by merging the entire spectrum of structural information into a complete picture of dynamic protein structure. Physics-based modeling allows us to expand static structures with a dynamic view and connect the different experimental methods in silico. Together with recent advances in structure prediction through artificial intelligence, data-assisted protein simulations can help push our understanding of the intricate relation between biomolecular structure and function by combining the multidisciplinary expertise of experimenters and modelers. Computational approaches as the one presented in this thesis provide a powerful means to fit the available puzzle pieces of structural information together, thus deepening our understanding of proteins as the ultimate building blocks of life. As the famous Chemistry Nobel laureate Linus Pauling put in a nutshell[1],

"Life... is a relationship between molecules."

[1] Quoted In T. Hager, *Force of Nature: The Life of Linus Pauling* (1997), 542.

PART IV

APPENDICES

A

Supplementary Information

A.1 Derivation of the Debye Equation

The atoms of a molecule illuminated by a monochromatic plane wave interact with the incident radiation and become sources of spherical waves in return[9]. The scattering amplitude for a single scattering event by an atom at position $\mathbf{r}_j$ is

$$A_j(\mathbf{q}) = f_j(q)\,\exp(i\mathbf{q}\cdot\mathbf{r}_j)\,. \tag{A.1}$$

f_j is the atomic form factor, a measure of the amplitude of a wave scattered from an isolated atom, and $\mathbf{q}$ the momentum transfer with $|\mathbf{q}| = q$. The interfering amplitudes of all N atoms are summed up to the overall molecular scattering amplitude,

$$\begin{aligned} A(\mathbf{q}) &= \sum_{j=1}^{N} A_j(\mathbf{q}) = \sum_{j=1}^{N} f_j(q)\,\exp(i\mathbf{q}\cdot\mathbf{r}_j) \\ &\equiv \int_{\mathrm{V}} f_e(q)\,\rho(\mathbf{r})\,\exp(i\mathbf{q}\cdot\mathbf{r})\,d\mathbf{r} = f_e(q)\,\mathfrak{F}[\rho(\mathbf{r})]\,. \end{aligned} \tag{A.2}$$

In the continuum limit, the scattering amplitude from an ensemble of atoms turns out to be the Fourier transform, $\mathfrak{F}[\cdot]$, of the scattering length density distribution, $\rho(\mathbf{r})$, weighted by the form factor, $f_e(q)$. The scattering intensity is

$$I(\mathbf{q}) = A(\mathbf{q})\cdot A^*(\mathbf{q}) = |A(\mathbf{q})|^2\,. \tag{A.3}$$

For X-rays, the scattering arises from the interaction between the incident wave and the electron clouds of the atoms. This means that the scattering power scales with the electron density. Assuming the Born

approximation to be valid, the corresponding form factor $f_e(q)$ is the Fourier transform of the molecule's normalized electric charge distribution. The amplitude, however, is experimentally inaccessible. For spatially isotropic particle distributions, the measured intensity is a rotational average in reciprocal space, representing the distinct molecular orientations in the solid angle Ω[205]:

$$I(q) = \langle I(\mathbf{q}) \rangle_\Omega \tag{A.4}$$

Using

$$\langle \exp(i\mathbf{q} \cdot \mathbf{r}) \rangle_\Omega = \frac{\sin(qr)}{qr} \tag{A.5}$$

yields the well-known Debye formula, which states the spherically averaged intensity for a particle described as discrete sum of elementary scatterers, e.g., atoms with form factors $f_j(q)$, to be[39]

$$I(q) = \sum_{j,k} f_j(q)\ f_k(q)\ \frac{\sin(qr_{jk})}{qr_{jk}}. \tag{A.6}$$

$r_{jk} = |\mathbf{r}_j - \mathbf{r}_k|$ is the distance between atoms at positions $\mathbf{r}_j$ and $\mathbf{r}_k$.

Spherical Harmonics Expansion

For spherical averaging in Eq. A.6, it is advantageous to introduce the multipole expansion in spherical coordinates by Stuhrmann (1970)[206]. Inserting the relation

$$\exp(i\mathbf{q} \cdot \mathbf{r}) = 4\pi \sum_{l=0}^{\infty} \sum_{m=-l}^{l} i^l j_l(qr)\, Y_{lm}^*(\omega)\, Y_{lm}(\Omega)\ , \tag{A.7}$$

where $j_l(qr)$ denote the spherical Bessel functions and $Y_{lm}(\Omega)$ the spherical harmonics, into the expression for the particle scattering in Eq. A.2 yields[206]

$$A(\mathbf{q}) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} A_{lm}(q)\, Y_{lm}(\Omega). \tag{A.8}$$

$A_{lm}(q)$ are the partial amplitudes given by[206]

$$A_{lm}(q) = 4\pi i^l \sum_{j=1}^{N} f_j(q)\, j_l(qr_j)\, Y_{lm}^*(\omega_j)\,. \tag{A.9}$$

Since the spherical harmonics form a complete set of orthogonal functions on the sphere, all cross terms cancel when taking the spherical average. As a result, one obtains a compact expression for the overall intensity

$$I(q) = \sum_{l=0}^{L} \sum_{m=-l}^{l} |A_{lm}(q)|^2\,. \tag{A.10}$$

The truncation value L determines the resolution of the particle representation. Similar expressions can be set up for the excluded volume and border-layer scattering. Apart from the Debye formula (see Eq. A.6) and the spherical harmonics expansion, common SAXS-averaging methods include numerical or spherical quadrature, Monte-Carlo sampling, the Cubature formula, or Zernike polynomials along with atomic, grid, and coarse-grained structural models[10].

A.2 Calculating SAXS Profiles from Protein Structures with CRYSOL

The popular PDB-oriented program CRYSOL[41] is an implicit-solvent method for evaluating the solution X-ray scattering from atomic macromolecular structures. It is part of the ATSAS[207,208] tool box for small-angle scattering data analysis.

CRYSOL uses multipole expansion to efficiently evaluate spherically averaged scattering patterns from biomolecular structures (see Supplementary Sec. A.1)[41]. The solvation shell is approximated by a border layer of 3 Å effective thickness with a 10 % to 15 % excess density with respect to the average density of free bulk water, $\rho_0 = 0.334\,e\text{Å}^{-3}$[41]. Originally, the solvation shell is represented by an envelope function[41]. As the accuracy of this representation may be limited for complex shapes, a newer version, CRYSOL 3[207], represents the solvation shell as dummy beads covering the molecular surface. The beads are divided into three classes whose mobility and thus contrast may vary depending on the location, i.e., internal water within cavities, the water shell on the outer convex surface, and water on the concave surface. This more sophisticated handling permits a better prediction of the wide-angle scattering[207].

The program can perform a fit of the calculated scattering curve to an experimental curve by minimizing the χ^2 discrepancy under variation of three parameters, i.e., the average displaced-solvent volume per atomic group, the solvation shell's contrast, and the relative background[41]. This dependency on several free parameters is a major downside of such commonly used implicit-solvent methods as it increases the risk of overfitting. However, CRYSOL has been shown to adequately evaluate SAXS profiles up to a momentum transfer of approximately $0.4\,\text{Å}^{-1}$[41]. Alternative explicit-solvent based methods have only one fitting parameter (see Sec. 5.1) but come at the cost of considerably increased computational demands[152].

B

Appendix to "PROJECT: Simulating the Mavirus Capsomer with Structure-Based Models"

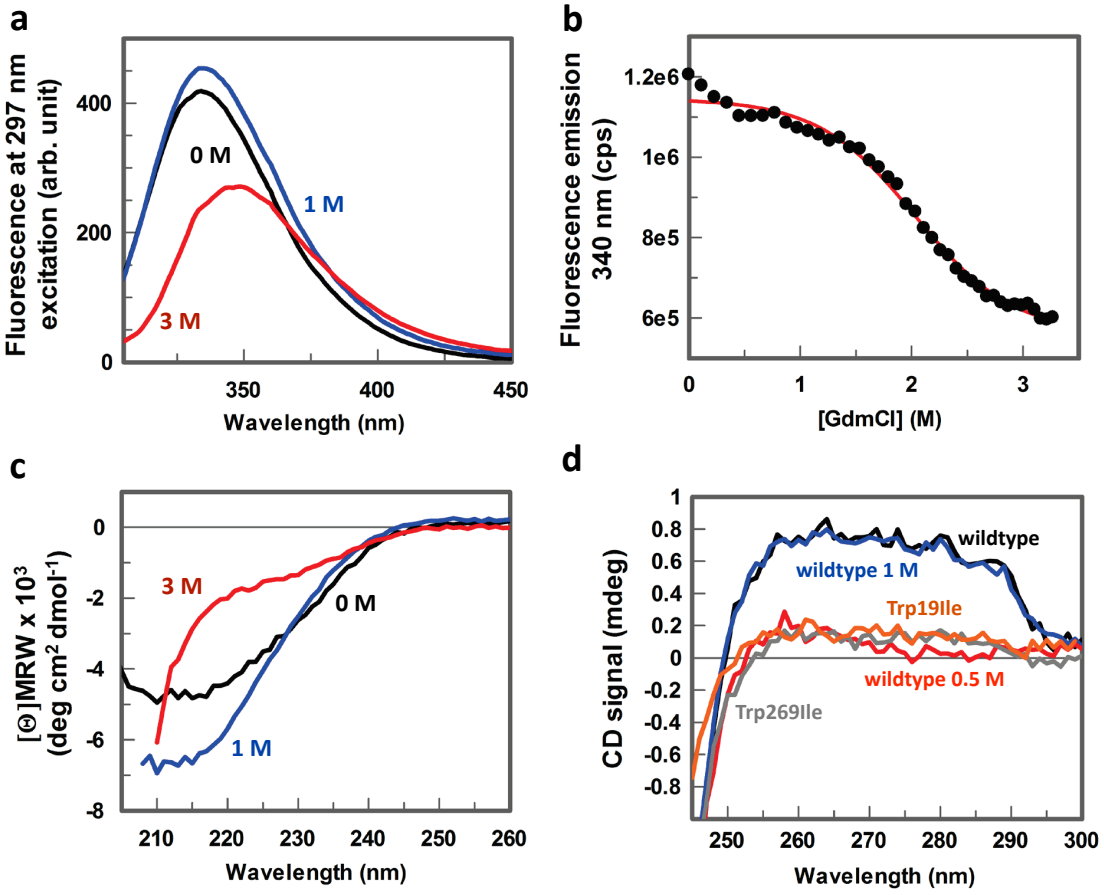


Figure B.1. CD and fluorescence spectra of MCP Trp19Ile at increasing [GdmCl]. Equilibrium spectra of 2.5 μM MCP Trp19Ile at different [GdmCl] for **a**. fluorescence (297 nm excitation) and **c**. far-UV CD. **b**. Chemical denaturation of MCP Trp19Ile after excitation at 297 nm followed at 340 nm. **d**. Near-UV CD spectra of MCP_3 wildtype (black), Trp269Ile (grey), and Trp19Ile (orange), recorded in 1 cm path-length cuvette with a protein concentration of 1 μM. For wildtype, additional spectra at 0.5 M (red) and 1 M (blue) GdmCl were recorded. Reproduced with permission from Ref.[28] under CC BY-NC-ND 4.0 (relabeled from original).

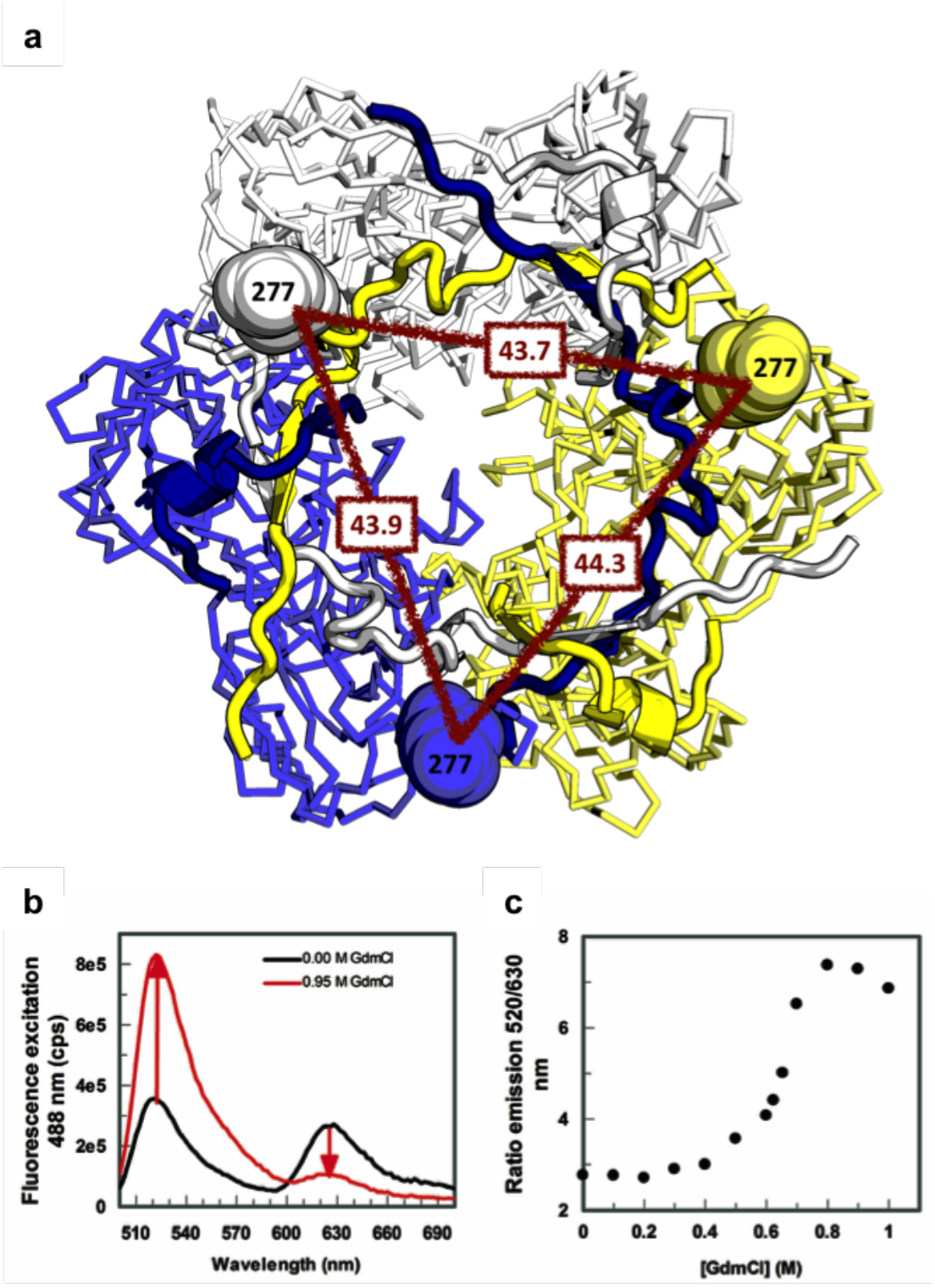


Figure B.2. FRET probe positions, equilibrium spectra, and GdmCl denaturation curve. **a**. FRET probes in MCP_3 with the mutated aspartic acid residues at sequence position 277 shown as spheres. The C_α distances to the corresponding residues in neighboring protomers are given in nm. The N-terminal arms and C-terminal clasps are highlighted as cartoons. Visualized in PyMOL[1]. **b**. Fluorescence spectra of dye-labeled MCP_3 at different [GdmCl] after donor excitation at 488 nm. **c**. FRET signal as function of [GdmCl] calculated as ratio of emission 520 nm / 630 nm at 488 nm excitation. Reproduced with permission from Ref.[28] under CC BY-NC-ND 4.0.

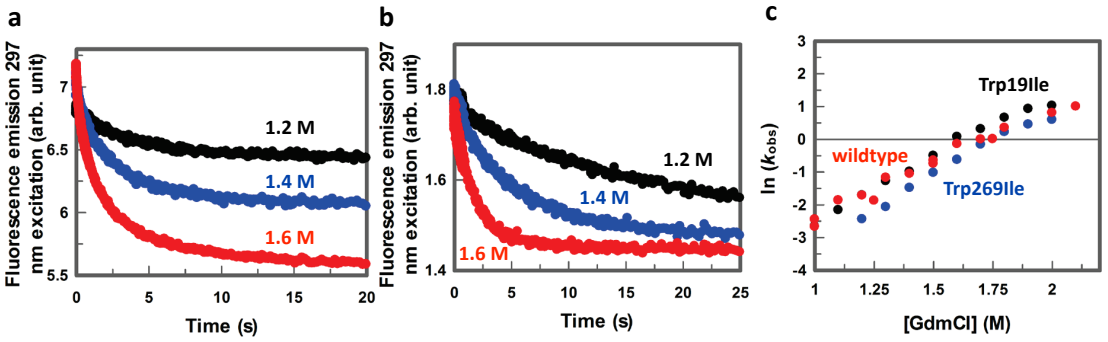


Figure B.3. Stopped-flow kinetics for MCP Trp19Ile and Trp269Ile. Exemplary kinetic traces of **a**. MCP Trp19Ile and **b**. MCP Trp269Ile after 297 nm excitation using a long pass filter (> 320 nm) for jumps to different [GdmCl], showing only one phase and a decrease in fluorescence intensity. **c**. Derived rate constants for the MCP mutants and the slow phase of MCP wildtype. Reproduced with permission from Ref.[28] under CC BY-NC-ND 4.0 (relabeled from original).

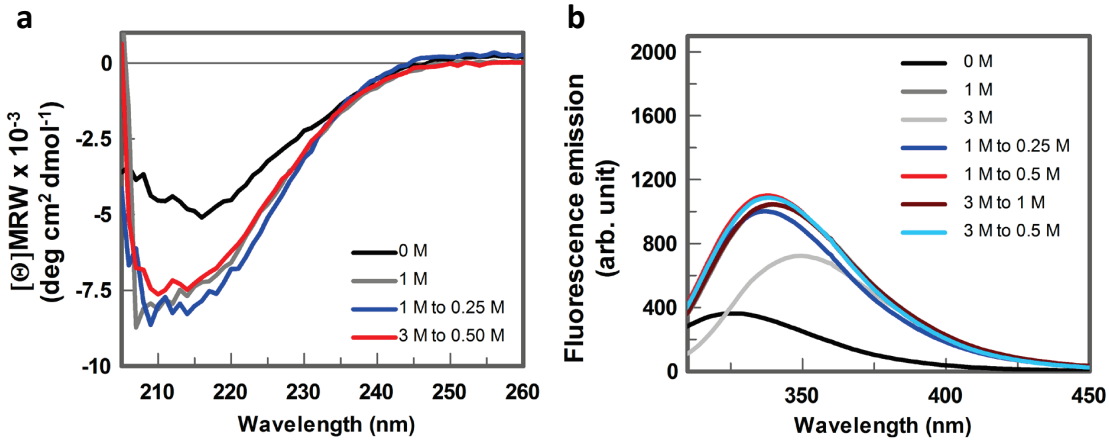


Figure B.4. Equilibrium CD and fluorescence spectra of MCP after refolding from 3 M GdmCl. We incubated the samples at either 3 M or 1 M GdmCl for 10 min and diluted them with buffer (pH 5) to a final [GdmCl] as indicated in the legend. After dilution, the final MCP concentration was 2.5 μM. **a**. CD spectra and **b**. fluorescence spectra (297 nm excitation). Reproduced with permission from Ref.[28] under CC BY-NC-ND 4.0.

C

Appendix to "PROJECT: Simulating the Interplay of FRET and SAXS with Structure-Based Models"

C.1 Dye-Labeled Proteins

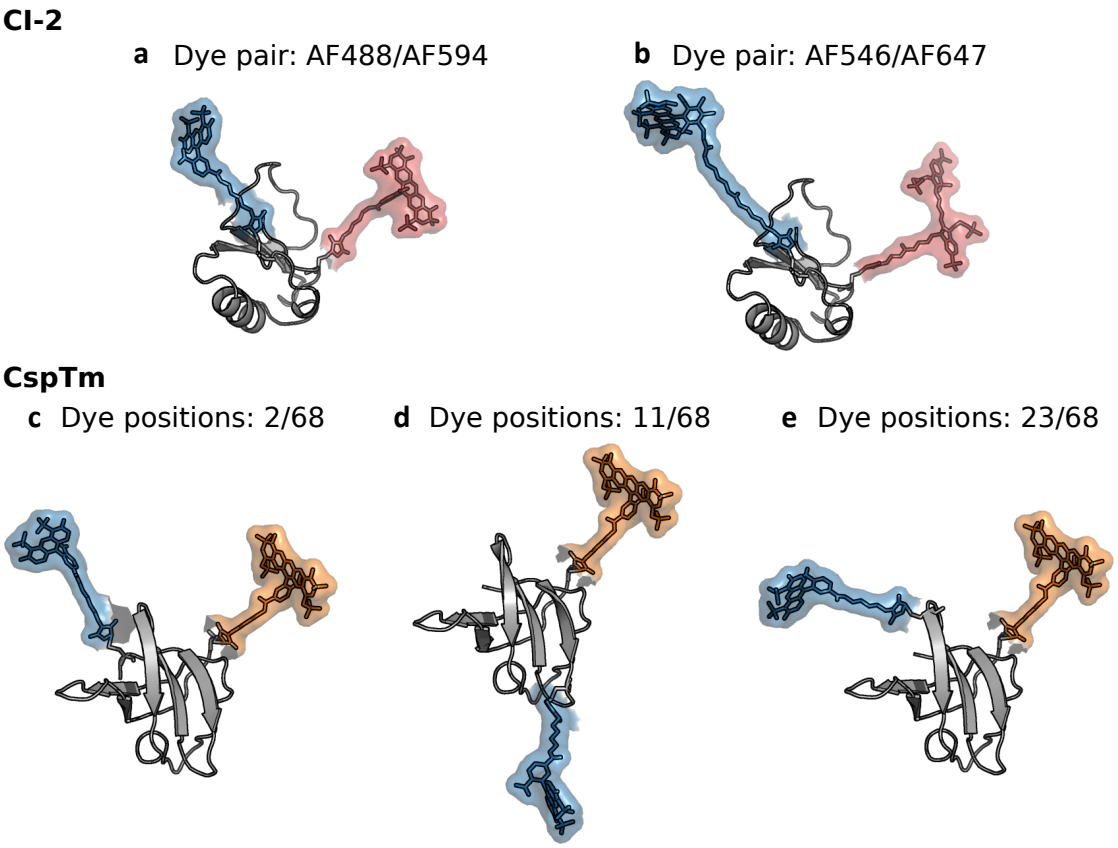


Figure C.1. CI-2 and CspTm with different dye pairs and labeling sites. **a**., **b**. CI-2 with AF dyes attached to residues 20 and 78. **c**., **d**., **e**. CspTm with AF dyes attached to given residues. Reproduced from Ref.[29] under CC BY 4.0.

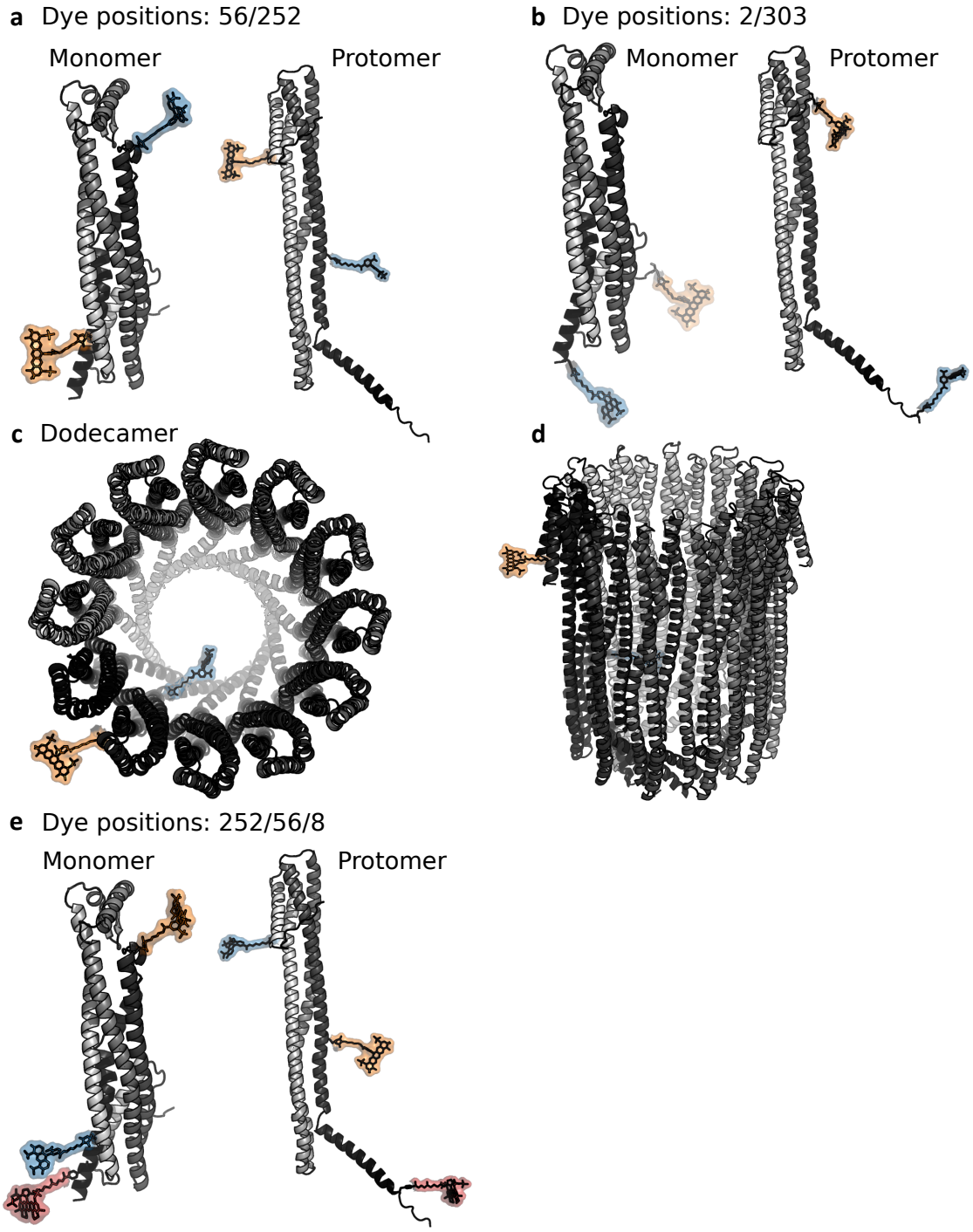


Figure C.2. Different ClyA conformations. **a**. Monomer and **b**. protomer conformations with AF dyes attached to the given residues. **c**. Top and **d**. side view of the dodecamer with AF dyes attached to residues 56 and 252. **e**. Monomer and protomer with three dyes attached to given residues. Reproduced from Ref.[29] under CC BY 4.0.

C.2 Radius of Gyration Analysis

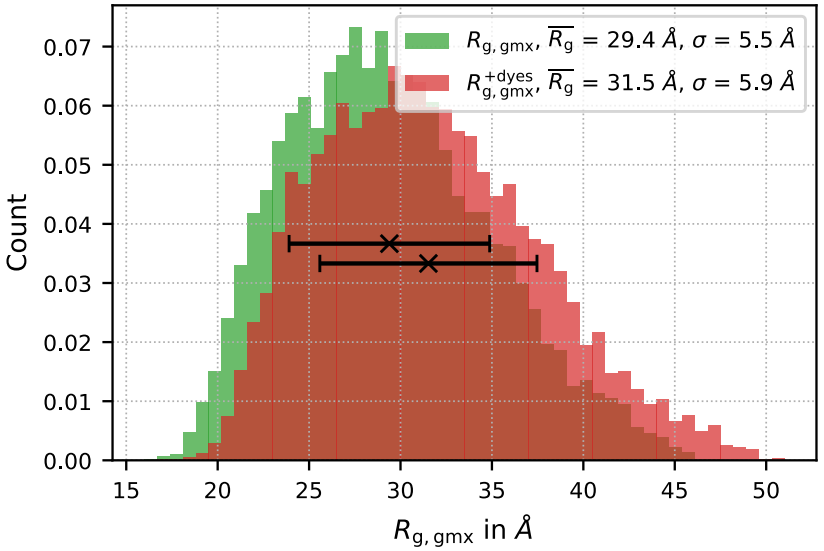


Figure C.3. $R_{g,gmx}$ distributions for folded ^{10}FNIII without (green) and with dyes (red). The narrower green distribution with its small standard deviation σ (black error bars) indicates a defined conformational ensemble for the system without dyes. The broader red distribution points to a more diverse ensemble for the (according to its mean $\overline{R_g}$) larger dye-labeled system. Reproduced from Ref.[29] under CC BY 4.0.

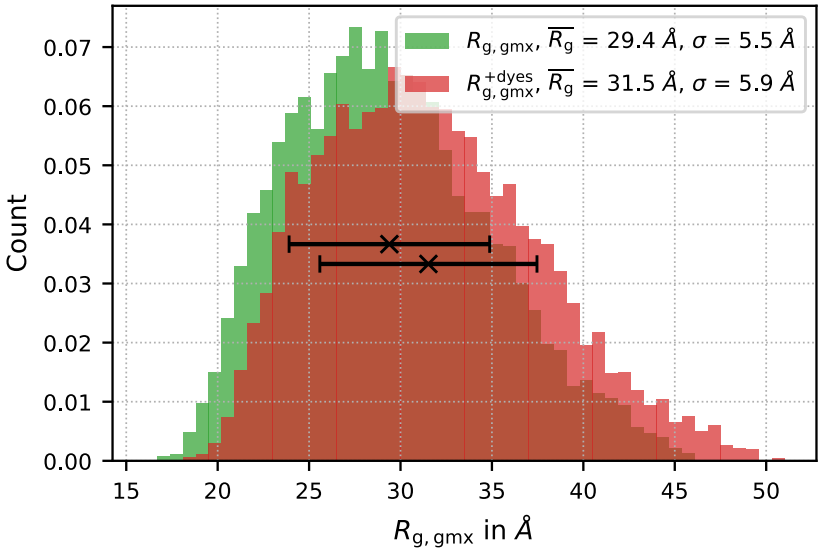


Figure C.4. $R_{g,gmx}$ distributions for unfolded ^{10}FNIII without (green) and with dyes (red). As anticipated, the distributions without and with dyes are almost identical for the unfolded system. The comparably larger size of the dye-labeled system is reflected by the slightly shifted mean $\overline{R_g}$ and the slightly different standard deviation σ (black error bars). Reproduced from Ref.[29] under CC BY 4.0.

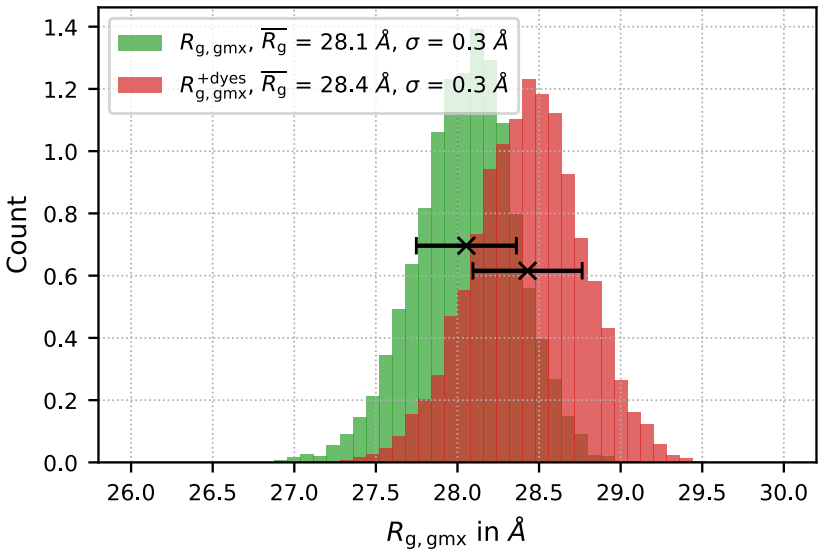


Figure C.5. $R_{g,gmx}$ distributions for folded ClyA monomer 56/252 without (green) and with dyes (red). As the dyes are comparatively smaller for ClyA monomer, their influence is less pronounced than for [10]FNIII. Nevertheless, analogous tendencies, such as a larger mean for the dye-labeled systems, are clearly visible. Reproduced from Ref.[29] under CC BY 4.0.

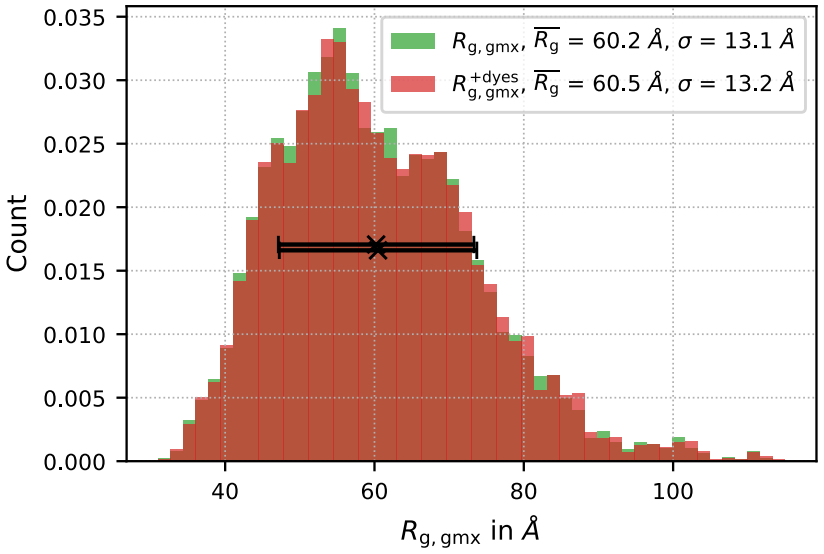


Figure C.6. $R_{g,gmx}$ distributions for unfolded ClyA monomer 56/252 without (green) and with dyes (red). In analogy to [10]FNIII, the distributions without and with dyes are virtually identical for the unfolded system. As the dyes' influence is less pronounced for the larger ClyA monomer, the distributions are even more similar. Reproduced from Ref.[29] under CC BY 4.0.

C.3 Solvation Shell Contrast Analysis

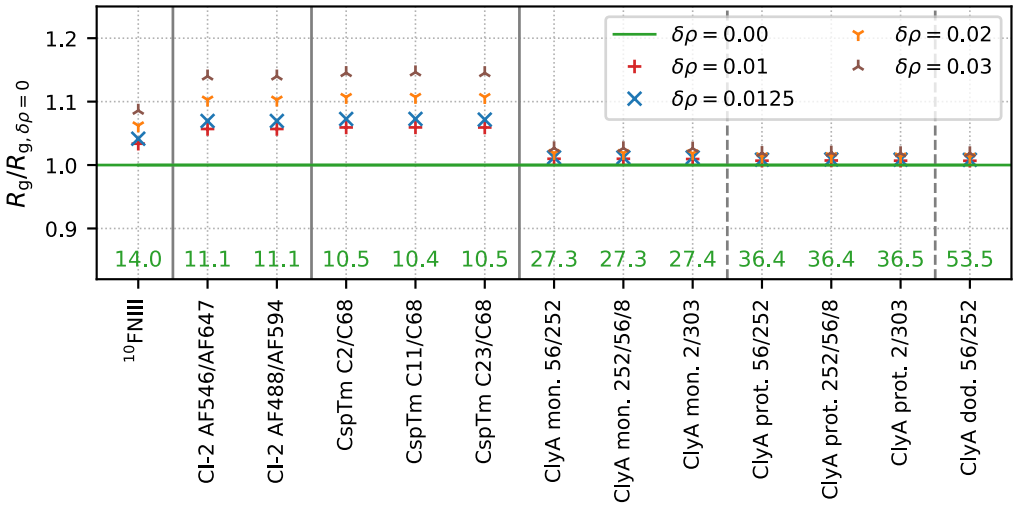


Figure C.7. Folded systems without dyes: R_g values for different solvation shell contrasts $\delta\rho$ (in e Å^{-3}) with respect to $R_{g,\delta\rho=0}$, given at the bottom in Å. We studied ^{10}FNIII, CI-2 with two different AF dye pairs, CspTm with AF dyes at three different labeling positions, and ClyA monomer, protomer, and dodecamer with two and three dyes at different labeling sites. Reproduced from Ref.[29] under CC BY 4.0.

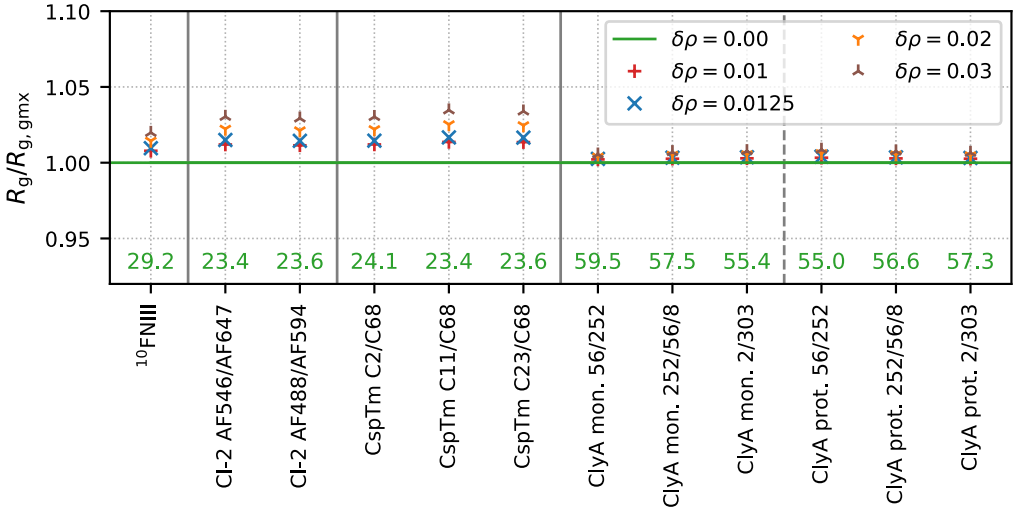


Figure C.8. Unfolded systems without dyes: R_g values for different solvation shell contrasts $\delta\rho$ (in e Å^{-3}) with respect to $R_{g,\delta\rho=0}$, given at the bottom in Å. We studied ^{10}FNIII, CI-2 with two different AF dye pairs, CspTm with AF dyes at three different labeling positions, and ClyA monomer, protomer, and dodecamer with two and three dyes at different labeling sites. Reproduced from Ref.[29] under CC BY 4.0.

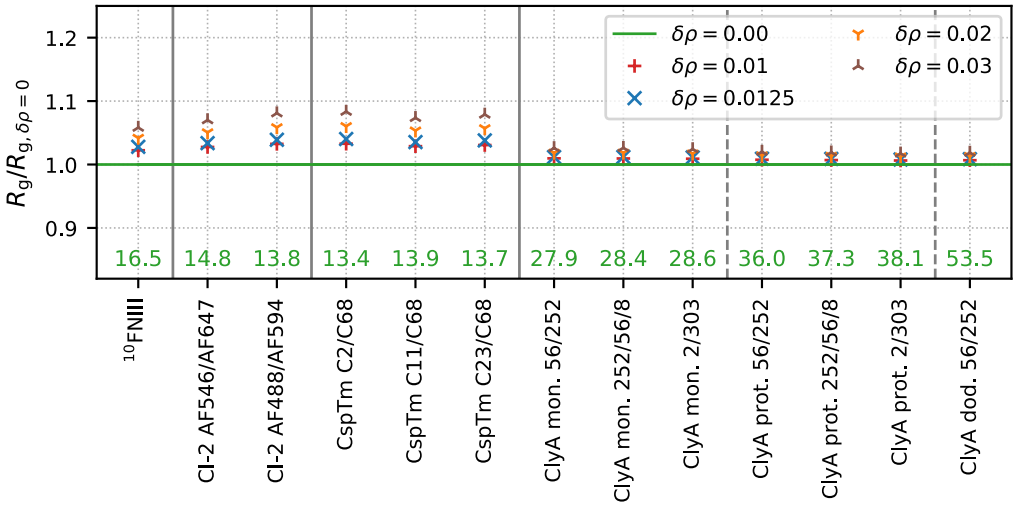


Figure C.9. Folded systems with dyes: R_g values for different solvation shell contrasts $\delta\rho$ (in e Å^{-3}) with respect to $R_{g,\delta\rho=0}$, given at the bottom in Å. We studied ^{10}FNIII, CI-2 with two different AF dye pairs, CspTm with AF dyes at three different labeling positions, and ClyA monomer, protomer, and dodecamer with two and three dyes at different labeling sites. Reproduced from Ref.[29] under CC BY 4.0.

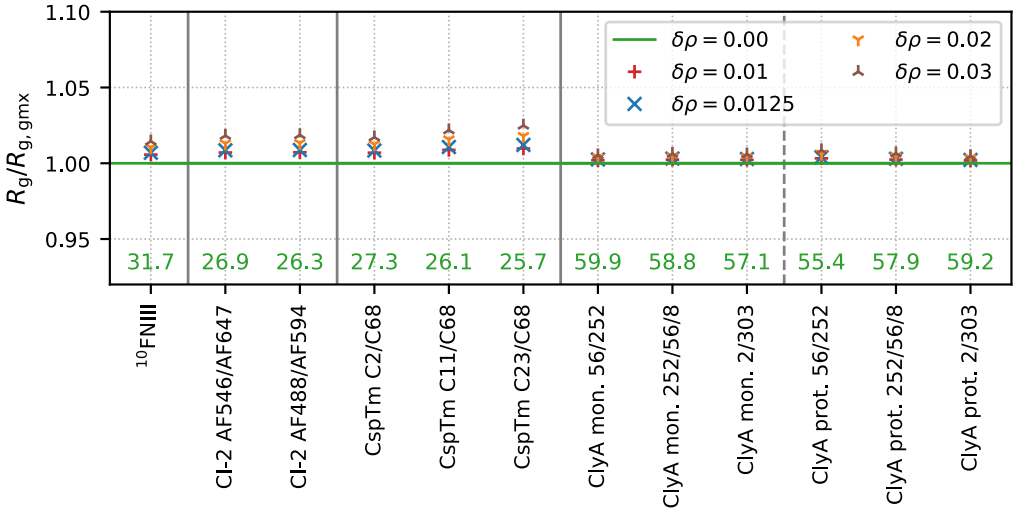


Figure C.10. Unfolded systems with dyes: R_g values for different solvation shell contrasts $\delta\rho$ (in e Å^{-3}) with respect to $R_{g,\delta\rho=0}$, given at the bottom in Å. We studied ^{10}FNIII, CI-2 with two different AF dye pairs, CspTm with AF dyes at three different labeling positions, and ClyA monomer, protomer, and dodecamer with two and three dyes at different labeling sites. Reproduced from Ref.[29] under CC BY 4.0.

D

Appendix to "PROJECT: Small-Angle X-Ray Scattering-Guided Structure-Based Protein Simulations"

D.1 SAXS-Restrained Ensemble Simulations with Commitment to the Principle of Maximum Entropy

This section is based on the Journal of Chemical Theory and Computation article "SAXS-Restrained Ensemble Simulations of Intrinsically Disordered Proteins with Commitment to the Principle of Maximum Entropy" (2019) by Markus Hermann and Jochen Hub[16]. Their multireplica refinement method is based on the principle of maximum entropy. Applying only a minimal bias towards agreement with the data, ensemble-averaged SAXS intensities are integrated into MD simulations of proteins. In the context of my work, it shall serve as another example of how biomolecular simulations and experimental data can complement each other with a special focus on the ensemble character of structurally heterogeneous disordered systems.

Intrinsically disordered proteins (IDPs) play key roles in pathologies such as amyloidoses, neurodegenerative diseases, and cancer, which is why there is a great interest in deriving accurate solution ensembles. As explained in Chapter 3, experimental data provide only a highly reduced view of such system's heterogeneous ensembles. While X-ray crystallography and cryo EM generate a single defined structure, solution scattering yields an average over the entire ensemble. Computational MD simulations are complementary to such experiments but suffer from sampling problems and force-field inaccuracies, in particular for disordered systems, suggesting to combine them with experimental data. Hermann and Hub developed a data-assisted method for refining protein ensembles towards agreement with ensemble-averaged SAXS data using a parallel-replica ensemble restraint [16]. Following the principle of maximum entropy, they couple a set of parallel replicas to the data, whereby only a minimal bias is applied [16]. Simulations are restrained at runtime to conformations $\mathbf{R}$ reproducing the ensemble-averaged experimental data I_{exp} by introducing an experiment-derived energy $E_{\mathrm{exp}}\left(\mathbf{R}; I_{\mathrm{exp}}\right)$. The simplest approach is to restrain a single simulation with a harmonic restraint to the experimental SAXS intensities $I_{\mathrm{exp}}\left(q_i\right)$ [16]:

$$E_{\mathrm{exp}}^{(1)} = \frac{k_{\mathrm{r}} k_{\mathrm{B}} T}{n_q} \sum_{i=1}^{n_q} \frac{\left[I_{\mathrm{calc}}\left(q_i, \mathbf{R}\right) - I_{\mathrm{exp}}\left(q_i\right)\right]^2}{\sigma_i^2} \tag{D.1}$$

k_{r} is a unitless force constant, i.e., the bias weight, k_{B} the Boltzmann constant, T the temperature, and $I_{\mathrm{calc}}\left(q_i, \mathbf{R}\right)$ the back-calculated scattering intensity. σ_i is the scattering intensity's overall uncertainty at momentum transfer q_i, which includes experimental uncertainties and uncertainties in the forward model of $I_{\mathrm{calc}}\left(q_i, \mathbf{R}\right)$. The hybrid energy is set up as [16]

$$E_{\mathrm{hybrid}}\left(\mathbf{R}; I_{\mathrm{exp}}\right) = E_{\mathrm{MD}}\left(\mathbf{R}\right) + E_{\mathrm{exp}}^{(1)}\left(\mathbf{R}; I_{\mathrm{exp}}\right) . \tag{D.2}$$

Eqs. D.1 and D.2 are only appropriate if the ensemble can be approximated by a single representative structure. For the heterogeneous ensembles of disordered systems, this is not the case. It thus is not meaningful to compare a single structure's scattering intensity, $I_{\mathrm{calc}}\left(q_i, \mathbf{R}\right)$, with the ensemble-averaged experimental data, $I_{\mathrm{exp}}\left(q_i\right)$. Instead, $I_{\mathrm{exp}}\left(q_i\right)$ should be subtracted against the scattering intensity $\int p\left(\mathbf{R}\right) I_{\mathrm{calc}}\left(q_i; \mathbf{R}\right) d\mathbf{R}$ averaged over the simulated ensemble distribution $p\left(\mathbf{R}\right)$. A statistically founded procedure for updating simulated ensembles with experimental data is Jaynes' maximum entropy principle [16]: The unbiased ensemble distribution $p_0\left(\mathbf{R}\right)$ should be modified as little as possible into a biased distribution $p_1\left(\mathbf{R}\right)$ that explains the data, i.e., the updated ensemble and the original unbiased ensemble should be as similar as possible. This can be quantified by the Kullback-Leibler divergence:

$$D_{\mathrm{KL}}\left(p_1 | p_0\right) = \int p_1\left(\mathbf{R}\right) \ln \frac{p_1\left(\mathbf{R}\right)}{p_0\left(\mathbf{R}\right)} d\mathbf{R} \tag{D.3}$$

Biasing a single simulation with a harmonic restraint towards the data obviously violates the maximum entropy principle, suggesting the use of alternative coupling schemes [16]. According to Pitera and Chodera, using a linear restraint instead of the harmonic restraint will generate a minimally biased ensemble. Alternatively, N parallel replica simulations can be coupled with a harmonic restraint to the data. Accordingly, the back-calculated scattering signal is first averaged over all parallel replicas [16],

$$\bar{I}_{\mathrm{calc}}\left(q_i, \mathbf{R}_1, ..., \mathbf{R}_N\right) = N^{-1} \sum_{\alpha=1}^{N} I_{\mathrm{calc}}\left(q_i, \mathbf{R}_\alpha\right) , \tag{D.4}$$

where α is the replica index. The restraint is designed to guide the simulation into agreement with the experimental SAXS data while applying a minimal bias [16]:

$$E_{\mathrm{exp}}\left(\mathbf{R}_1, ..., \mathbf{R}_N; I_{\mathrm{exp}}\right) = \frac{k_{\mathrm{r}} N k_{\mathrm{B}} T}{n_q} \sum_{i=1}^{n_q} \frac{\left[\bar{I}_{\mathrm{calc}}\left(q_i, \mathbf{R}_1, ..., \mathbf{R}_N\right) - I_{\mathrm{exp}}\left(q_i\right)\right]^2}{\sigma_i^2} \tag{D.5}$$

k_r is an empirical parameters expressing the degree of confidence in the experimental data versus the unbiased force field[16]. Practically, N and k_r can be determined in a reasonable way by analyzing D_{KL} and the residuals χ^2 between experimental and calculated data versus increasing N and k_r and choosing the smallest values for giving satisfactory χ^2 and D_{KL}.

Hermann and Hub demonstrated their multireplicate method for generating ensembles with commitment to the maximum entropy principle by refining ensembles of the disordered RS peptide against experimental SAXS data. SAXS data were calculated from explicit solvent as described before[12,152]. They analyzed the influence of the number of replicas, the scaling of the SAXS force constant with the number of replicas, and the force field. They find that the refined ensemble greatly improves when switching from a single to four replicas and suggest using a force constant of unity[16]. As the force field seemed to have only a minor influence on the derived ensemble, the data are believed to reduce the force-field bias, rendering such hybrid methods useful for obtaining physically precise simulations of disordered systems[16].

D.2 Simulation Setups

All simulations were carried out on a standard workstation with an Intel Core i7-6700 CPU comprising eight cores at a frequency of 3.40 GHz. I used a version of `GROMACS 5` modified by the scattering-guided MD extension[18,61] and molecular dynamics parameters listed below. The simulations differed only in their couplings to the scattering data, k_χ, and their reduced `GROMACS` temperatures, T. As the information on crucial structural features, i.e., the molecular shape and global conformational changes, is contained in the small-angle regime, I included q values up to a maximum of $0.5\,\text{Å}^{-1}$. I used theoretical scattering data calculated from pure initial and final states for method validation. In a SAXS experiment, the measured intensity pattern might reflect a linear combination of scattering intensities from a mixture of conformations in the sample. Starting from the pure initial state, I assumed conformational transitions to take place entirely in the simulations. This means, in a corresponding experiment, all protein molecules would undergo the structural transition of interest from initial to final state. Consequently, I set α to 1 in all simulations.

Setup of `XSBM` Simulations

As a starting point, I constructed all-atom SBMs from the considered system's initial structure with `eSBMTools`[99] to obtain suitable coordinate and topology files. In `XSBM`, the Debye scattering terms are encoded as a special type of bonded interaction in the topology file[18]. I constructed the scattering topology and the related extended coordinate file with `gmx genrestr`. This command creates half a matrix of virtual-site type-3 pairs, i.e., Debye terms, for the input coordinate file. I used amino-acid scatterers centered on virtual interaction sites at the respective residue's center of mass. I added the resulting topology include file to the system's topology directly after the atoms section and appended the corresponding atom type "MW" manually to the atom types table. The `XSBM` run parameters are listed below. I set T and k_χ as described in Sec. 5.2. Finally, I preprocessed the SBMs with `gmx grompp` and run the simulations with `gmx mdrun`.

```
; Sample mdp file for XSBM simulations in GROMACS 5

; Run Control
integrator = sd      ;leap-frog integrator
dt         = 0.0005  ;time step / ps
nsteps     = 4000000 ;number of steps

; XSBM parameters
waxs-type        = Debye ;Debye scattering intensities
waxs-fc          = ?     ;bias weight
waxs-nstout      = 1     ;steps between dumping simulated intensities
waxs-nstcalc     = 10    ;steps between calculation/application of XS forces
debye-alpha-mode = 0     ;constant alpha at
debye-alpha-min  = 1.0   ;(debye-alpha-min + debye-alpha-max)/2
debye-alpha-max  = 1.0

; Output Control
nstxout            = 0       ;Only write final coordinates to trajectory.
nstvout            = 0       ;Only write final velocities to trajectory.
nstxout-compressed = 100     ;steps between dumping compressed coordinates
compressed-x-grps  = Protein ;groups in compressed trajectory
nstcalcenergy      = 105     ;steps between calculation of energies
nstenergy          = 105     ;steps between dumping energies
nstlog             = 100     ;steps between writing energies to log
nstcomm            = 105     ;frequency for center of mass motion removal

; Neighbor Searching
cutoff-scheme = group ;Generate pair list for groups of atoms.
ns-type       = grid  ;Make grid in box, only check atoms in neighboring cells.
nstlist       = 15    ;frequency to update neighbor list
pbc           = xyz   ;periodic BC in all directions

; Electrostatics
coulombtype = Cut-off ;twin range cut-off
rcoulomb    = 1.5     ;Coulomb cut-off / nm

; VdW
vdwtype = Cut-off ;twin range cut-off
rvdw    = 1.5     ;VdW cut-off / nm

; Temperature Coupling
tcoupl  = V-rescale ;stochastic temperature coupling using velocity rescaling
tc-grps = Protein   ;groups to couple separately to temperature bath
tau-t   = 0.5       ;time constant for coupling
ref-t   = ?         ;reference temperature for coupling

; Pressure Coupling
pcoupl = no ;no pressure coupling
```

```
; Velocity Generation
gen-vel  = yes ;velocity generation according to Maxwell distribution
gen-temp = ?   ;temperature for Maxwell distribution
gen-seed = 9   ;initialize random generator for random velocities

; Bonds
continuation         = no        ;Constrain start configuration.
constraint-algorithm = lincs     ;LINear Constraint Solver
constraints          = all-bonds ;Constrain all bonds.
lincs-iter           = 1         ;iterations for LINCS
lincs-order          = 4         ;matrices in expansion for LINCS inversion
```

Setup of Scattering-Guided Explicit-Solvent MD Simulations

The setup of explicit-solvent MD simulations followed the common steps of adding hydrogen atoms, choosing potential and water model, neutralizing electric charge, minimizing energy, and equilibrating temperature and pressure. I used the `CHARMM27` force field[209], `TIP3P` water model[210], Verlet cut-off scheme, and a constant temperature of 300 K. Electrostatics were treated with the Particle Mesh Ewald method. I applied a Parrinello-Rahman pressure coupling and a V-rescale temperature coupling. To obtain the coordinate and topology files, I preprocessed and protonated the initial models with `gmx pdb2gmx`. I constructed a periodic cubic box exceeding twice the longest inter-protein distance with `gmx editconf`. The structure was initially energy-minimized using the `GROMACS` preprocessor `gmx grompp` and simulation command `gmx mdrun`. After solvation and electric neutralization, I energy-minimized the structure again. Subsequently, I equilibrated the systems in the canonical and isothermal-isobaric ensemble until temperature and pressure converged, where I position-restrained all heavy atoms to their initial positions. I constructed a half-matrix of Debye terms with `gmx genrestr` for the NPT-equilibrated structure, using amino-acid scattering factors. This created the scattering topology, which I manually included into the system's topology. After preprocessing with `gmx grompp`, I performed the scattering-guided MD simulation with `gmx mdrun`.

```
; Sample mdp file for scattering-guided
; explicit-solvent MD simulations in GROMACS5

; Run parameters
integrator = md      ;leap-frog integrator
dt         = 0.002   ;time step / ps
nsteps     = 5000000 ;number of steps (10 ns)

; WAXS parameters
waxs-type        = Debye ;Debye scattering intensities
waxs-fc          = ?     ;bias weight
waxs-nstout      = 1     ;steps between dumping simulated intensities
waxs-nstcalc     = 10    ;steps between calculation/application of XS forces
debye-alpha-mode = 0     ;constant alpha at
debye-alpha-min  = 1.0   ;(debye-alpha-min + debye-alpha-max)/2
debye-alpha-max  = 1.0
```

```
; Output control
nstxout            = 0       ;Only write final coordinates to trajectory.
nstvout            = 0       ;Only write final velocities to trajectory.
nstxout-compressed = 100     ;steps between dumping compressed coordinates
compressed-x-grps  = Protein ;groups in compressed trajectory
nstenergy          = 100     ;steps between dumping energies
nstlog             = 100     ;steps between writing energies to log

; Bond parameters
continuation         = yes
constraint-algorithm = lincs
constraints          = all-bonds
lincs-iter           = 1
lincs-order          = 4

; Neighborsearching
cutoff-scheme           = verlet ;Generate pair list with buffering.
verlet-buffer-tolerance = 0.005
ns-type                 = grid
nstlist                 = 15     ;frequency to update neighbor list
pbc                     = xyz    ;Periodic BC in all directions
periodic-molecules      = no     ;Molecules are finite.

; Electrostatics
coulombtype    = PME ;particle-mesh Ewald
pme-order      = 4
fourierspacing = 0.12
rcoulomb       = 1.0 ;Coulomb cutoff / nm

; Van der Waals
rvdw     = 1.0      ;VdW cutoff / nm
DispCorr = EnerPres ;account for cut-off vdW scheme

; Temperature coupling
tcoupl  = V-rescale ;stochastic temperature coupling using velocity rescaling
tc-grps = Protein Water_and_ions
tau-t   = 0.5 0.5
ref-t   = 300  300

; Pressure coupling
pcoupl           = Parrinello-Rahman ;extended-ensemble pressure coupling
pcoupltype       = isotropic
tau-p            = 2.0    ;time constant for coupling / ps
ref-p            = 1.0    ;reference pressure / bar
compressibility  = 4.5e-5 ;compressibility / 1/bar
refcoord-scaling = com
```

```
; Velocity generation
gen-vel = no ;Take initial velocities from NPT run.
```

For each conformational transition, I present results for parameter combinations (T, k_χ) of temperature and bias weight determined via grid search, where I considered eight bias weights between 10^{-11} kJ/mol and $5 \cdot 10^{-8}$ kJ/mol at two temperatures $T = 300$ K and 330 K.

D.3 Structural Conformity Analysis

To study the structural conformity of the ensembles generated by scattering-guided structure-based and explicit-solvent MD simulations, I calculated the radius of gyration, R_g, and the asphericity, Δ, as a function of simulated time. A molecule's gyration tensor is given by[211]

$$T_{\alpha\beta} = \frac{1}{2N^2} \sum_{i,j=1}^{N} (r_{i\alpha} - r_{j\alpha}) \cdot (r_{i\beta} - r_{j\beta}) \,, \tag{D.6}$$

where N is the number of atoms at Cartesian positions $\mathbf{r}_i$,$i = 1, ..., N$, $r_{i\alpha}$ is the α^{th} component of $\mathbf{r}_i$, and α, β are x, y, and z. The asphericity can be calculated as

$$\Delta = \frac{3 \sum_{i=1}^{3} \left(\lambda_i - \bar{\lambda}\right)^2}{2 \left(\operatorname{tr} T\right)^2} \,, \tag{D.7}$$

where λ_i are the eigenvalues of T with $\operatorname{tr} T = \sum_i \lambda_i$ and $\bar{\lambda} = \operatorname{tr} T/3$. I computed the radii of gyration and the eigenvalues with `gmx gyrate` and `gmx polystat` in `GROMACS`, respectively. As can be seen in Supplementary Figs. D.1 to D.8, both shape parameters suggest the structural ensembles generated by SBM and explicit-solvent MD to be sufficiently similar. As expected, I observed comparatively large differences for VHP_{54}^{74}. This can be explained by its globally changing tertiary structure, which makes the conformational transitions difficult to sample in a geometry-derived SBM with its intrinsic bias towards the native (initial) state.

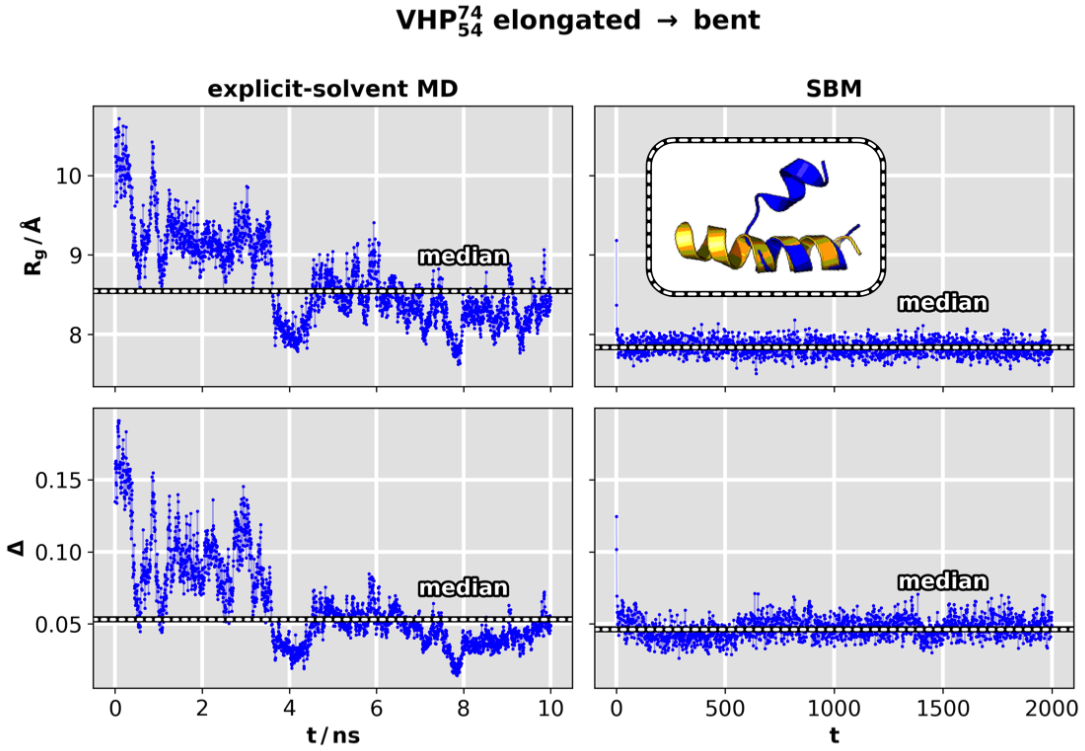


Figure D.1. VHP$_{54}^{74}$ e → b transition: Structural conformity of the ensembles generated by SBM and explicit-solvent MD. Radius of gyration, R_g, and asphericity, Δ, versus simulated time.

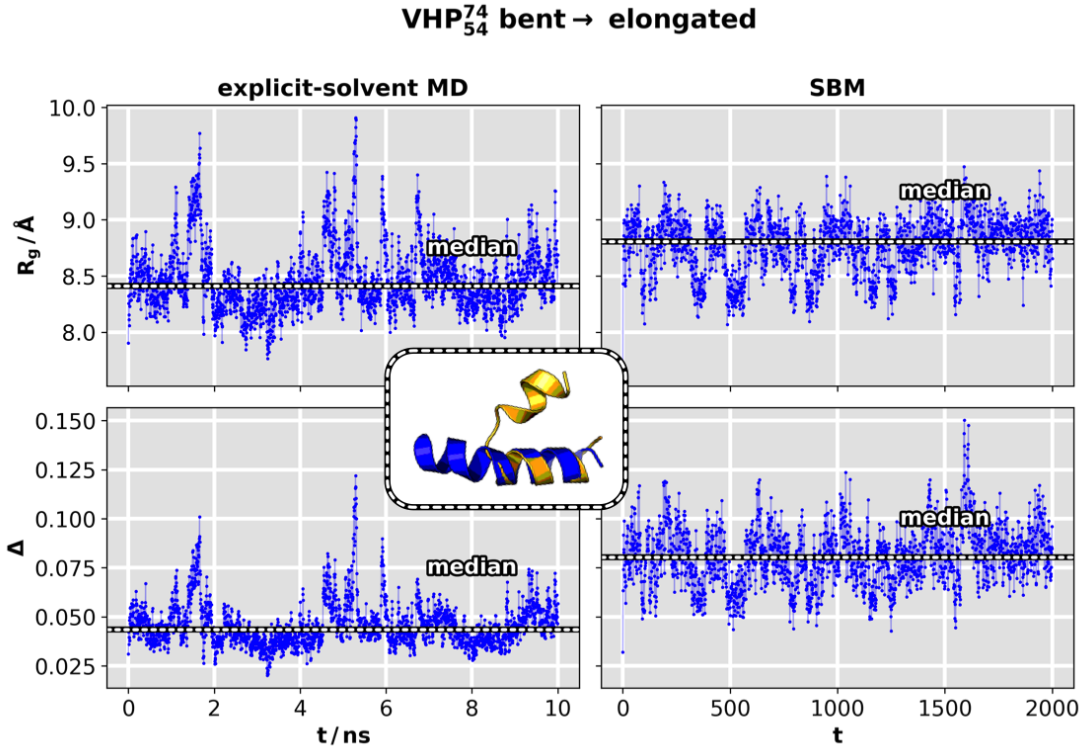


Figure D.2. VHP$_{54}^{74}$ b → e transition: Structural conformity of the ensembles generated by SBM and explicit-solvent MD. Radius of gyration, R_g, and asphericity, Δ, versus simulated time.

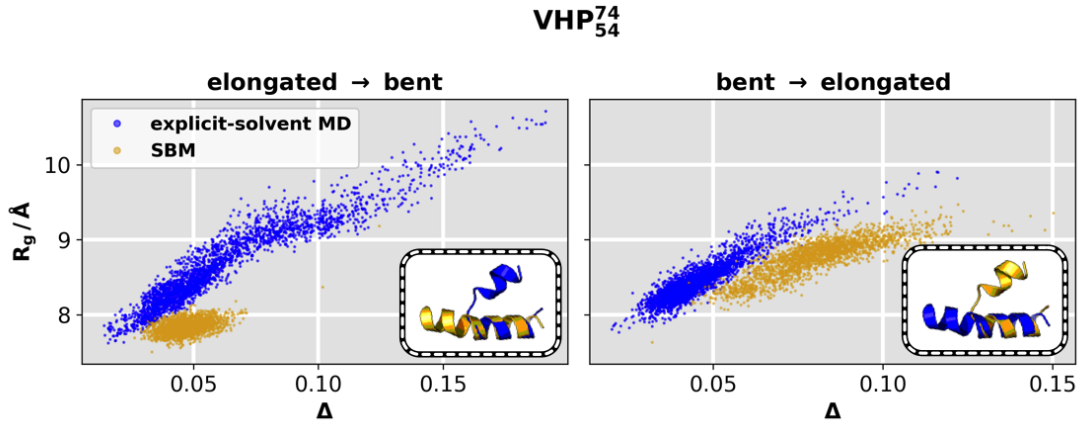


Figure D.3. VHP$_{54}^{74}$: Structural conformity of the ensembles generated by SBM and explicit-solvent MD. Radius of gyration, R_g, versus asphericity, Δ.

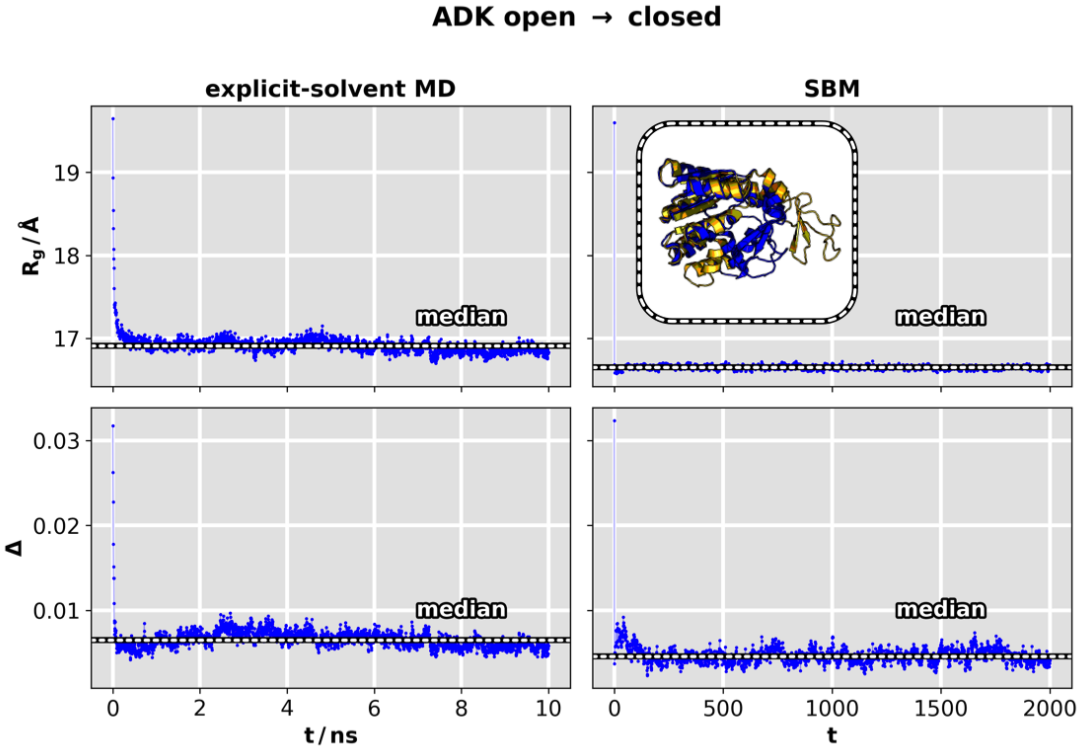


Figure D.4. ADK o → c transition: Structural conformity of the ensembles generated by SBM and explicit-solvent MD. Radius of gyration, R_g, and asphericity, Δ, versus simulated time.

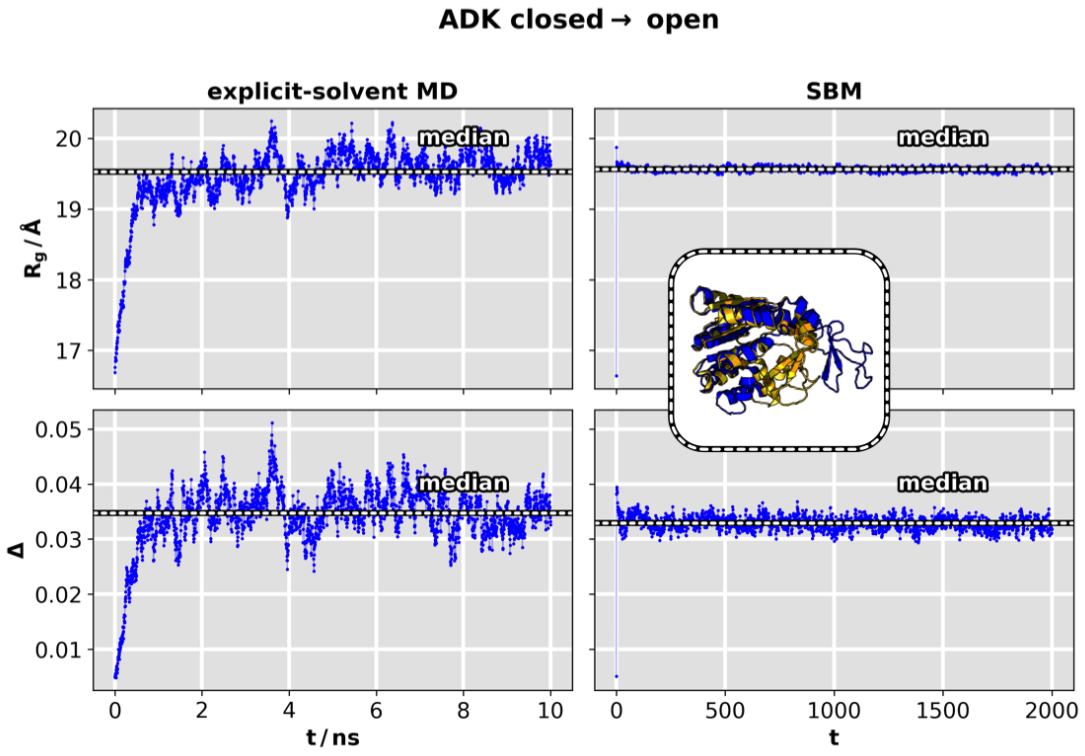


Figure D.5. ADK c → o transition: Structural conformity of the ensembles generated by SBM and explicit-solvent MD. Radius of gyration, R_g, and asphericity, Δ, versus simulated time.

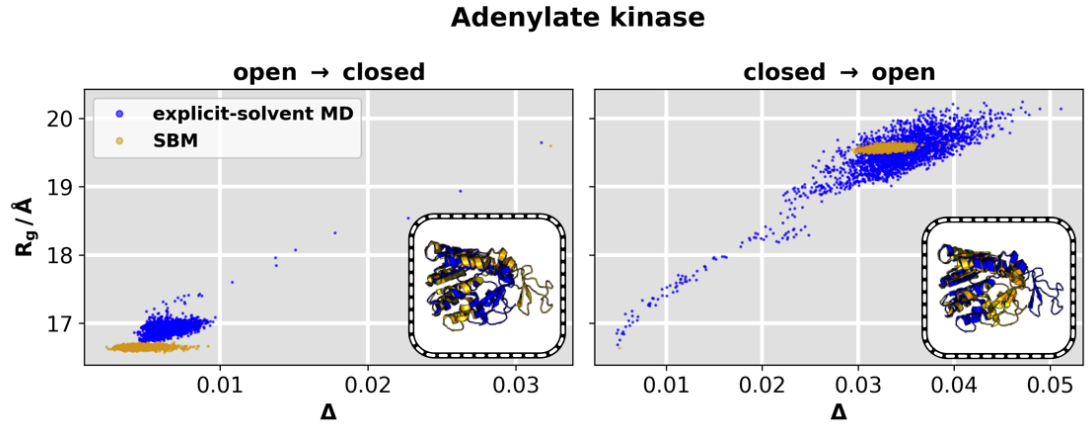


Figure D.6. ADK: Structural conformity of the ensembles generated by SBM and explicit-solvent MD. Radius of gyration, R_g, versus asphericity, Δ.

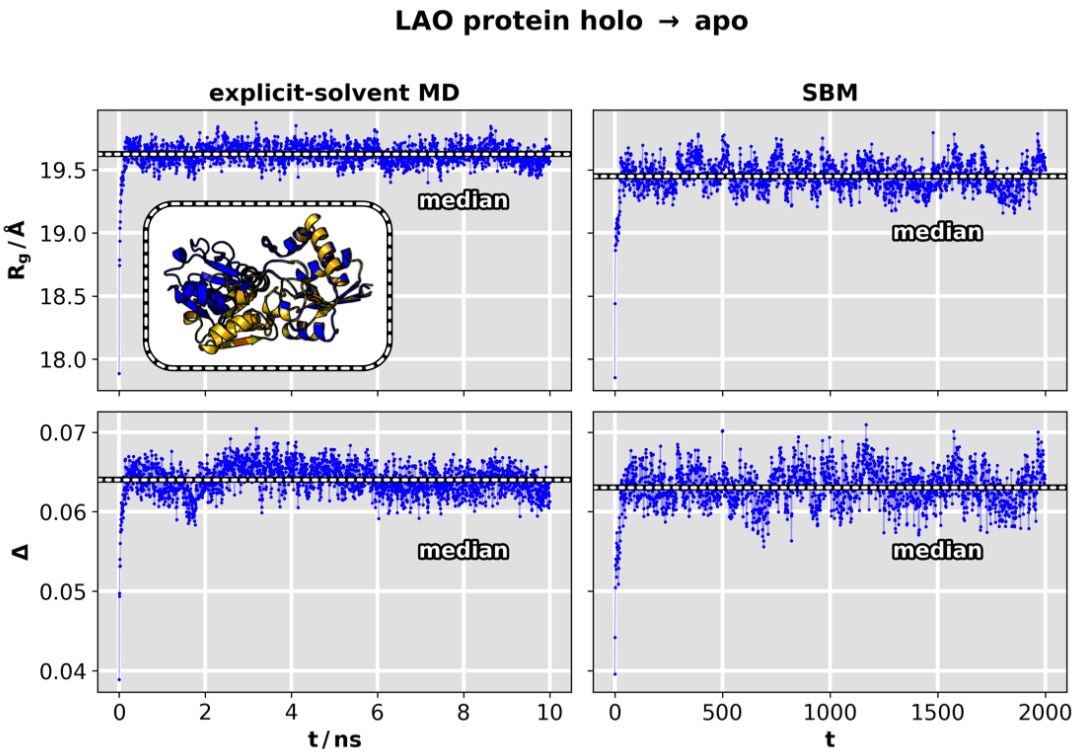


Figure D.7. LAO protein h → a transition: Structural conformity of the ensembles generated by SBM and explicit-solvent MD. Radius of gyration, R_g, and asphericity, Δ, versus simulated time.

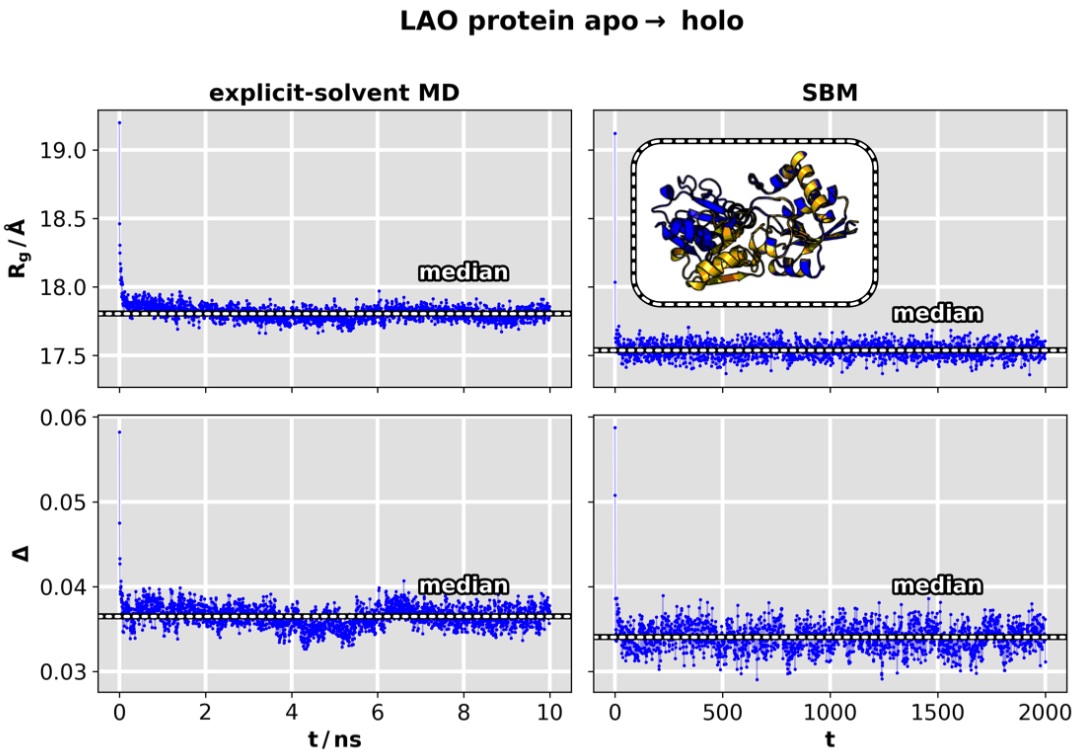


Figure D.8. LAO protein a → h transition: Structural conformity of the ensembles generated by SBM and explicit-solvent MD. Radius of gyration, R_g, and asphericity, Δ, versus simulated time.

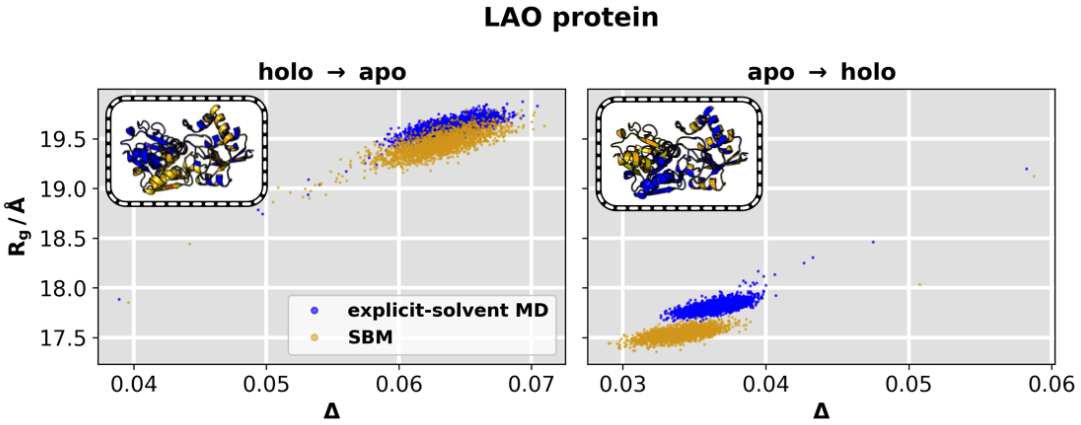


Figure D.9. LAO protein: Structural conformity of the ensembles generated by SBM and explicit-solvent MD. Radius of gyration, R_g, versus asphericity, Δ.

D.4 Villin Headpiece

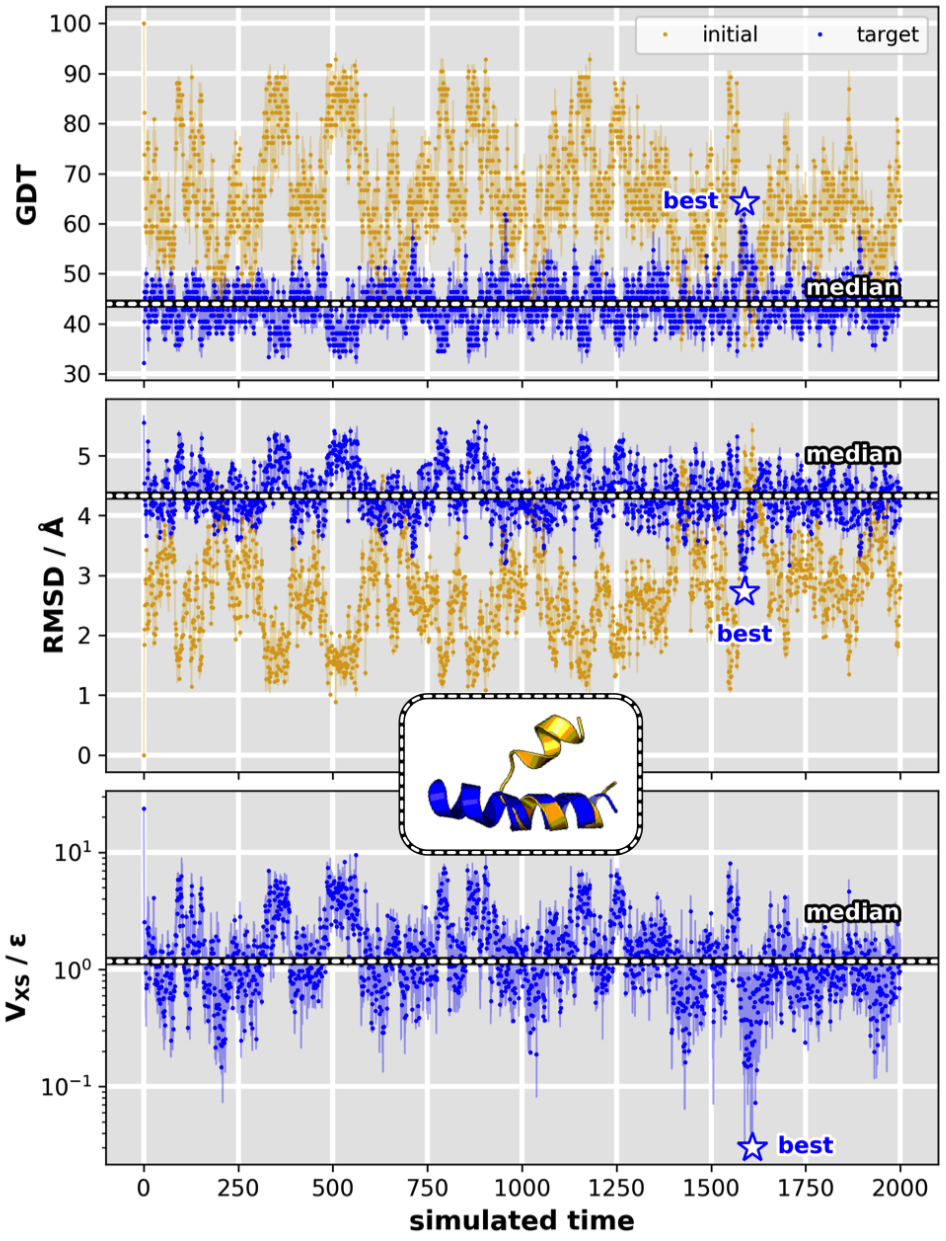


Figure D.10. `XSBM` **simulation results for VHP$^{74}_{54}$ b → e transition.** Results are shown for parameters $(T,\ k_\chi) = (50,\ 7 \cdot 10^{-8}\,\varepsilon)$. Initial and target GDT (top), initial and target RMSD (middle), and bias energy (bottom) versus simulated time.

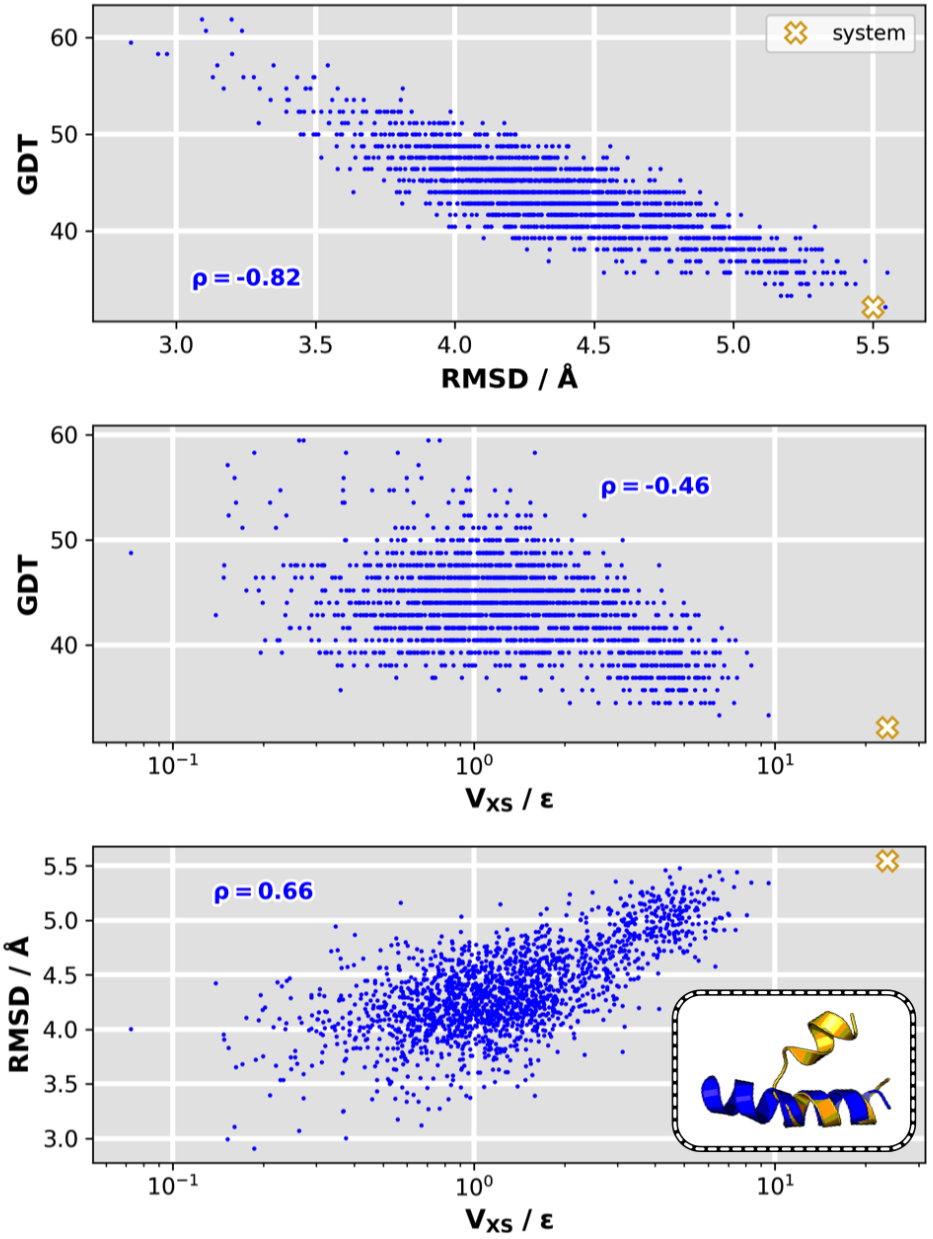


Figure D.11. VHP^{74}_{54} b → e transition in `XSBM`: Mutual correlations among GDT, RMSD, and bias potential. Target GDT versus target RMSD (top), target GDT versus V_{XS} (middle), target RMSD versus V_{XS} (bottom). ρ is the Pearson correlation of each two plotted quantities.

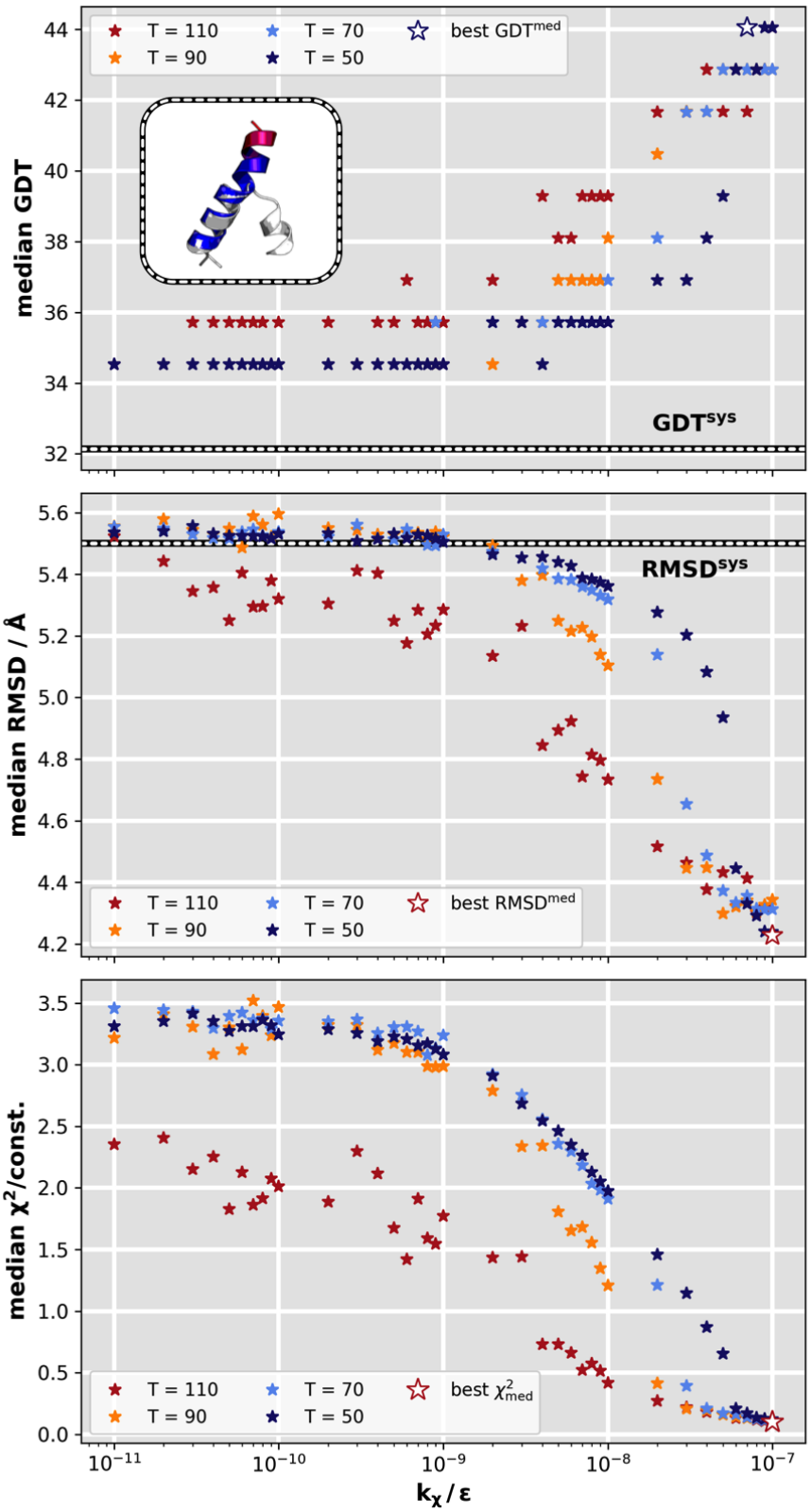


Figure D.12. VHP$^{74}_{54}$: Variational grid search for the b → e transition. Median GDT, median target RMSD, and median χ^2 deviation versus bias weight k_χ at different temperatures T. The variational series comprised 148 simulations in total. Best (maximum) GDT, best (minimum) RMSD, and best (minimum) V_{XS} are marked by a white star, each outlined in the color of the related temperature.

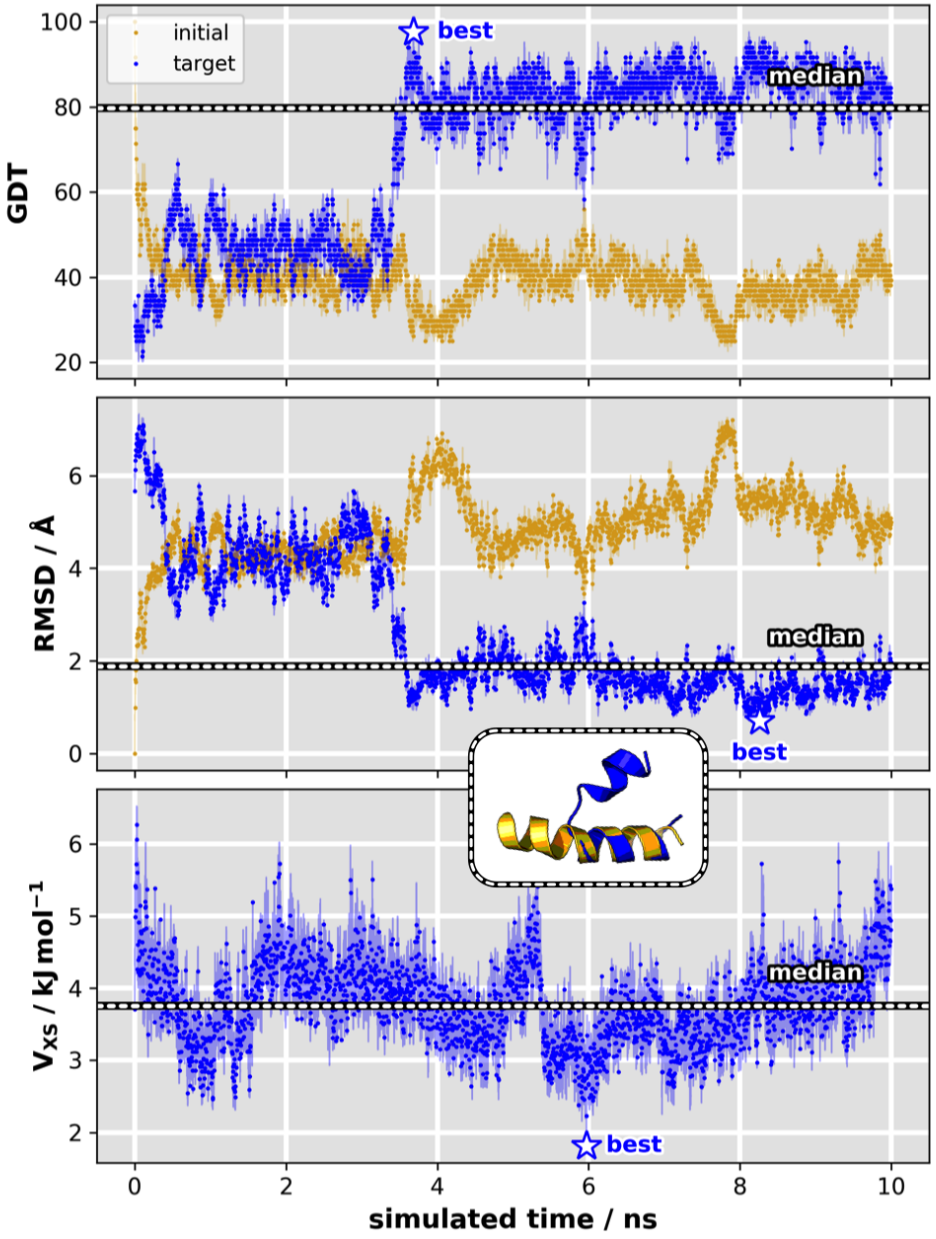


Figure D.13. Explicit-solvent MD simulation results for VHP_{54}^{74} e $\rightarrow$ b transition. Results are shown for parameters $(T,\, k_\chi) = (330\,\mathrm{K},\, 5 \cdot 10^{-9}\,\mathrm{kJ/mol})$. Initial and target GDT (top), initial and target RMSD (middle), and bias energy (bottom) versus simulated time.

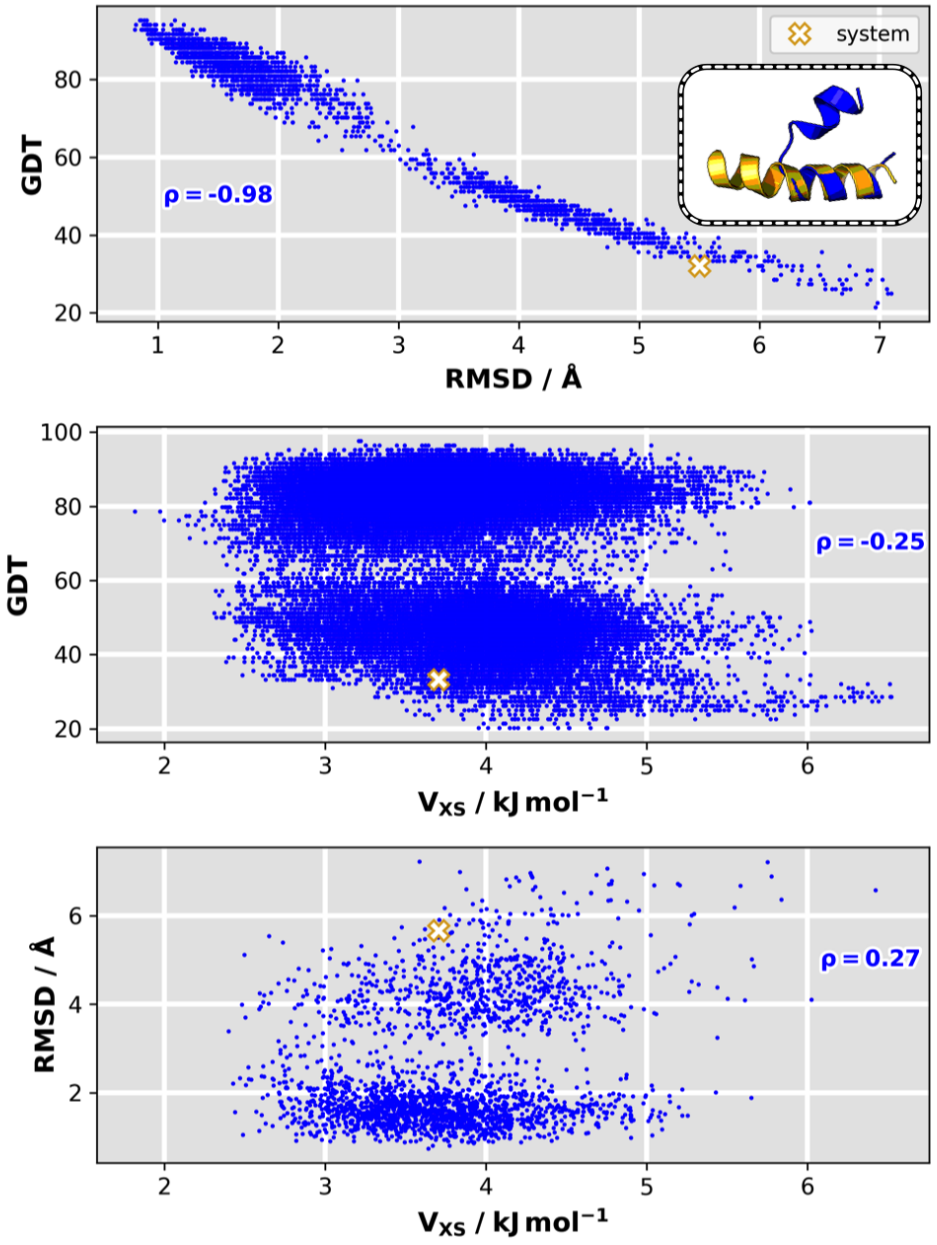


Figure D.14. VHP^{74}_{54} e → b transition in explicit-solvent MD: Mutual correlations among GDT, RMSD, and bias potential. Target GDT versus target RMSD (top), target GDT versus V_{XS} (middle), target RMSD versus V_{XS} (bottom). ρ is the Pearson correlation of each two plotted quantities.

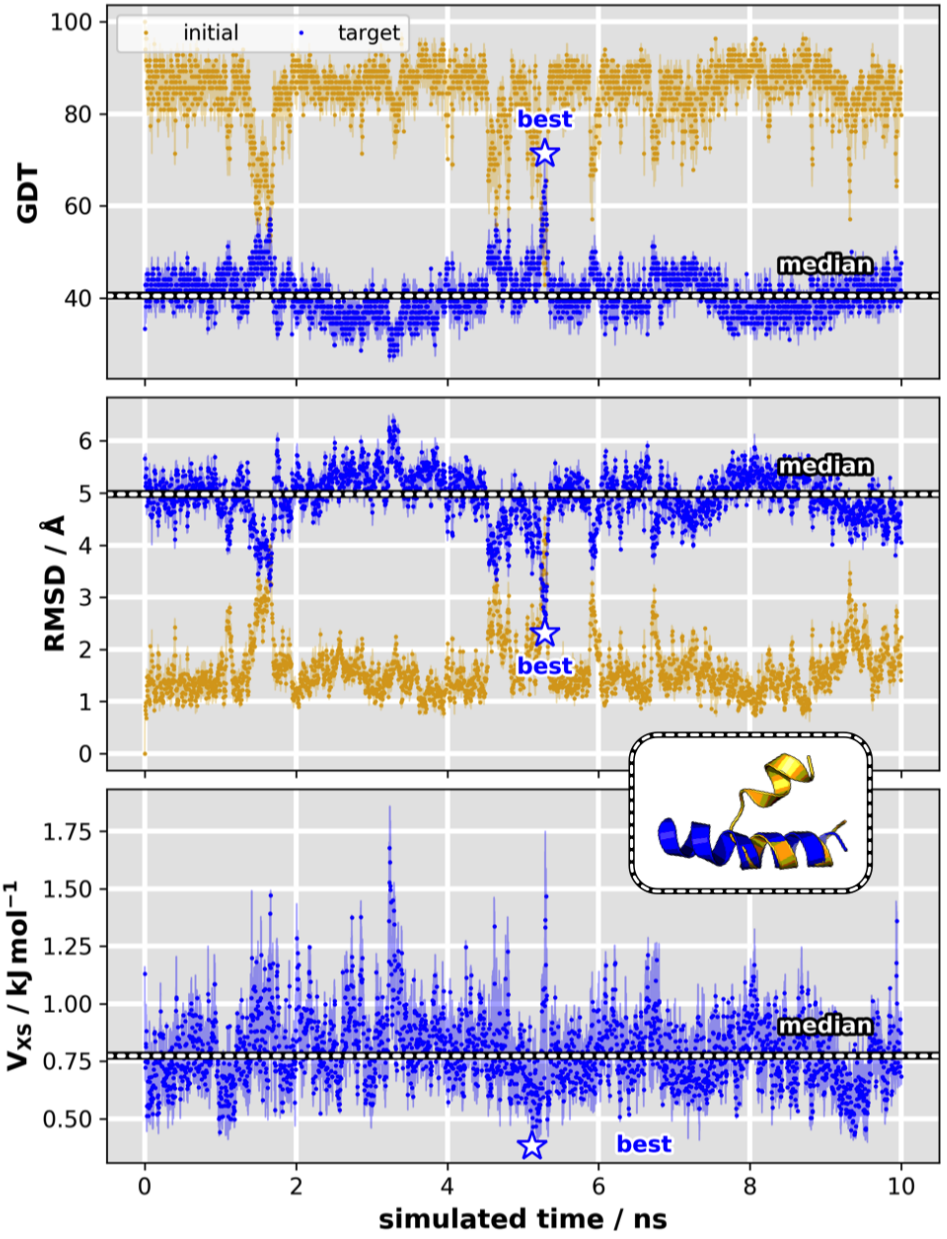


Figure D.15. Explicit-solvent MD simulation results for VHP_{54}^{74} b $\rightarrow$ e transition. Results are shown for parameters $(T,\ k_\chi) = (330\,\mathrm{K},\ 5 \cdot 10^{-9}\,\mathrm{kJ/mol})$. Initial and target GDT (top), initial and target RMSD (middle), and bias energy (bottom) versus simulated time.

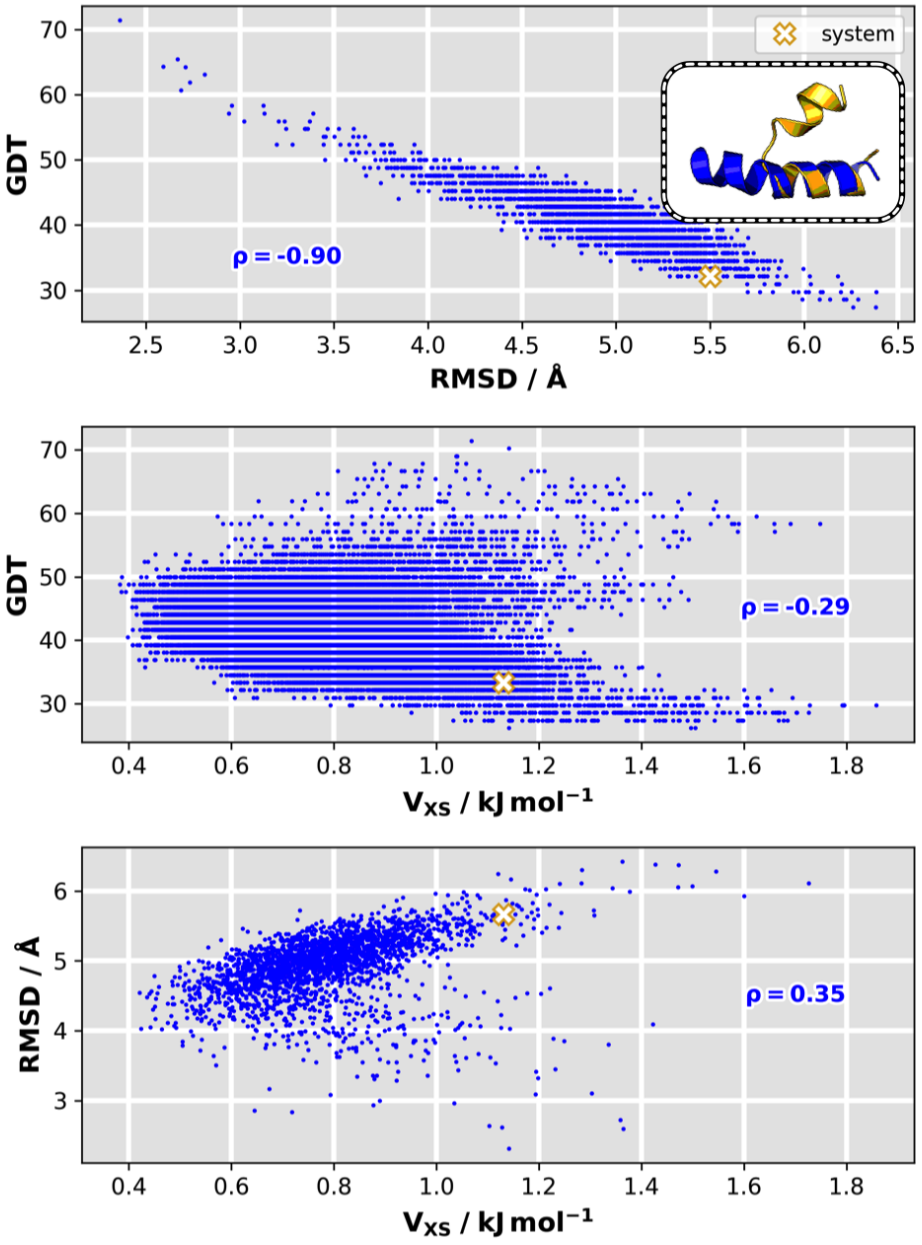


Figure D.16. VHP^{74}_{54} b → e transition in explicit-solvent MD: Mutual correlations among GDT, RMSD, and bias potential. Target GDT versus target RMSD (top), target GDT versus V_{XS} (middle), target RMSD versus V_{XS} (bottom). ρ is the Pearson correlation of each two plotted quantities.

D.5 Lysine-, Arginine-, Ornithine-Binding Protein

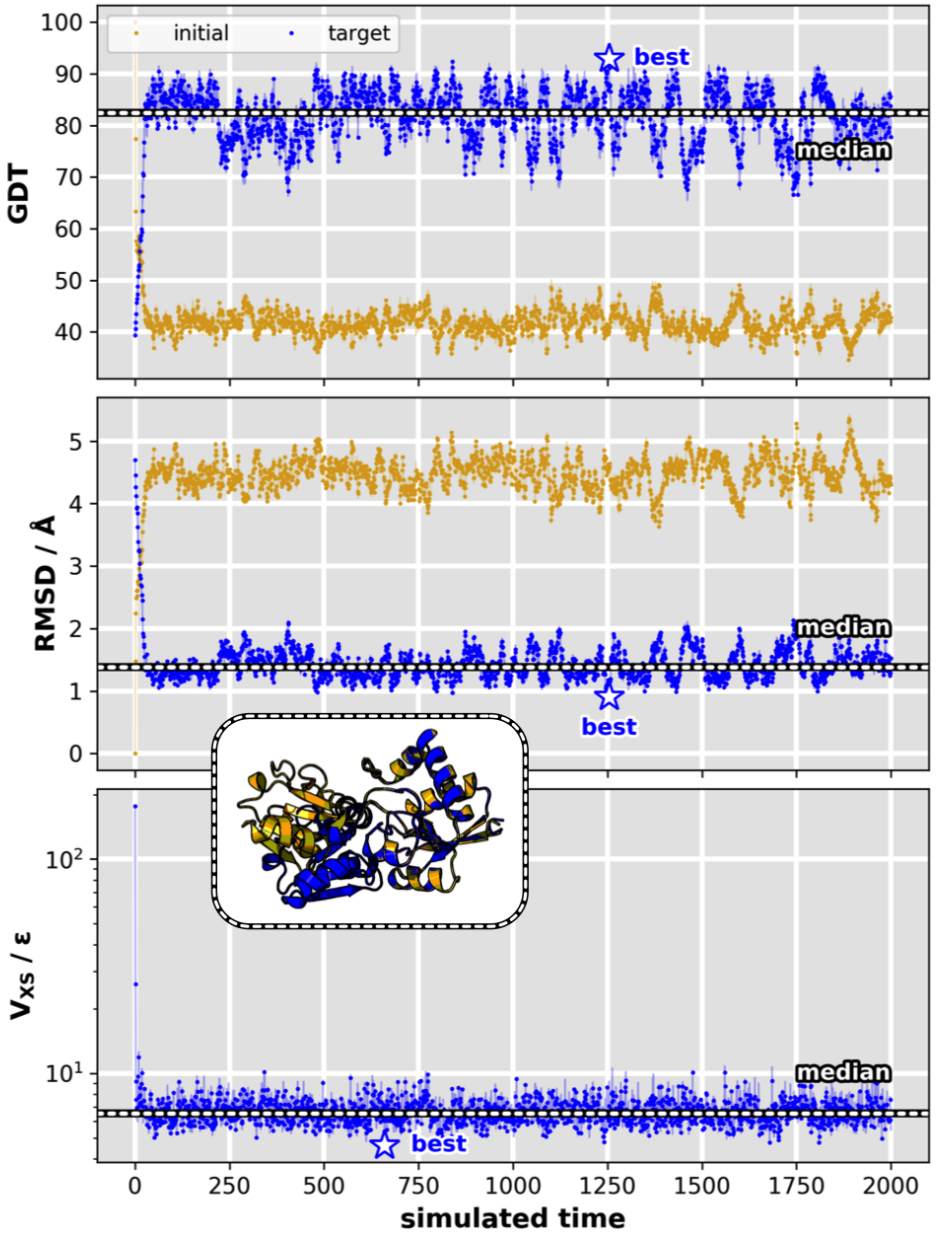


Figure D.17. `XSBM` **simulation results for LAO protein a $\rightarrow$ h transition.** Results are shown for parameters $(T,\ k_\chi) = \left(50,\ 2 \cdot 10^{-10}\,\varepsilon\right)$. Initial and target GDT (top), initial and target RMSD (middle), and bias energy (bottom) versus simulated time.

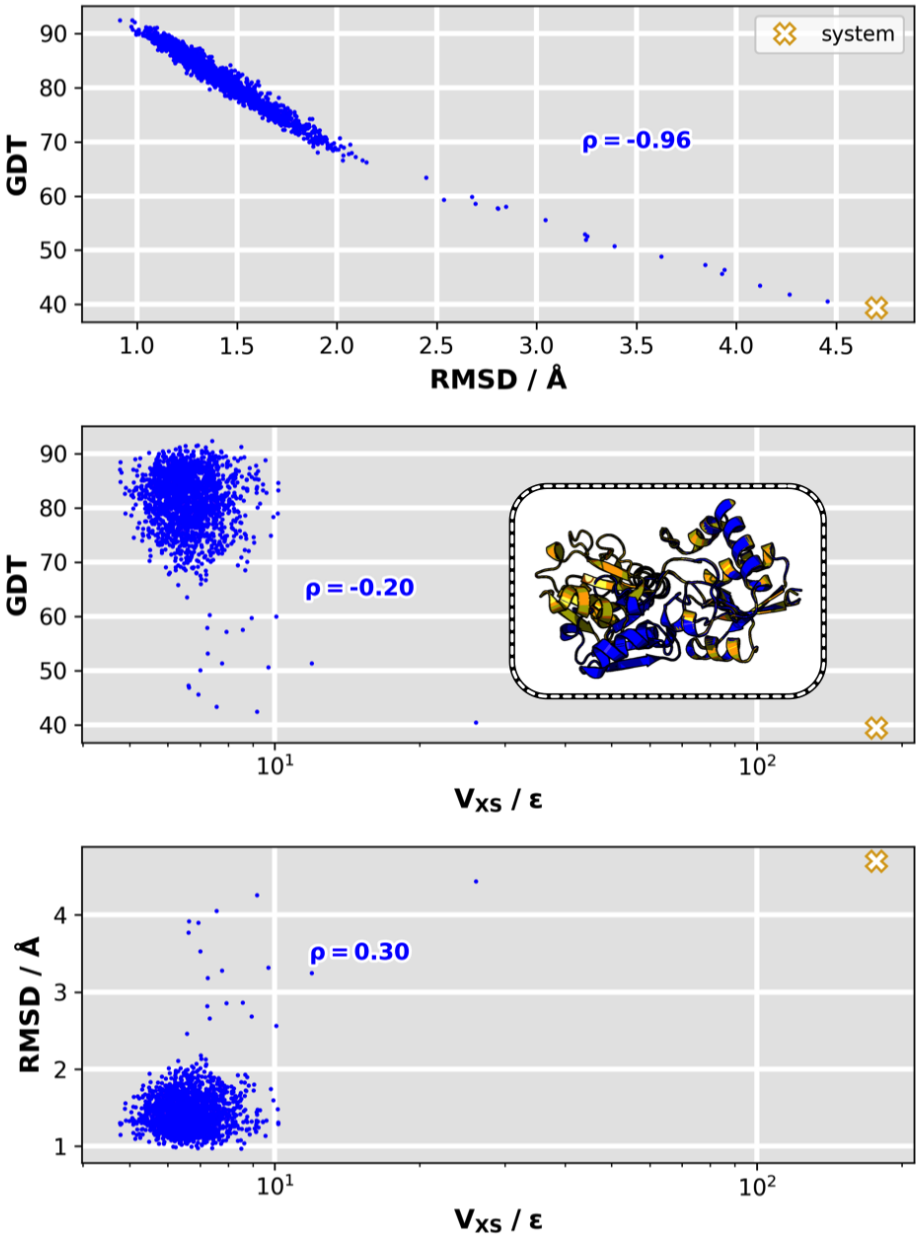


Figure D.18. LAO protein a → h transition in `XSBM`: Mutual correlations among GDT, RMSD, and bias potential. Target GDT versus target RMSD (top), target GDT versus V_{XS} (middle), target RMSD versus V_{XS} (bottom). ρ is the Pearson correlation of each two plotted quantities.

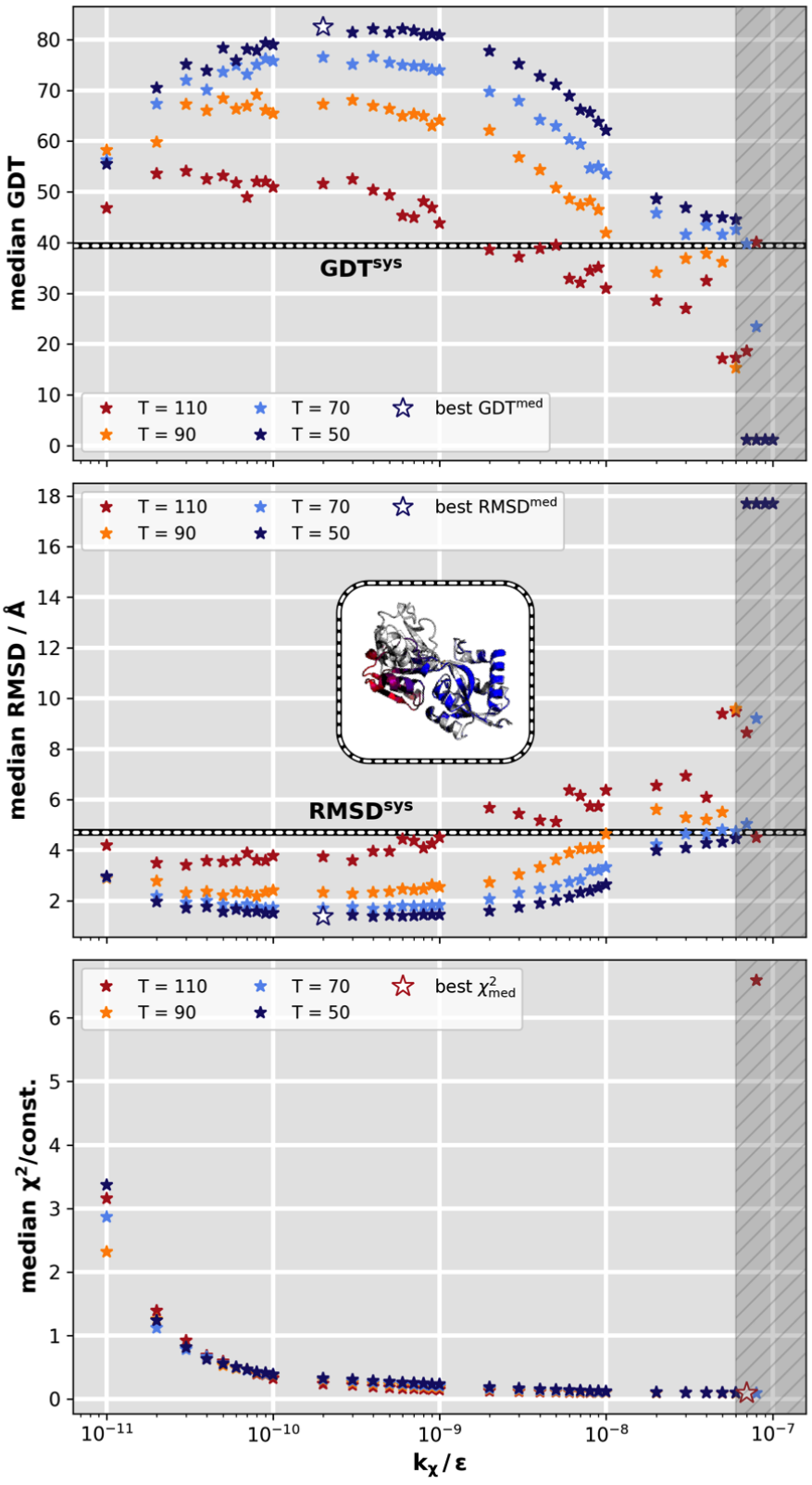


Figure D.19. Variational grid search for LAO protein a → h transition. Median GDT, median target RMSD, and median χ^2 deviation versus bias weight k_χ at different temperatures T. The variational series comprised 148 simulations in total. Best (maximum) GDT, best (minimum) RMSD, and best (minimum) V_{XS} are marked by a white star, each outlined in the color of the related temperature.

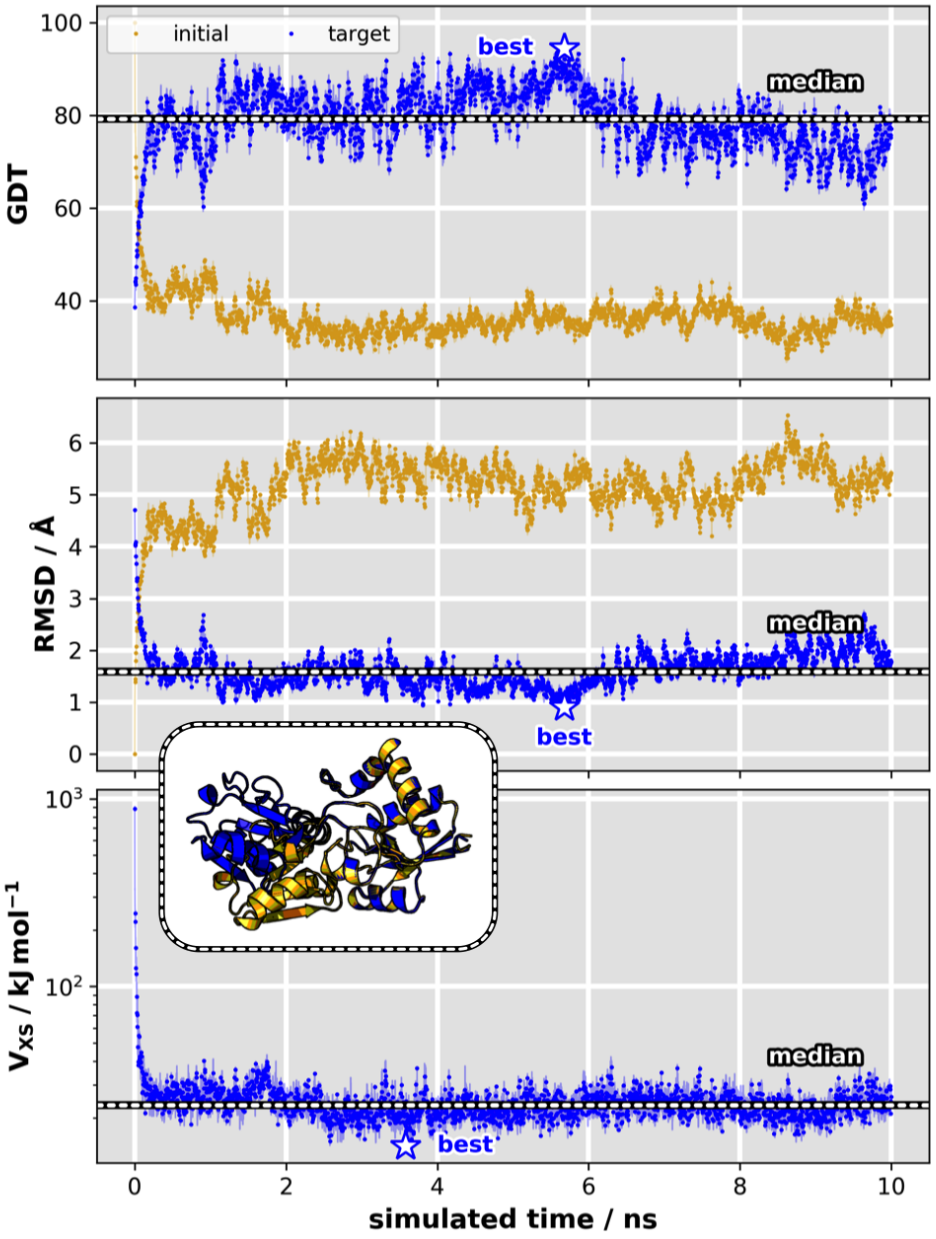


Figure D.20. Explicit-solvent MD simulation results for LAO protein h → a transition. Results are shown for parameters $(T,\, k_\chi) = \left(300\,\mathrm{K},\, 1 \cdot 10^{-9}\,\mathrm{kJ/mol}\right)$. Initial and target GDT (top), initial and target RMSD (middle), and bias energy (bottom) versus simulated time.

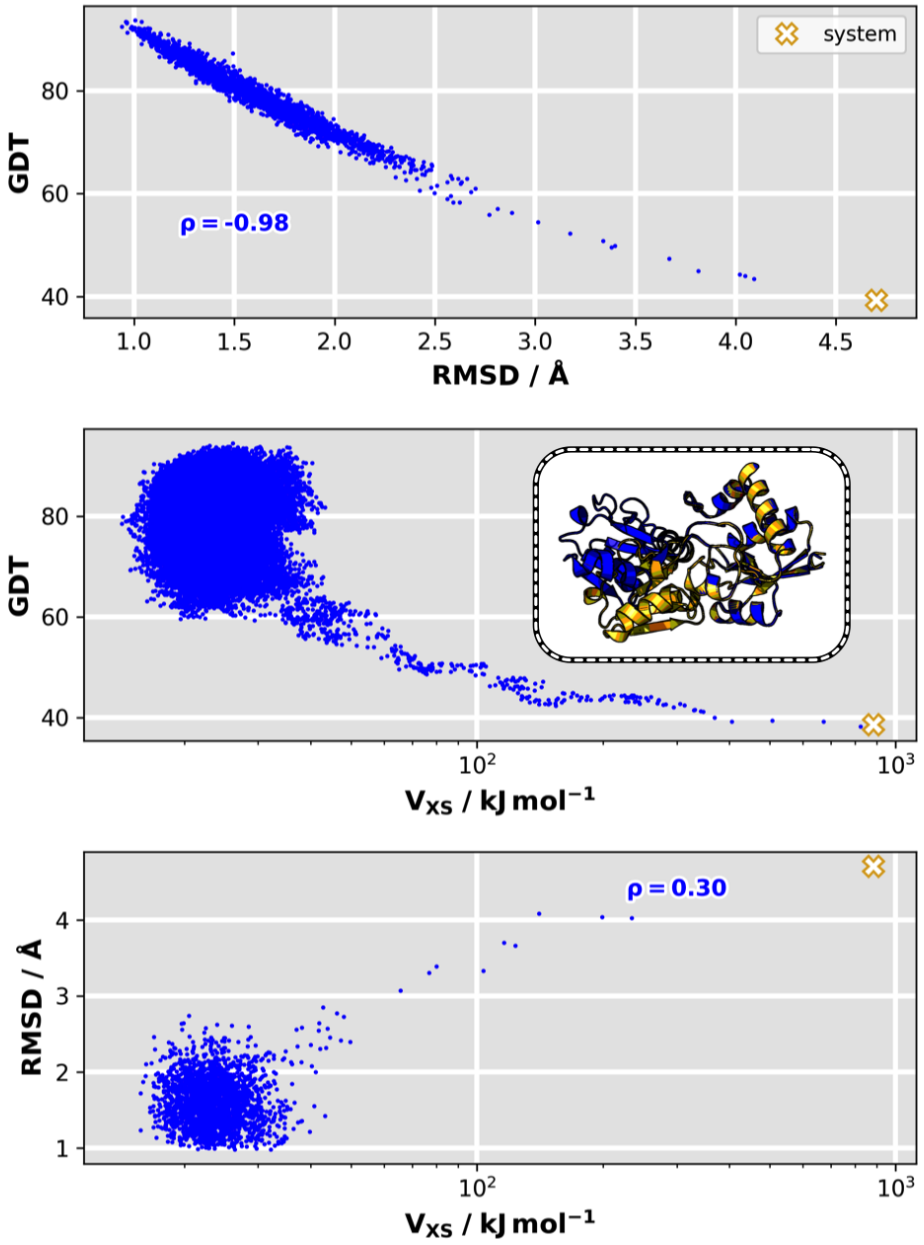


Figure D.21. LAO protein h $\rightarrow$ a transition in explicit-solvent MD: Mutual correlations among GDT, RMSD, and bias potential. Target GDT versus target RMSD (top), target GDT versus V_{XS} (middle), target RMSD versus V_{XS} (bottom). ρ is the Pearson correlation of each two plotted quantities.

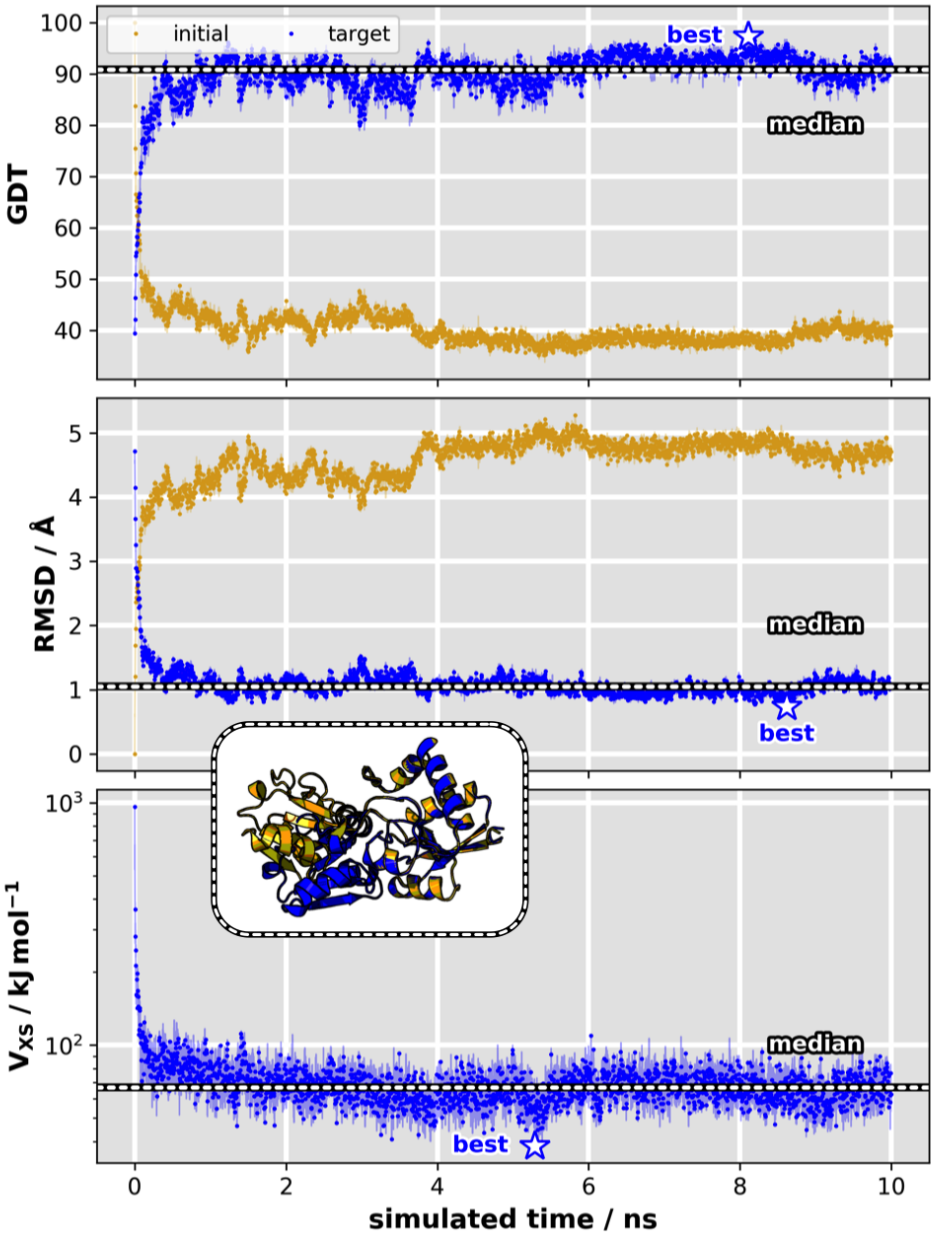


Figure D.22. Explicit-solvent MD simulation results for LAO protein a → h transition. Results are shown for parameters $(T,\, k_\chi) = \left(300\,\mathrm{K},\, 1 \cdot 10^{-9}\,\mathrm{kJ/mol}\right)$. Initial and target GDT (top), initial and target RMSD (middle), and bias energy (bottom) versus simulated time.

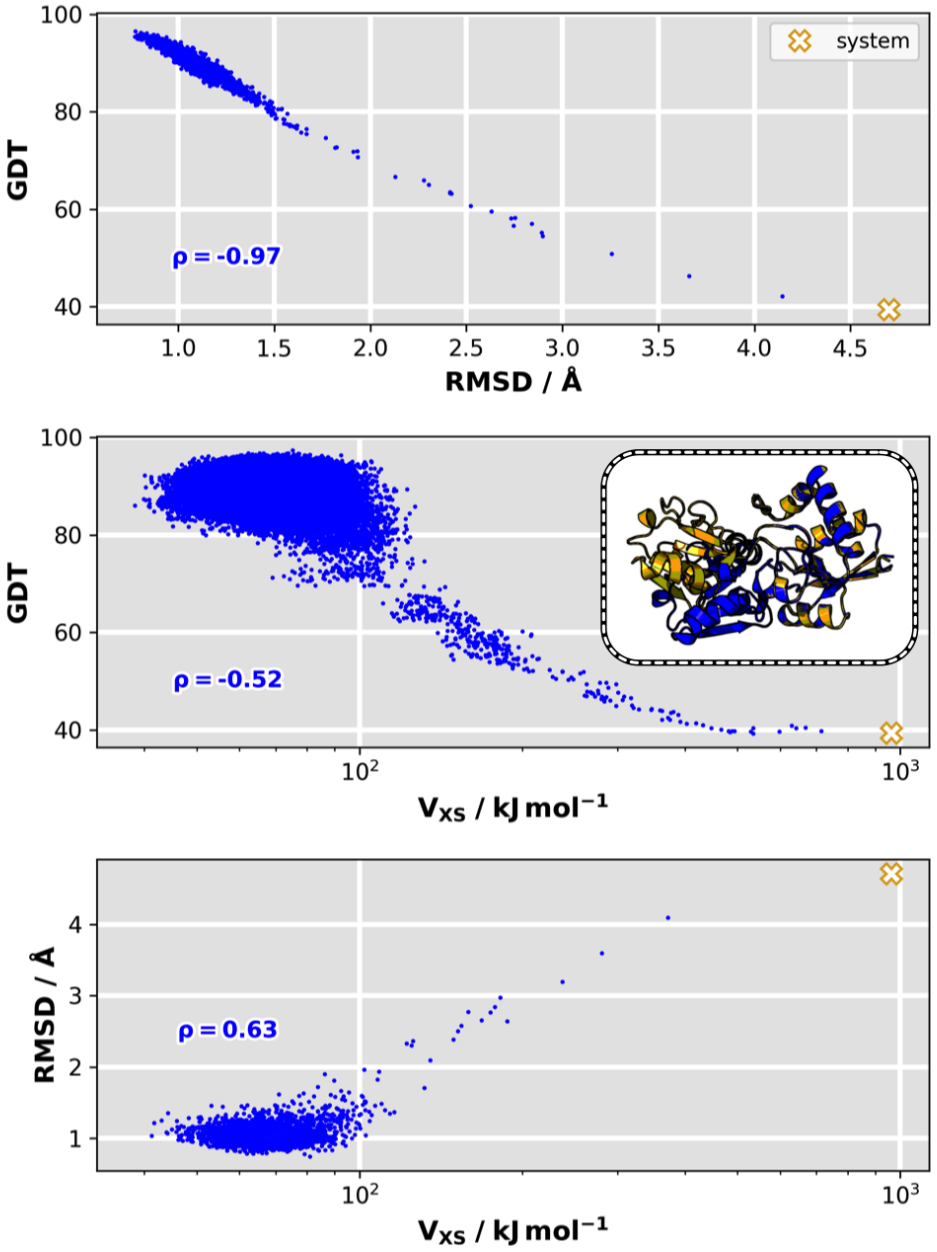


Figure D.23. LAO protein a → h transition in explicit-solvent MD: Mutual correlations among GDT, RMSD, and bias potential. Target GDT versus target RMSD (top), target GDT versus V_{XS} (middle), target RMSD versus V_{XS} (bottom). ρ is the Pearson correlation of each two plotted quantities.

D.6 Adenylate Kinase

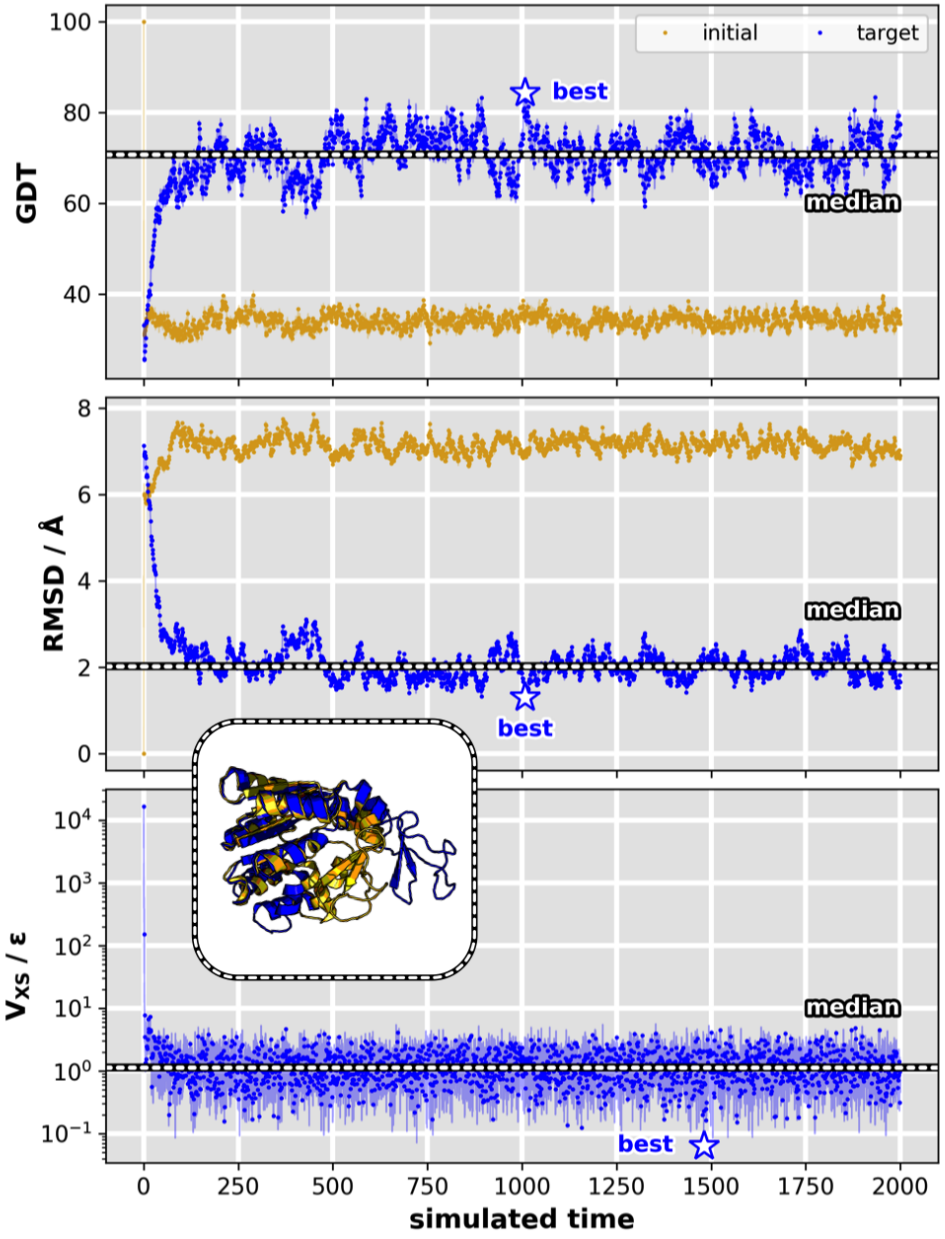


Figure D.24. `XSBM` **simulation results for ADK c → o transition.** Results are shown for parameters $(T,\, k_\chi) = (50,\, 7 \cdot 10^{-9}\,\varepsilon)$. Initial and target GDT (top), initial and target RMSD (middle), and bias energy (bottom) versus simulated time.

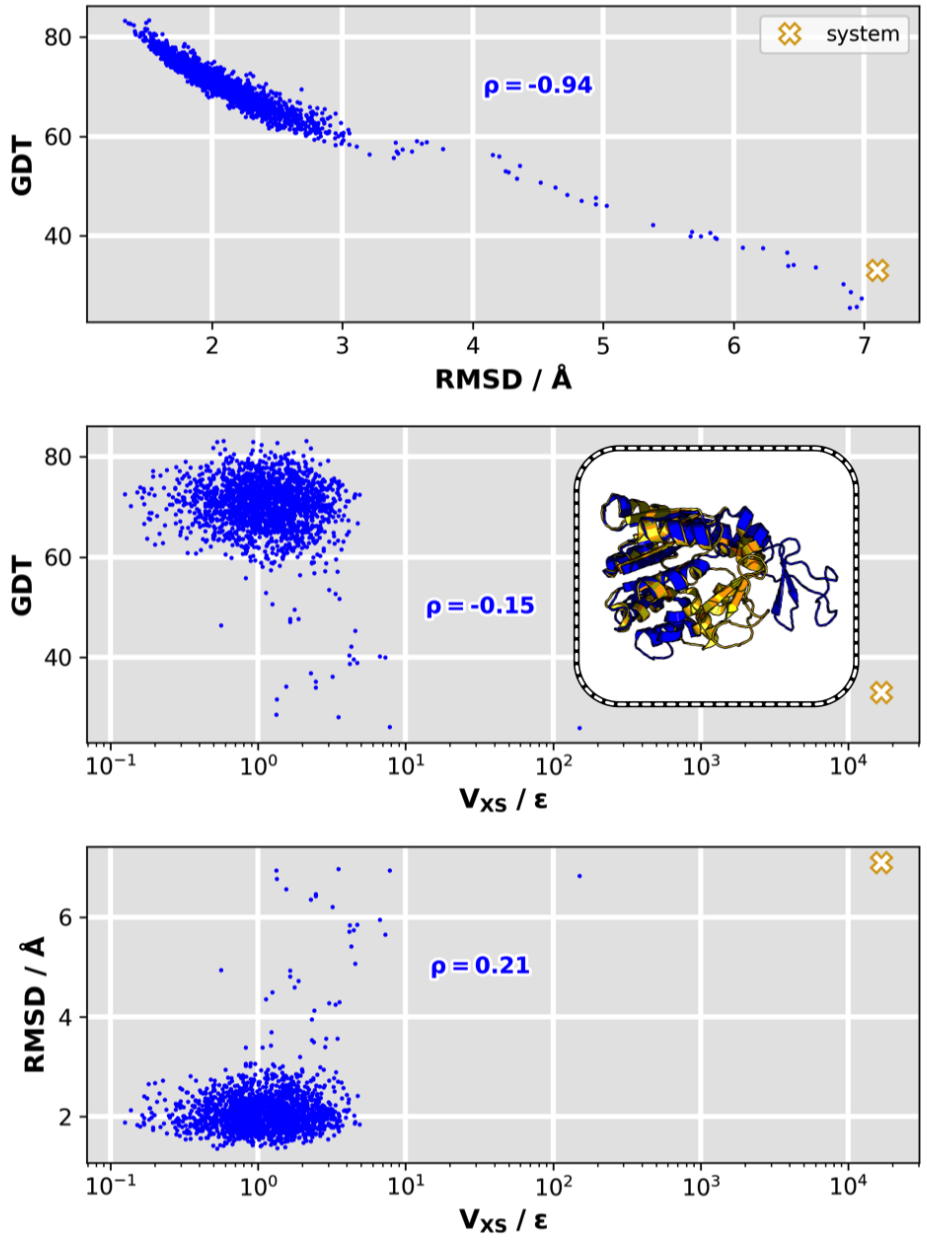


Figure D.25. ADK c → o transition in `XSBM`: Mutual correlations among GDT, RMSD, and bias potential. Target GDT versus target RMSD (top), target GDT versus V_{XS} (middle), target RMSD versus V_{XS} (bottom). ρ is the Pearson correlation of each two plotted quantities.

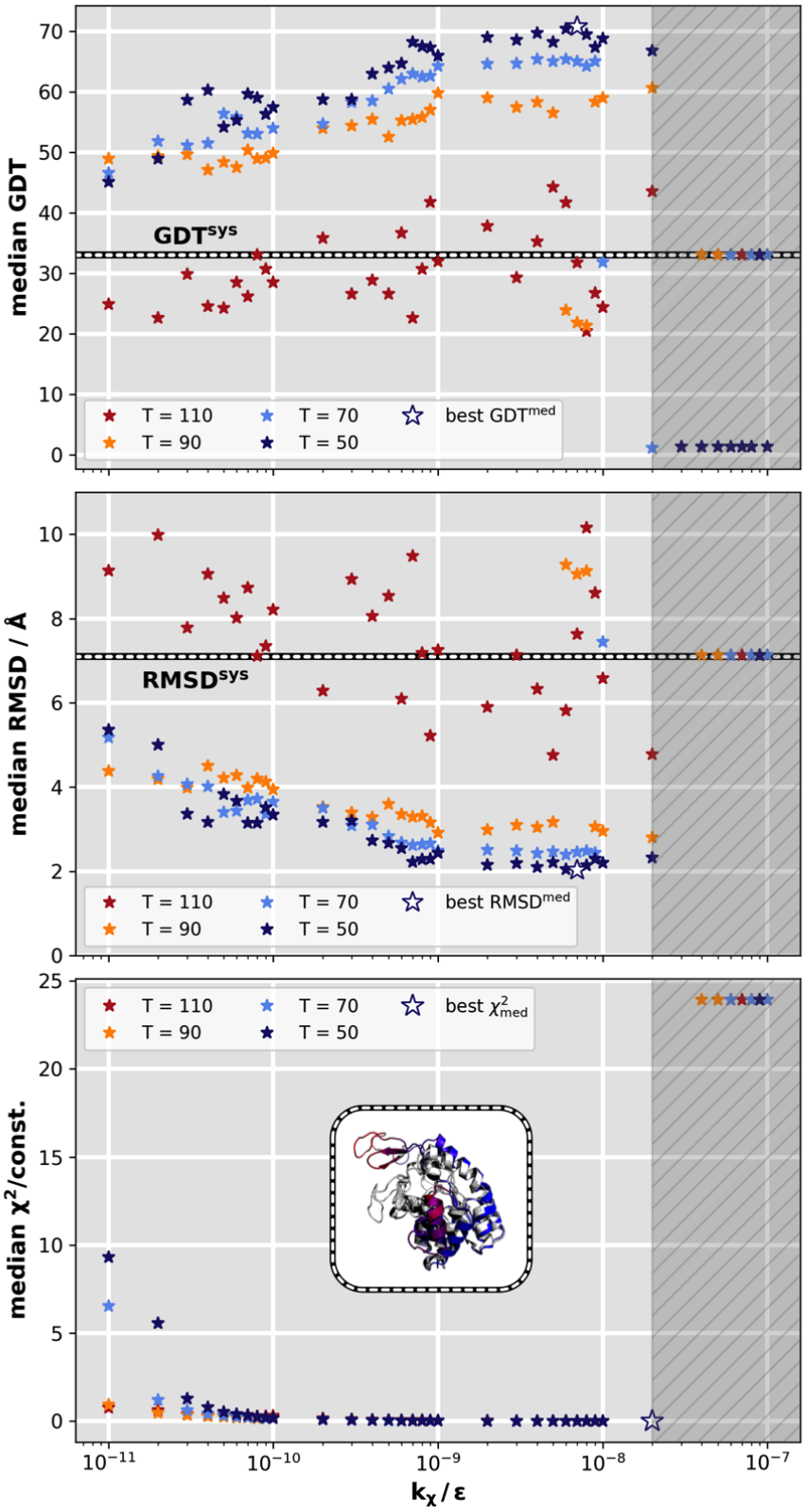


Figure D.26. Variational grid search for ADK c → o transition. Median GDT, median target RMSD, and median χ^2 deviation versus bias weight k_χ at different temperatures T. The variational series comprised 148 simulations in total. Best (maximum) GDT, best (minimum) RMSD, and best (minimum) V_{XS} are marked by a white star, each outlined in the color of the related temperature.

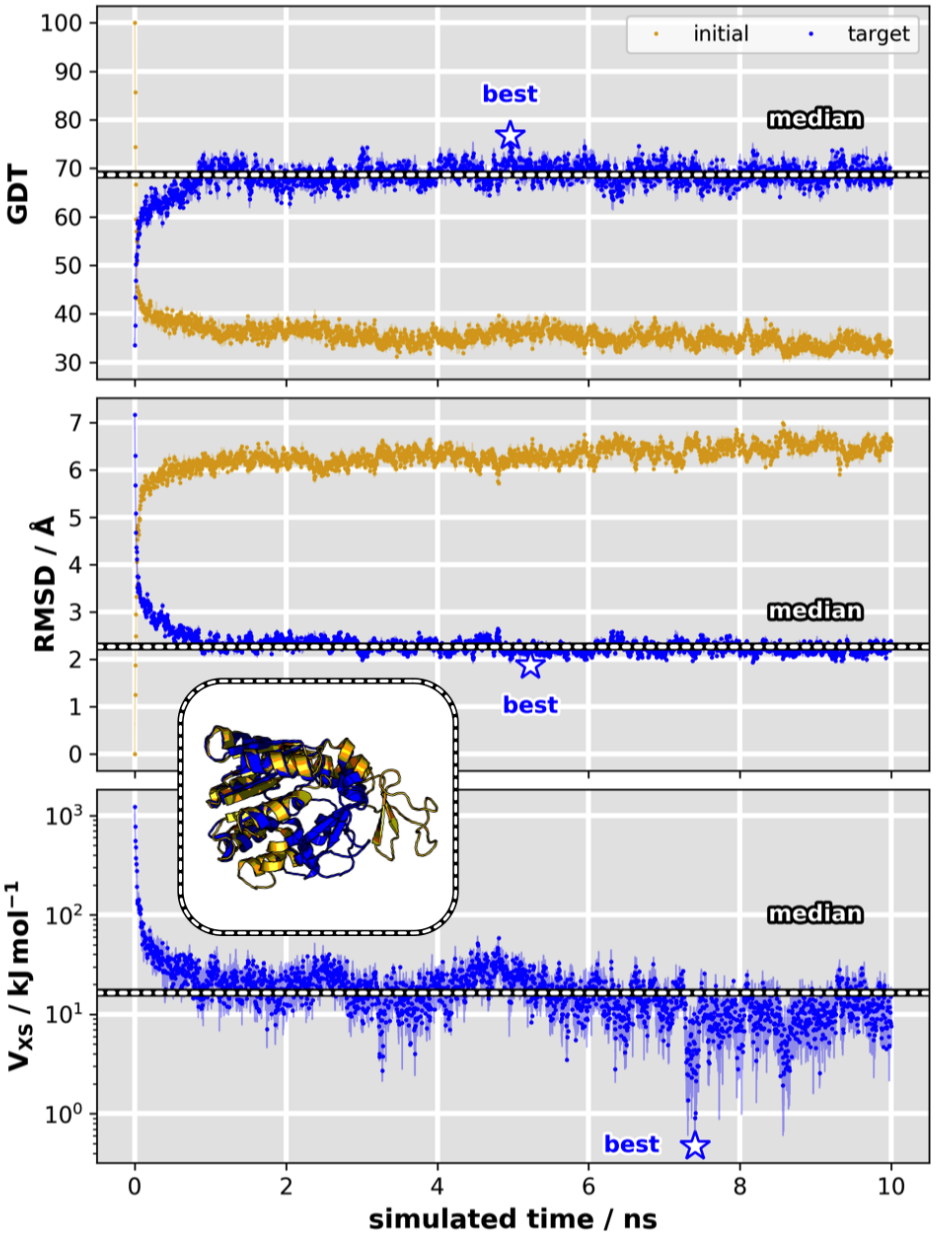


Figure D.27. Explicit-solvent MD simulation results for ADK o → c transition. Results are shown for parameters $(T,\, k_\chi) = \left(300\,\mathrm{K},\, 5 \cdot 10^{-10}\,\mathrm{kJ/mol}\right)$. Initial and target GDT (top), initial and target RMSD (middle), and bias energy (bottom) versus simulated time.

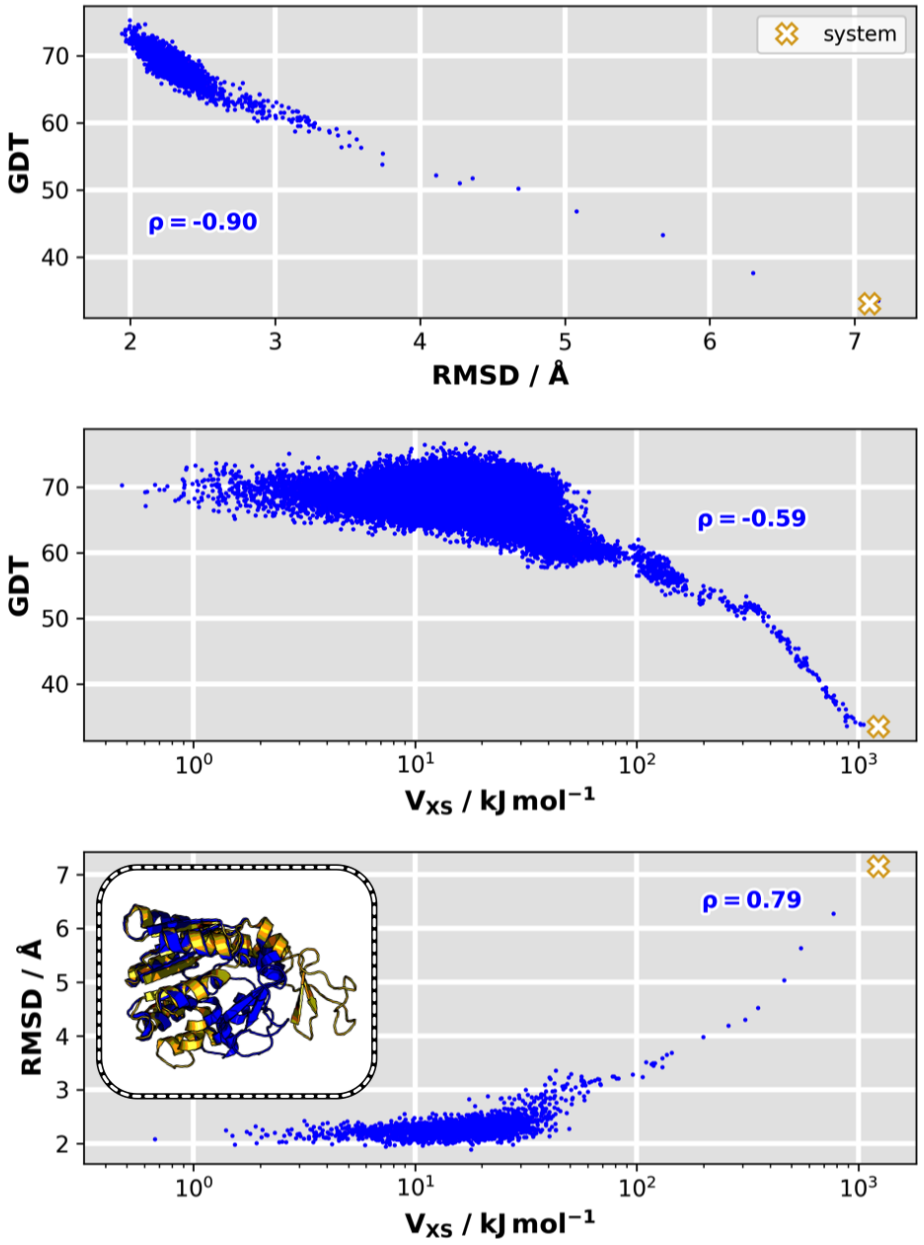


Figure D.28. ADK o → c transition in explicit-solvent MD: Mutual correlations among GDT, RMSD, and bias potential. Target GDT versus target RMSD (top), target GDT versus V_{XS} (middle), target RMSD versus V_{XS} (bottom). ρ is the Pearson correlation of each two plotted quantities.

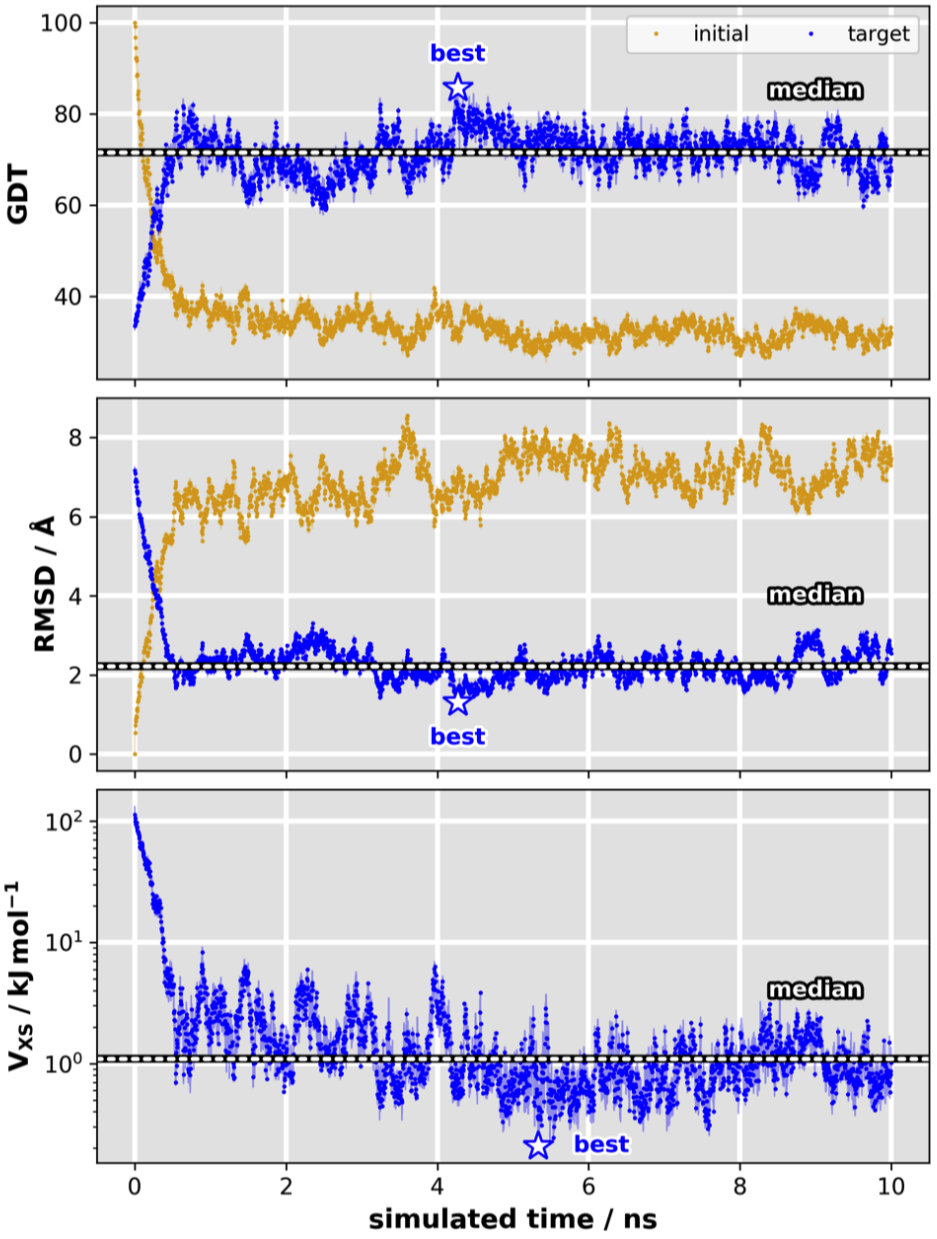


Figure D.29. Explicit-solvent MD simulation results for ADK c → o transition. Results are shown for parameters $(T, k_\chi) = (300\,\mathrm{K}, 5 \cdot 10^{-10}\,\mathrm{kJ/mol})$. Initial and target GDT (top), initial and target RMSD (middle), and bias energy (bottom) versus simulated time.

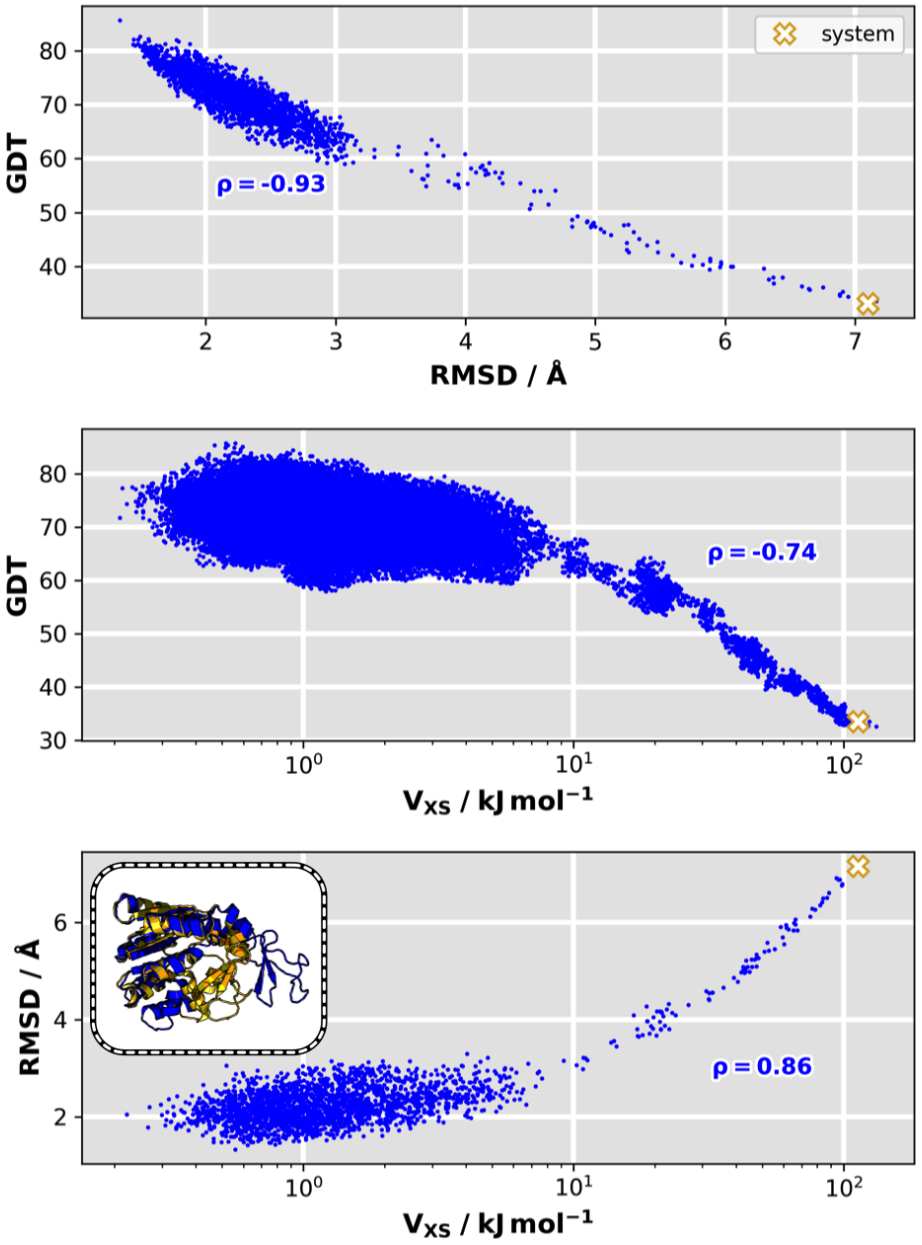


Figure D.30. ADK c → o transition in explicit-solvent MD: Mutual correlations among GDT, RMSD, and bias potential. Target GDT versus target RMSD (top), target GDT versus V_{XS} (middle), target RMSD versus V_{XS} (bottom). ρ is the Pearson correlation of each two plotted quantities.

E

Appendix to "PROJECT: Optimizing Biomolecular Simulation Parameters with Computational Intelligence"

E.1 FLAPS Results

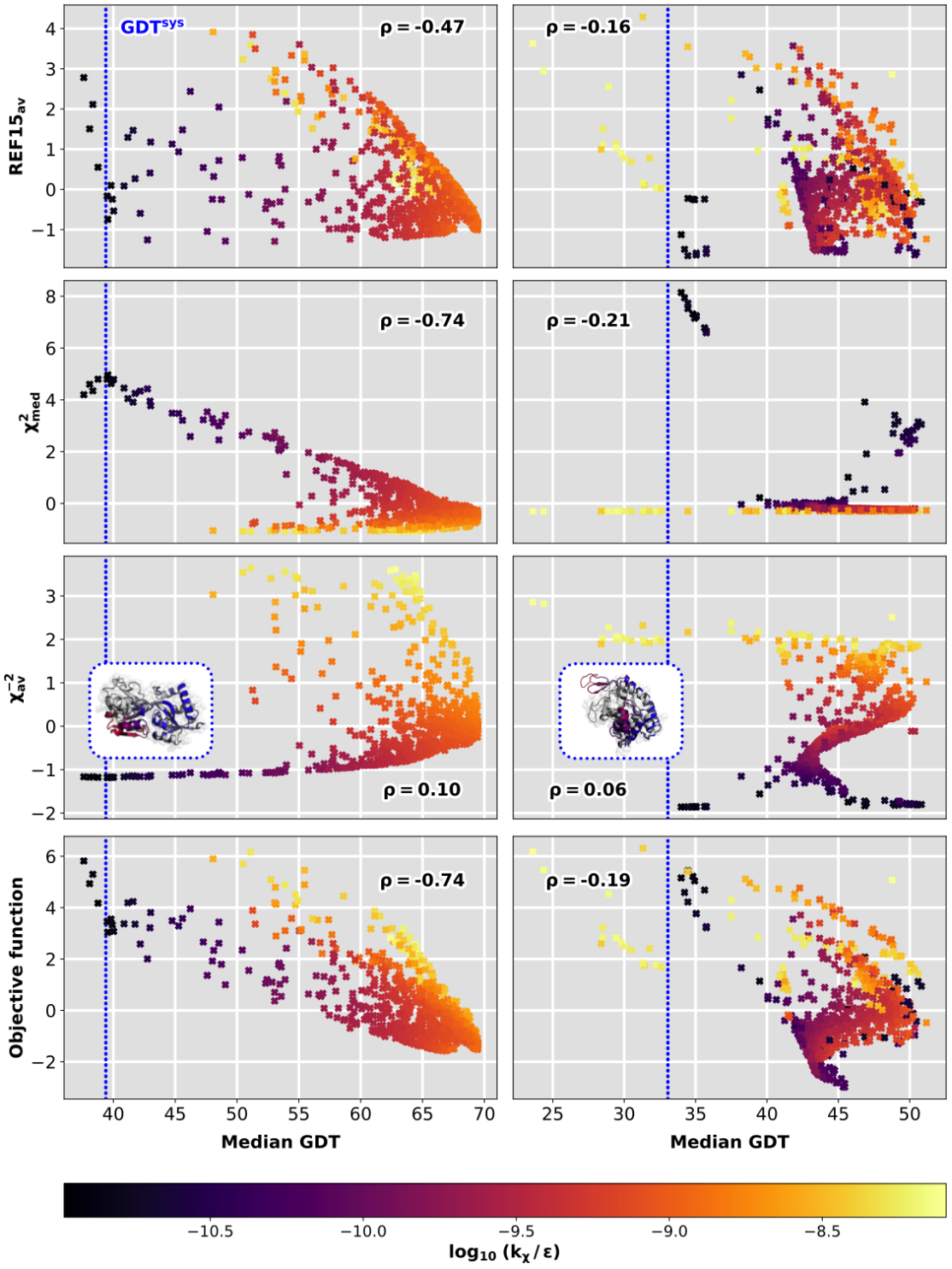


Figure E.1. OF and its standardized responses versus median GDT. LAO protein's a → h transition (left) and ADK's c → o transition (right). GDT^{sys}, GDT between initial and target structure of each test system, ρ, Pearson correlation of each quantity and median GDT, k_χ, bias weight. Data from all runs combined[173].

Seed	1790954	1791103	1791104	1791105	1791106	1792508	1792509
ρ^{RMSD}	0.85	0.70	0.88	0.63	0.84	0.64	0.79
ρ^{GDT}	-0.94	-0.84	-0.87	-0.68	-0.87	-0.74	-0.80
$\min(f)$	-2.34	-2.03	-1.79	-2.15	-1.99	-1.60	-1.78
$\max(f)$	8.32	8.31	5.86	4.38	4.41	4.72	6.27
Best simulation in terms of OF							
k_χ / ε	2.170 e-10	1.073 e-10	3.339 e-11	6.654 e-11	5.080 e-11	1.493 e-09	5.869 e-11
T	13.19	10.36	28.82	10.47	11.05	13.60	14.65
$\mathrm{RMSD}^{\mathrm{med}}$ / Å	2.1	2.1	2.2	2.2	2.2	2.9	2.2
$\mathrm{GDT}^{\mathrm{med}}$	70.59	69.96	69.22	69.44	69.44	64.81	69.22
Best simulation in terms of $\mathrm{RMSD}^{\mathrm{med}}$							
$\mathrm{RMSD}^{\mathrm{med}}$ / Å	2.1	2.1	2.2	2.1	2.1	2.2	2.1
$f(\mathrm{RMSD}^{\mathrm{med}})$	-2.03	-2.02	-1.73	-1.79	-1.55	-0.90	-1.45
k_χ / ε	3.001 e-10	2.320 e-10	3.422 e-11	2.401 e-10	4.190 e-10	5.652 e-10	2.395 e-10
T	11.98	10.43	29.63	10.18	10.03	23.67	12.93
$\mathrm{GDT}^{\mathrm{med}}$	70.69	70.80	69.54	70.91	70.69	68.17	70.69
Best simulation in terms of $\mathrm{GDT}^{\mathrm{med}}$							
$\mathrm{GDT}^{\mathrm{med}}$	70.69	70.80	69.54	70.91	70.69	68.17	70.69
$f(\mathrm{GDT}^{\mathrm{med}})$	-2.03	-2.02	-1.73	-1.79	-1.55	-1.08	-1.45
k_χ / ε	3.001 e-10	2.320 e-10	3.422 e-11	2.401 e-10	4.190 e-10	6.429 e-10	2.395 e-10
T	11.98	10.43	29.63	10.18	10.03	20.89	12.93
$\mathrm{RMSD}^{\mathrm{med}}$ / Å	2.1	2.1	2.2	2.1	2.1	2.2	2.1

Table E.1. `FLAPS` **optimization results for** `XSBM` **h → a transition of LAO protein**[173]**.** OF, objective function f, ρ^{RMSD}, Pearson correlation of OF and median RMSD, ρ^{GDT}, Pearson correlation of OF and median GDT, k_χ, bias weight, T, temperature, $\mathrm{RMSD}^{\mathrm{med}}$, median RMSD, $\mathrm{GDT}^{\mathrm{med}}$, median GDT.

Seed	1795691	1797335	1797338	1797339	1798723	1801054	1810891
ρ^{RMSD}	0.58	0.33	0.40	0.41	0.49	0.03	0.42
ρ^{GDT}	-0.85	-0.61	-0.77	-0.76	-0.81	-0.32	-0.74
$\min(f)$	-1.42	-1.54	-1.20	-1.50	-1.57	-1.44	-1.62
$\max(f)$	6.92	6.38	7.09	6.63	9.05	5.45	7.47
Best simulation in terms of OF							
k_χ / ε	1.969 e-09	1.283 e-09	1.858 e-09	1.810 e-09	1.970 e-09	8.218 e-10	1.819 e-09
T	16.90	10.02	10.17	10.28	10.56	10.51	10.09
$\mathrm{RMSD}^{\mathrm{med}}$ / Å	2.8	3.1	2.9	2.9	2.8	3.3	2.9
$\mathrm{GDT}^{\mathrm{med}}$	63.20	62.38	63.67	63.55	63.78	60.63	63.55
Best simulation in terms of $\mathrm{RMSD}^{\mathrm{med}}$							
$\mathrm{RMSD}^{\mathrm{med}}$ / Å	2.6	2.6	2.6	2.6	2.6	2.7	2.6
$f(\mathrm{RMSD}^{\mathrm{med}})$	-1.19	-0.46	-0.59	-0.81	-1.21	2.15	-1.28
k_χ / ε	2.910 e-09	2.996 e-09	3.477 e-09	2.488 e-09	3.213 e-09	3.311 e-09	3.419 e-09
T	12.40	16.32	12.02	15.51	10.52	31.12	10.01
$\mathrm{GDT}^{\mathrm{med}}$	63.20	62.97	63.20	63.09	63.44	62.38	63.32
Best simulation in terms of $\mathrm{GDT}^{\mathrm{med}}$							
$\mathrm{GDT}^{\mathrm{med}}$	63.78	63.90	63.90	63.90	63.78	62.62	63.90
$f(\mathrm{GDT}^{\mathrm{med}})$	-1.28	-1.29	-1.17	-1.35	-1.57	-1.34	-1.59
k_χ / ε	2.030 e-09	2.171 e-09	2.134 e-09	2.171 e-09	1.970 e-09	1.377 e-09	2.170 e-09
T	10.84	10.67	10.07	10.45	10.56	10.22	10.23
$\mathrm{RMSD}^{\mathrm{med}}$ / Å	2.8	2.8	2.8	2.8	2.8	3.0	2.8

Table E.2. `FLAPS` **optimization results for** `XSBM` **o → c transition of ADK**[173]**.** OF, objective function f, ρ^{RMSD}, Pearson correlation of OF and median RMSD, ρ^{GDT}, Pearson correlation of OF and median GDT, k_χ, bias weight, T, temperature, $\mathrm{RMSD}^{\mathrm{med}}$, median RMSD, $\mathrm{GDT}^{\mathrm{med}}$, median GDT.

Seed	1800990	1800994	1800995	1800996	1805228	1805229	1805230
ρ^{RMSD}	0.78	0.88	0.83	0.84	0.69	0.87	0.59
ρ^{GDT}	-0.75	-0.86	-0.80	-0.81	-0.67	-0.88	-0.56
min (f)	-1.33	-1.34	-1.27	-1.51	-1.82	-1.65	-1.58
max (f)	6.28	5.45	6.76	6.00	3.59	5.53	5.25
Best simulation in terms of OF							
k_χ / ε	1.035 e-09	1.245 e-09	9.913 e-10	9.194 e-10	6.466 e-10	1.252 e-09	7.810 e-10
T	10.25	10.53	10.46	10.01	10.74	10.90	13.23
$\mathrm{RMSD}^{\mathrm{med}}$ / Å	2.2	2.0	2.2	2.2	2.3	2.0	2.2
$\mathrm{GDT}^{\mathrm{med}}$	67.02	69.33	66.91	66.81	65.44	69.33	66.17
Best simulation in terms of $\mathrm{RMSD}^{\mathrm{med}}$							
$\mathrm{RMSD}^{\mathrm{med}}$ / Å	2.0	2.0	2.0	2.0	2.0	2.0	2.0
$f(\mathrm{RMSD}^{\mathrm{med}})$	-1.13	-1.24	-1.02	-1.37	-1.50	-1.50	-0.82
k_χ / ε	1.448 e-09	1.473 e-09	1.388 e-09	1.399 e-09	1.503 e-09	1.580 e-09	1.389 e-09
T	11.94	11.66	11.65	10.67	12.74	11.76	17.35
$\mathrm{GDT}^{\mathrm{med}}$	69.44	69.44	69.44	69.44	69.33	69.44	69.22
Best simulation in terms of $\mathrm{GDT}^{\mathrm{med}}$							
$\mathrm{GDT}^{\mathrm{med}}$	69.44	69.44	69.44	69.44	69.33	69.44	69.44
$f(\mathrm{GDT}^{\mathrm{med}})$	-1.13	-1.25	-0.97	-1.48	-1.41	-1.50	-1.35
k_χ / ε	1.448 e-09	1.285 e-09	1.328 e-09	1.070 e-09	1.386 e-09	1.580 e-09	1.160 e-09
T	11.94	12.62	13.65	10.92	16.06	11.76	12.35
$\mathrm{RMSD}^{\mathrm{med}}$ / Å	2.0	2.0	2.0	2.0	2.0	2.0	2.0

Table E.3. `FLAPS` **optimization results for** `XSBM` **a** $\rightarrow$ **h transition of LAO protein**[173]**.** OF, objective function f, ρ^{RMSD}, Pearson correlation of OF and median RMSD, ρ^{GDT}, Pearson correlation of OF and median GDT, k_χ, bias weight, T, temperature, $\mathrm{RMSD}^{\mathrm{med}}$, median RMSD, $\mathrm{GDT}^{\mathrm{med}}$, median GDT.

Seed	1799347	1799348	1801030	1801031	1801032	1801033	1801034
ρ^{RMSD}	0.41	0.65	0.79	0.62	0.65	0.71	0.56
ρ^{GDT}	0.29	-0.15	-0.53	-0.17	-0.31	-0.29	-0.11
min (f)	-1.86	-2.50	-1.98	-2.61	-1.84	-2.37	-1.49
max (f)	4.75	7.42	6.30	6.08	4.76	6.54	5.68
Best simulation in terms of OF							
k_χ / ε	1.071 e-10	3.888 e-11	6.902 e-11	4.917 e-11	3.563 e-11	6.541 e-11	3.634 e-10
T	26.58	41.99	11.82	37.45	45.43	15.06	10.53
$\mathrm{RMSD}^{\mathrm{med}}$ / Å	4.4	4.4	4.3	4.4	4.4	4.3	4.2
$\mathrm{GDT}^{\mathrm{med}}$	43.11	45.33	45.44	44.27	44.57	45.44	43.46
Best simulation in terms of $\mathrm{RMSD}^{\mathrm{med}}$							
$\mathrm{RMSD}^{\mathrm{med}}$ / Å	3.5	3.5	3.5	3.6	3.6	3.5	3.4
$f(\mathrm{RMSD}^{\mathrm{med}})$	-0.36	-0.50	0.26	-0.38	-0.52	0.38	-1.00
k_χ / ε	8.490 e-10	8.294 e-10	8.194 e-10	6.801 e-10	6.623 e-10	1.039 e-09	1.117 e-09
T	31.22	34.12	27.42	43.01	44.02	28.40	17.73
$\mathrm{GDT}^{\mathrm{med}}$	49.71	49.65	49.65	48.83	48.95	50.12	49.94
Best simulation in terms of $\mathrm{GDT}^{\mathrm{med}}$							
$\mathrm{GDT}^{\mathrm{med}}$	50.23	49.65	50.47	50.81	50.59	50.35	51.17
$f(\mathrm{GDT}^{\mathrm{med}})$	4.75	-0.50	-0.63	0.43	0.34	-0.11	-0.73
k_χ / ε	3.122 e-11	8.294 e-10	3.134 e-11	2.229 e-11	5.289 e-09	3.515 e-10	2.163 e-09
T	22.77	34.12	13.00	42.58	27.35	30.33	13.43
$\mathrm{RMSD}^{\mathrm{med}}$ / Å	4.9	3.5	4.9	4.9	5.3	4.0	3.5

Table E.4. `FLAPS` **optimization results for** `XSBM` **c** $\rightarrow$ **o transition of ADK**[173]**.** OF, objective function f, ρ^{RMSD}, Pearson correlation of OF and median RMSD, ρ^{GDT}, Pearson correlation of OF and median GDT, k_χ, bias weight, T, temperature, $\mathrm{RMSD}^{\mathrm{med}}$, median RMSD, $\mathrm{GDT}^{\mathrm{med}}$, median GDT.

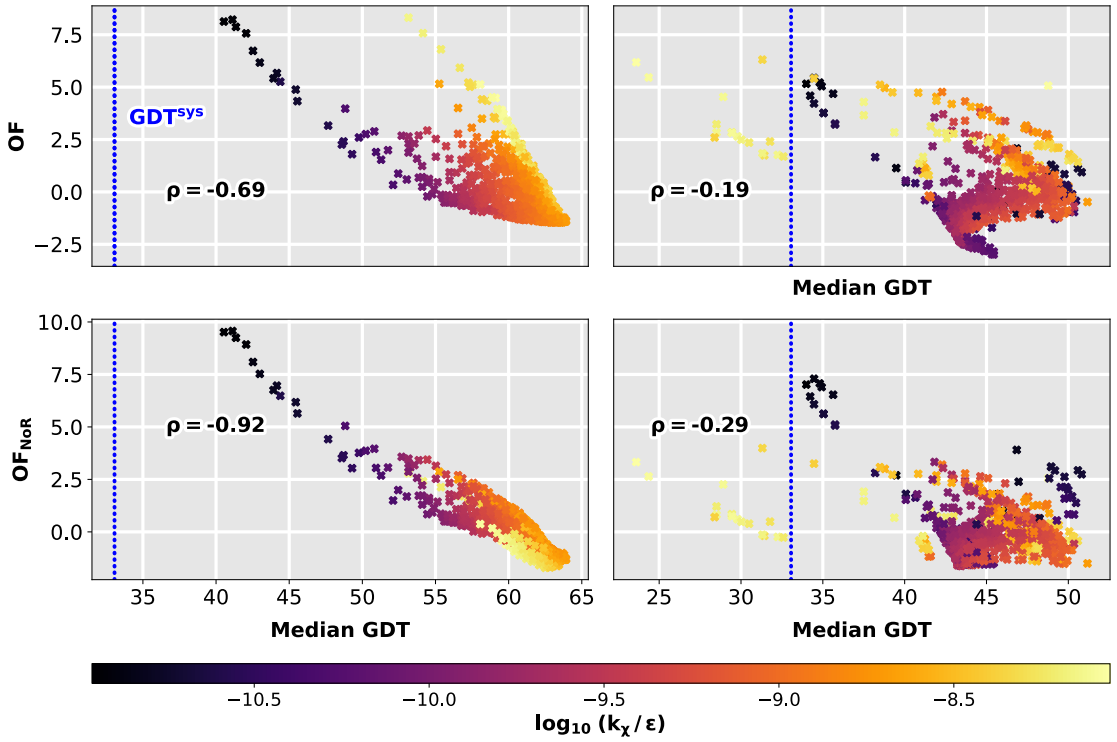


Figure E.2. OF with and without regularizer versus median GDT for ADK. o → c transition (left) and c → o transition (right). GDT^{sys}, GDT between initial and target structure, OF, objective function with regularizing χ_{av}^{-2}, OF_{NoR}, objective function without regularizing χ_{av}^{-2}, ρ, Pearson correlation of each quantity and median GDT, k_χ, bias weight. Data from all runs combined[173].

E.2 Analyzing Swarm Convergence

In the context of PSO, convergence can refer to (i) the sequential convergence of solutions, i.e., all particles have reached a particular and possibly (but not necessarily) optimal position, or (ii) convergence of either the personal bests or the global best to a local optimum, independent of the swarm's behavior as a whole. I considered the first convergence concept. Analyzing sequential convergence has issued guidelines for selecting PSO parameters that presumably cause the particles to converge to some point in the search space[177]. A common view is that the swarm varies between exploration and exploitation, which involves adapting the algorithm and its parameters to properly balance these behaviors. This is important to avoid early entrapment in local optima yet allow for a reasonable convergence rate. A primary focus of past research thus was increasing the algorithm's adaptability by making it more complex. However, it is still not clear how the swarm's behavior affects the actual optimization performance, in particular for dynamic environments. This renders the opposite approach, i.e., implementing PSO setups that perform well regardless of how the swarm's behavior is to be interpreted, equally valid. Inspired by Occam's razor, this view is based on the principle that PSO should be simplified to the greatest possible extent without compromising its performance. As a metaheuristic, PSO can only be proven correct in the sense of demonstrating its efficacy empirically by a finite number of computational experiments. This poses the risk of making errors in its description and implementation. Simplifying PSO was first suggested by Kennedy in 1997[175]. The concept has been studied more extensively[184,212], where the optimization performance was found to improve across different problems and the parameters were easier to tune.

Following the "simplifying PSO" paradigm, `FLAPS` rests on a slim PSO core with only few parameters. In the velocity update, cognitive and social influence have random components limited by their associated acceleration coefficients ϕ_1 and ϕ_2. The latter are hyperparameters of the algorithm (see Algorithm 2). Originally, ϕ_1 and ϕ_2 were chosen as 2 because this leads to unity weights for cognitive and social influence on average, which makes the swarm contract to the current g_{best}. Adding this velocity to a particle's current position introduces a stochastic tendency to return towards the previous best positions that have demonstrated promise[175]. This randomness keeps the particles from moving directly toward the global and personal best. It facilitates the exploration of new solutions near the current best positions and diversifies the particles for more effective searches. I find that further diversity enhancement is not needed for the presented application. Supplementary Figs. E.3 to E.6 show the swarm in the current OF topology after each generation. For all protein systems, the swarm converged to a stable topology and contracted around a functional parameter combination. Quantitative swarm diversity measures include position diversity, velocity diversity, and cognitive diversity. Based on the particles' positions, I calculated a relative swarm spread with respect to the swarm's initial state,

$$SwS = \frac{\operatorname{std}(\operatorname{norm} T) + \operatorname{std}(\operatorname{norm} k_\chi)}{SwS_0}\,, \tag{E.1}$$

after each generation (see Supplementary Figs. E.3 to E.6, bottom left). $\operatorname{std}(x)$ is the standard deviation of quantity x, $\operatorname{norm} x$ is the min-max scaled quantity $(x - \min x) / (\max x - \min x)$, and SwS_0 is the absolute initial swarm spread. Starting from a value of 1 by definition, the swarm spread significantly dropped down in the first two to five generations. Depending on the system, it stabilized at fractions of 0.2 to 0.4 of each initial swarm spread with a tendency to further decrease.

In addition, I considered the Euclidean distance of the current normalized global best, $\operatorname{norm} g^g_{\text{best}} = (\operatorname{norm} T, \operatorname{norm} k_\chi)^g_{g_{\text{best}}}$, from that of each previous generation, $\operatorname{norm} g^{g-1}_{\text{best}}$ (see Supplementary Figs. E.3 to E.6, bottom right). This g_{best} fluctuation minimized clearly for only one of the considered systems (see

Supplementary Fig. E.5). As different parameter combinations can equally yield useful results in physico-empirical SBMs, this is no surprise. Many factors influence the convergence behavior and performance of particle-swarm based algorithms, including selection of the acceleration coefficients, velocity clamping, and the swarm's communication network. As a metaheuristic that implements a form of stochastic optimization, PSO is not guaranteed to find the globally optimal solution. It rather is a practical strategy that guides the optimization process in order to efficiently explore the search space and find near-optimal solutions. While built on the "simplifying PSO" paradigm, `FLAPS` can easily be complemented by concepts such as inertia weight[176] and swarm constriction[177], diversity increasing mechanisms, or flexible termination criteria based on, e.g, the swarm spread or the global best's fluctuation. If desired, the communication pattern can also be adapted towards local geometrical or social topologies.

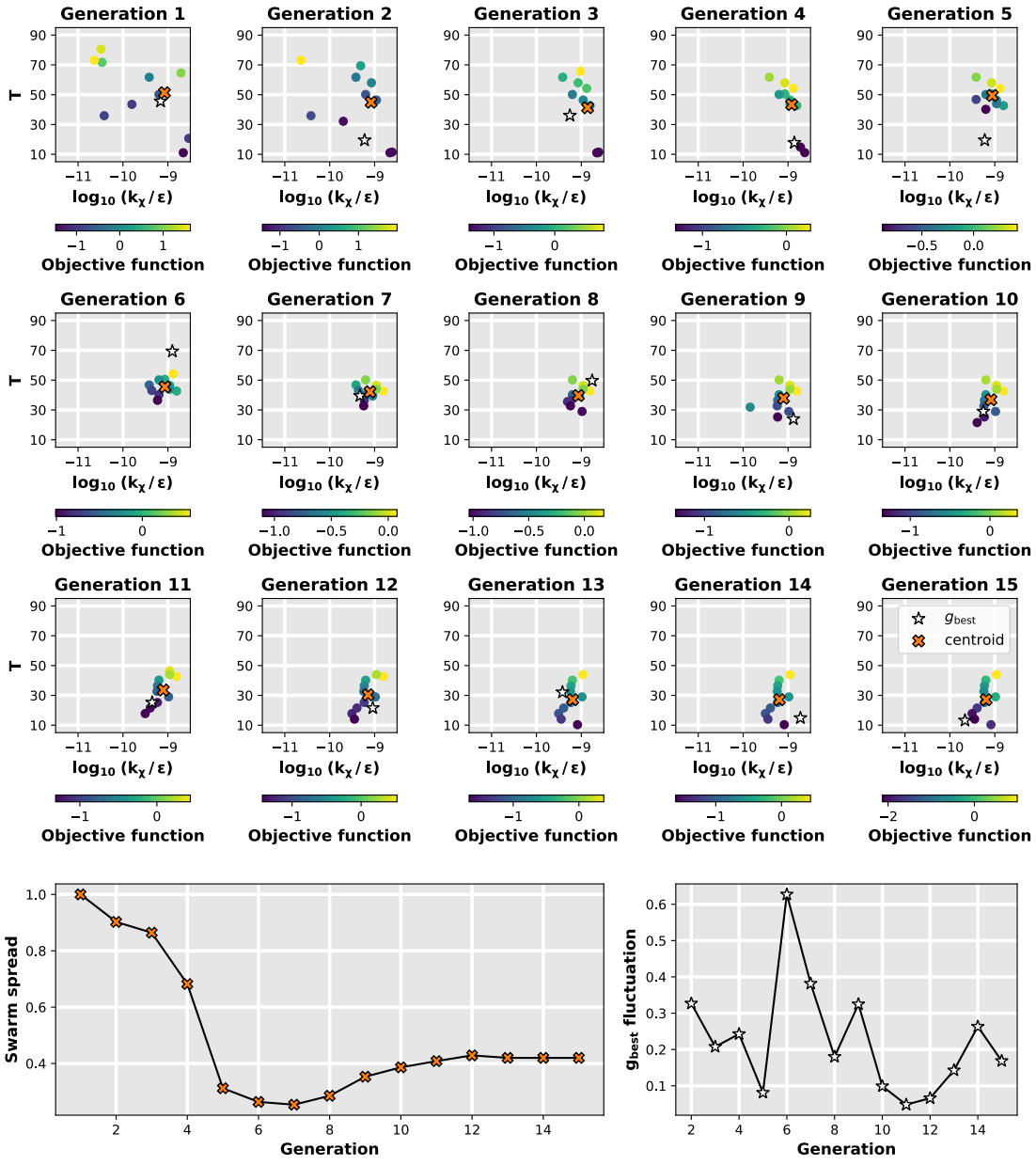


Figure E.3. Swarm convergence for `XSBM` h → a transition of LAO protein (seed 1790954). **Top:** Dynamically evolving OF topology after each generation. The current global best position, g_{best}, and the swarm's centroid are marked. k_χ, bias weight, T, temperature. **Bottom:** Evolution of swarm spread SwS (left) and g_{best} fluctuation (right) during the `FLAPS` optimization. Reproduced from Ref.[173] under CC BY 4.0.

Figure E.4. Swarm convergence for `XSBM` o → c transition of ADK (seed 1795691). Top: Dynamically evolving OF topology after each generation. The current global best position, g_{best}, and the swarm's centroid are marked. k_χ, bias weight, T, temperature. **Bottom:** Evolution of swarm spread SwS (left) and g_{best} fluctuation (right) during the `FLAPS` optimization. Reproduced from Ref.[173] under CC BY 4.0.

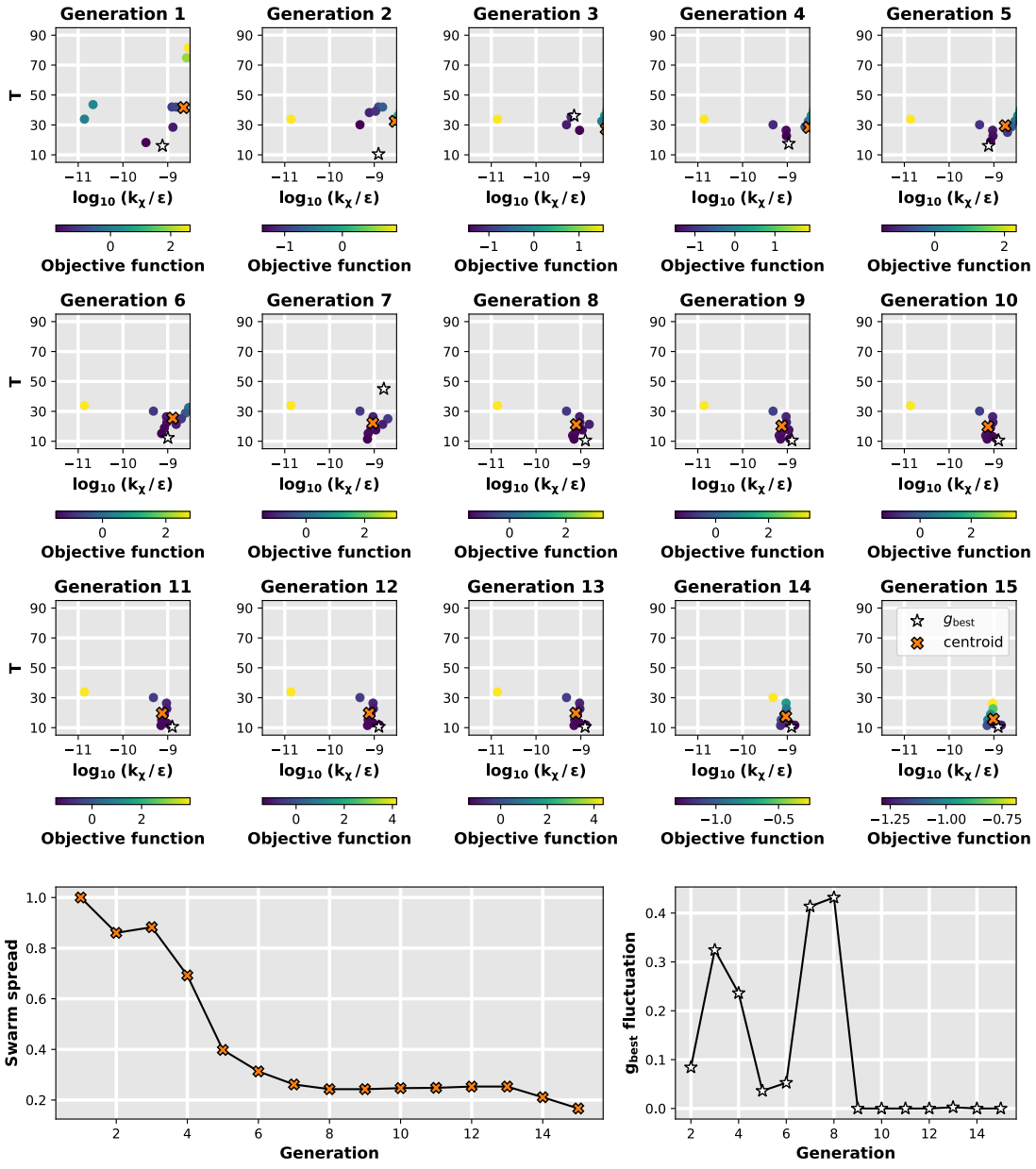


Figure E.5. Swarm convergence for `XSBM` a → h transition of LAO protein (seed 1800994). **Top:** Dynamically evolving OF topology after each generation. The current global best position, g_{best}, and the swarm's centroid are marked. k_χ, bias weight, T, temperature. **Bottom:** Evolution of swarm spread *SwS* (left) and g_{best} fluctuation (right) during the `FLAPS` optimization. Reproduced from Ref.[173] under CC BY 4.0.

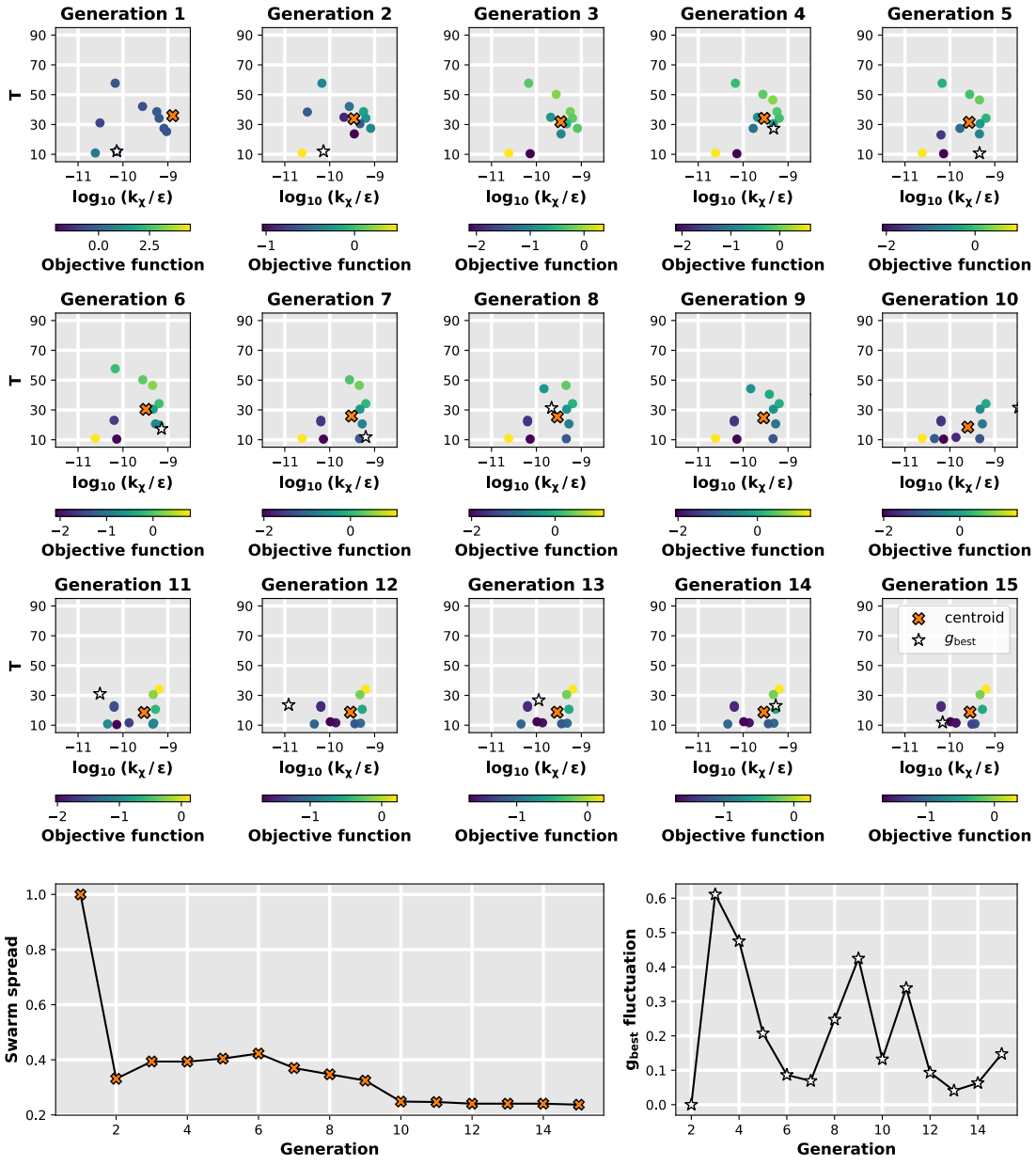


Figure E.6. Swarm convergence for `XSBM` c → o transition of ADK (seed 1801030). Top: Dynamically evolving OF topology after each generation. The current global best position, g_{best}, and the swarm's centroid are marked. k_χ, bias weight, T, temperature. **Bottom:** Evolution of swarm spread SwS (left) and g_{best} fluctuation (right) during the `FLAPS` optimization. Reproduced from Ref.[173] under CC BY 4.0.

E.3 Comparison to Grid Search

I performed comparative grid-search optimizations for the presented protein systems. In swarm-based approaches like `FLAPS`, each generation depends on the particles' positions and fitnesses attained in the previous generation (see velocity update in Algorithm 2). This recursive dependency limits the intrinsic parallelizability of such algorithms. In contrast, a grid search is embarrassingly parallel, which hinders a direct speed-up comparison. To ensure comparability of the two approaches, I considered $n_{k_\chi} \times n_T = 15 \times 10 = 150$ equidistant grid points within $-11 \leq \log_{10}(k_\chi) \leq -8$ and $10 \leq T \leq 90$. Such a grid equates to the sample size and computational demands of the presented `FLAPS` optimizations. A grid search does not include prior information about the optimization task. Hence, it lacks an intrinsic quality measure to rank different parameter combinations in terms of their performance. The best solutions must be manually identified from the set of evaluated grid points using independently defined quality criteria. PSO implements an intrinsic quality measure in form of the OF that is key in selecting the parameter combinations to be tested during the optimization process.

I employed my flexible OF (see Eq. 6.2) to evaluate the simulated ensemble of protein structures at each grid point. As anticipated, the grid searches yielded several acceptable or equally functional MD parameter combinations as `FLAPS` (see Table E.5). As proteins are intrinsically dynamic, I am interested in conformational ensembles rather than in single static structures. A PSO search guided by the collective experience of all particles cooperating in a swarm will always yield an overall greater proportion of usable parameter combinations, i.e., meaningful simulations and thus molecular structures, than an exhaustive grid search on a predefined set of parameter combinations. While PSO tends to remember and return to promising regions in the search space, a grid search always takes the risk of evaluating a significant number of ill-suited parameter combinations that would be dismissed on the basis of the swarm's experience.

To illustrate the advantages of a swarm-based approach over classical grid search, I considered the distributions of the simulations' median GDT, $\mathrm{GDT}^{\mathrm{med}}$, in each grid search and `FLAPS` optimization. The distributions are shown in Supplementary Figs. E.7 to E.14. For all systems, I find the `FLAPS` distribution to be shifted to higher GDTs compared to the grid search, indicating an overall greater proportion of accurate structural ensembles in `FLAPS`. In addition, I calculated the fraction of simulations with a median GDT equal to or greater than $\mathrm{GDT}_{75}^{\mathrm{max}} = 0.75 \max\left(\mathrm{GDT}^{\mathrm{med}}\right)$ for each optimization run, which was consistently larger in `FLAPS`.

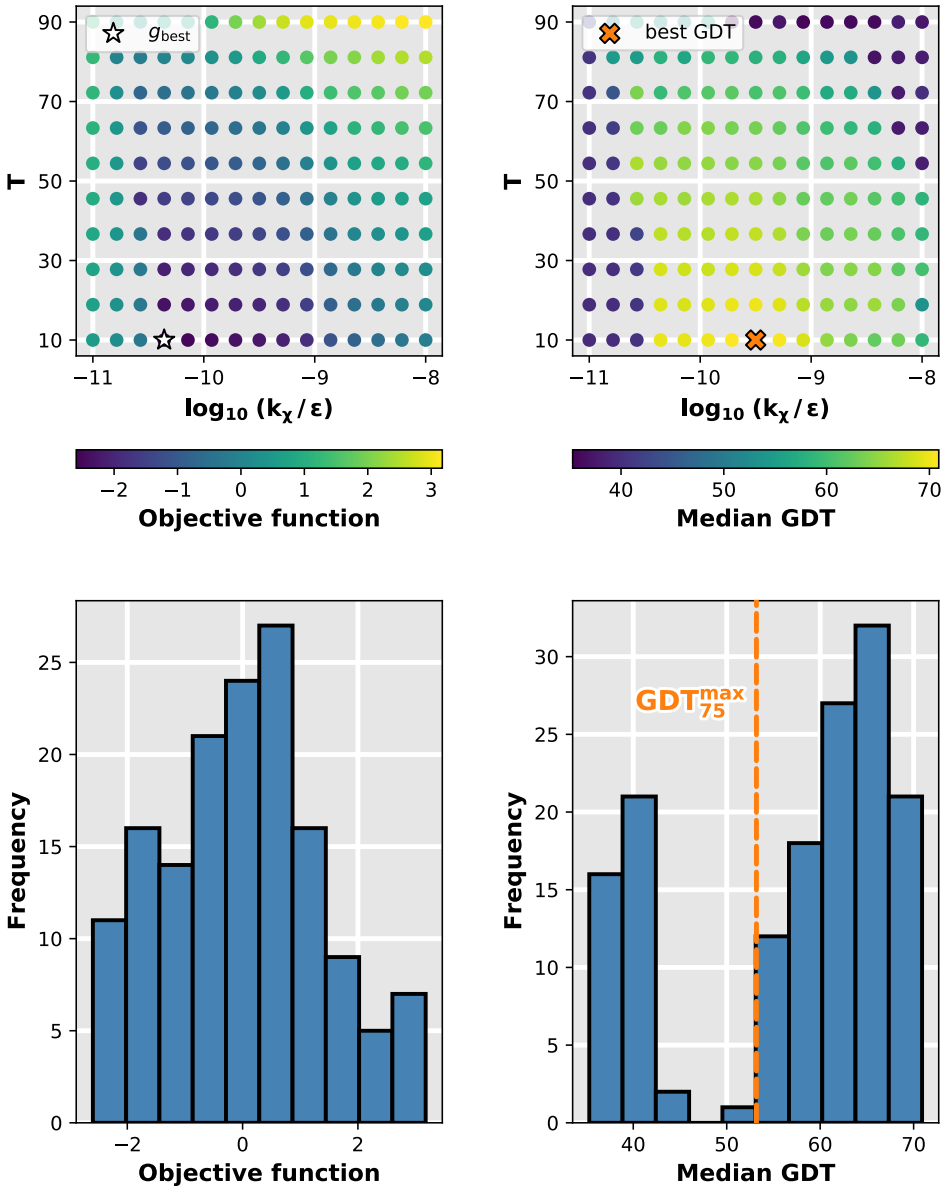


Figure E.7. Grid-search optimization for `XSBM` h → a transition of LAO protein. **Top:** Final topologies of OF (left) and median GDT (right). The global best position according to the OF, g_{best}, and the best position according to the median GDT are marked. **Bottom:** Frequency distributions of OF (left) and median GDT (right). 73 % of all simulations had a median GDT greater than $GDT_{75}^{max} = 0.75 \max\left(GDT^{med}\right)$. Reproduced from Ref.[173] under CC BY 4.0.

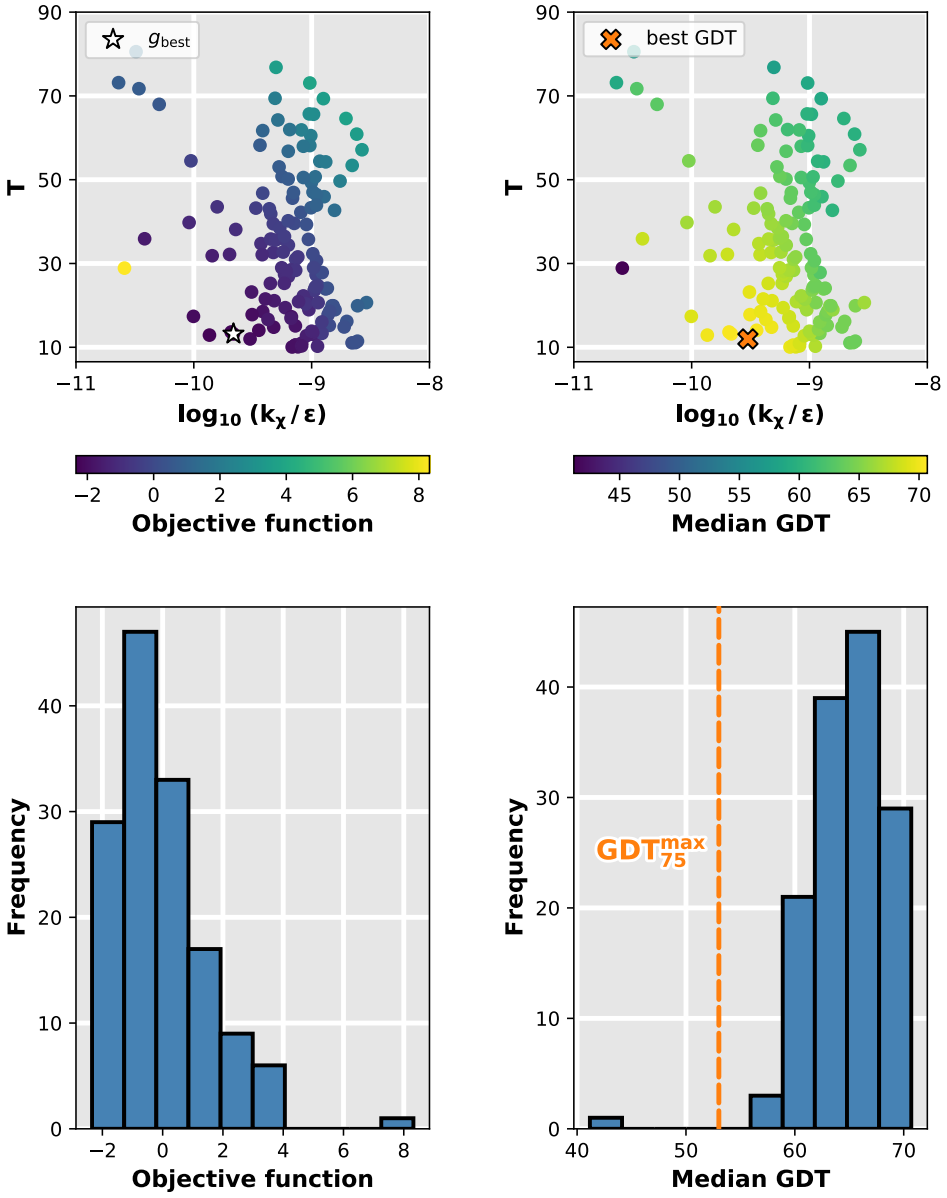


Figure E.8. `FLAPS` **optimization for** `XSBM` **h → a transition of LAO protein.** **Top:** Final topologies of OF (left) and median GDT (right). The global best position according to the OF, g_{best}, and the best position according to the median GDT are marked. **Bottom:** Frequency distributions of OF (left) and median GDT (right). 99 % of all simulations had a median GDT greater than $GDT_{75}^{max} = 0.75 \max\left(GDT^{med}\right)$. Reproduced from Ref.[173] under CC BY 4.0.

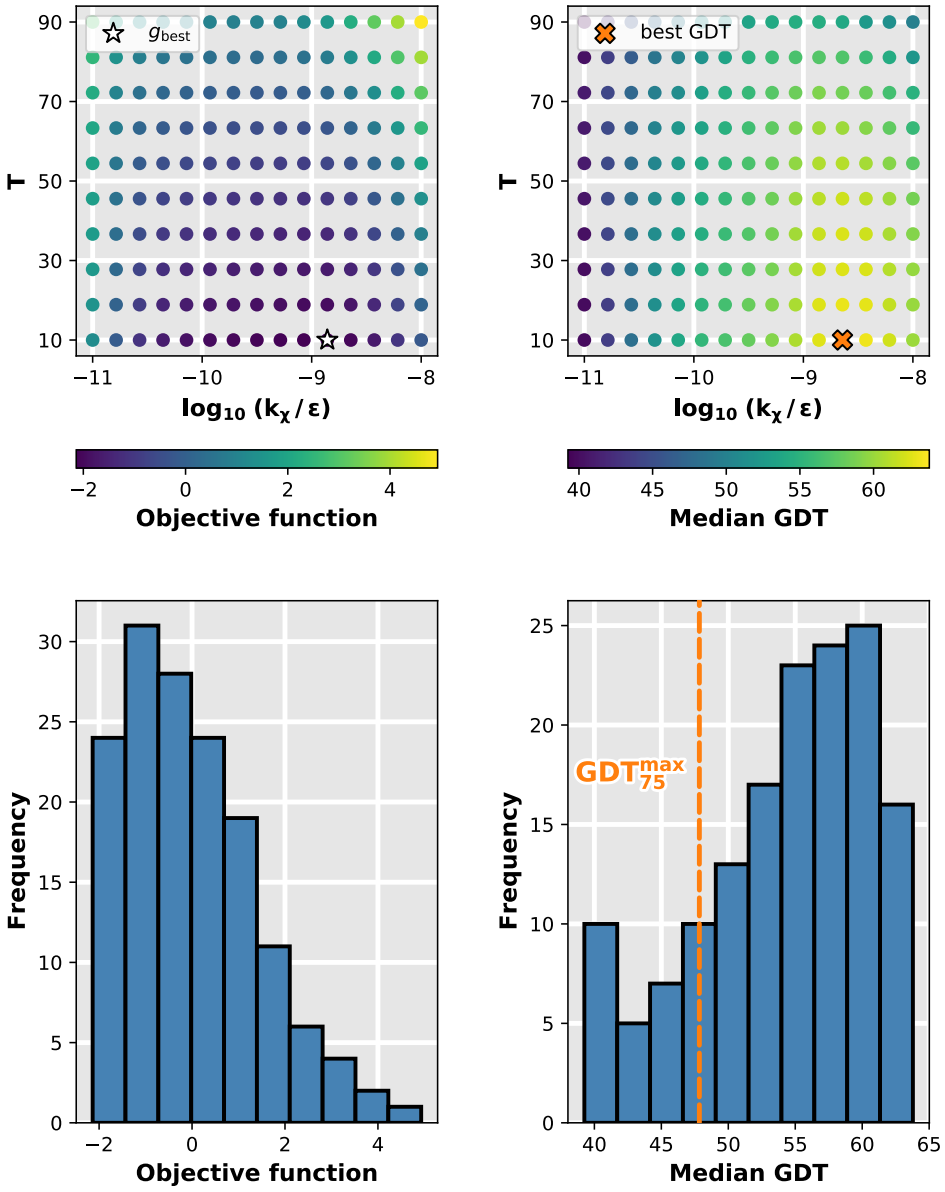


Figure E.9. Grid-search optimization for XSBM o → c transition of ADK. Top: Final topologies of OF (left) and median GDT (right). The global best position according to the OF, g_{best}, and the best position according to the median GDT are marked. **Bottom:** Frequency distributions of OF (left) and median GDT (right). 83 % of all simulations had a median GDT greater than $GDT_{75}^{max} = 0.75 \max\left(GDT^{med}\right)$. Reproduced from Ref.[173] under CC BY 4.0.

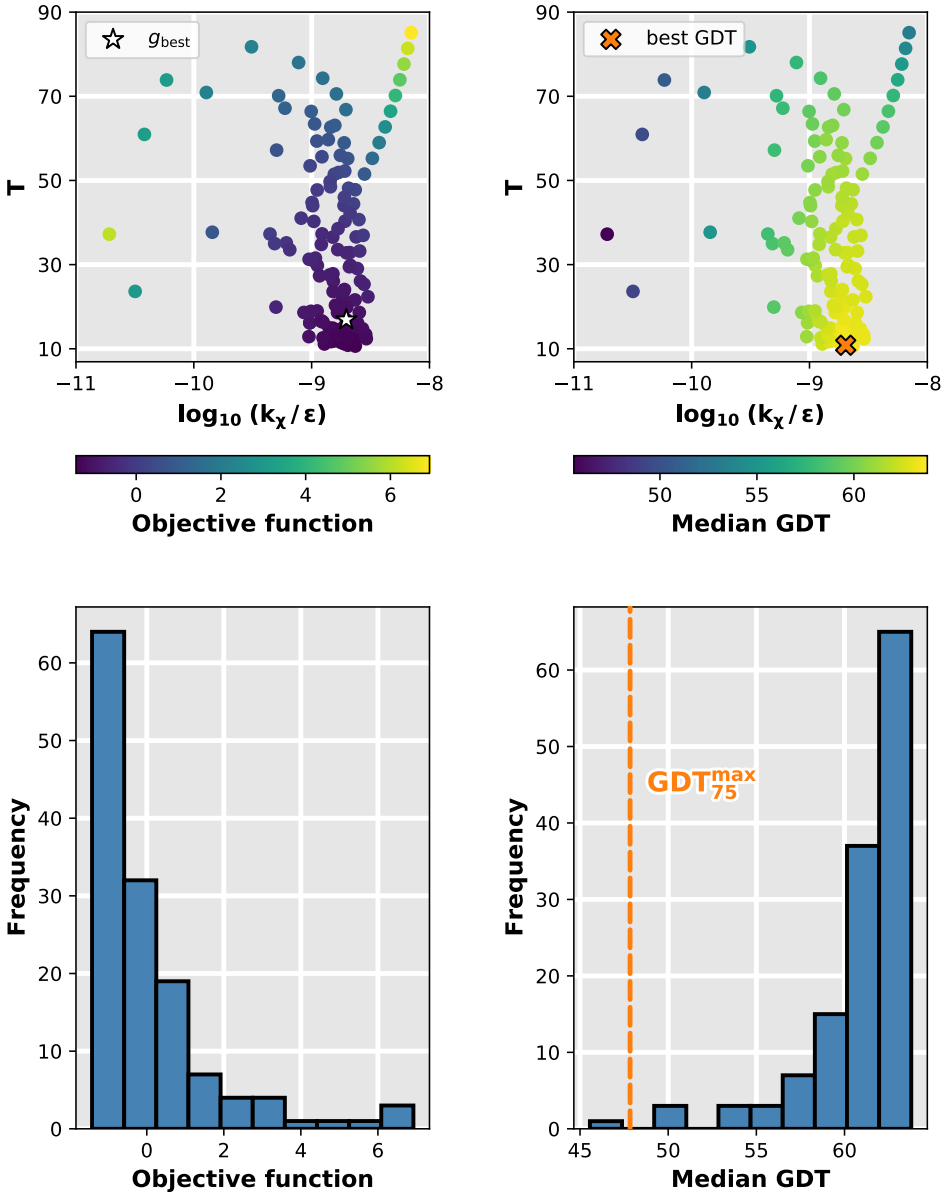


Figure E.10. `FLAPS` **optimization for** `XSBM` **o → c transition of ADK. Top:** Final topologies of OF (left) and median GDT (right). The global best position according to the OF, g_{best}, and the best position according to the median GDT are marked. **Bottom:** Frequency distributions of OF (left) and median GDT (right). 99 % of all simulations had a median GDT greater than $GDT_{75}^{max} = 0.75 \max\left(GDT^{med}\right)$. Reproduced from Ref.[173] under CC BY 4.0.

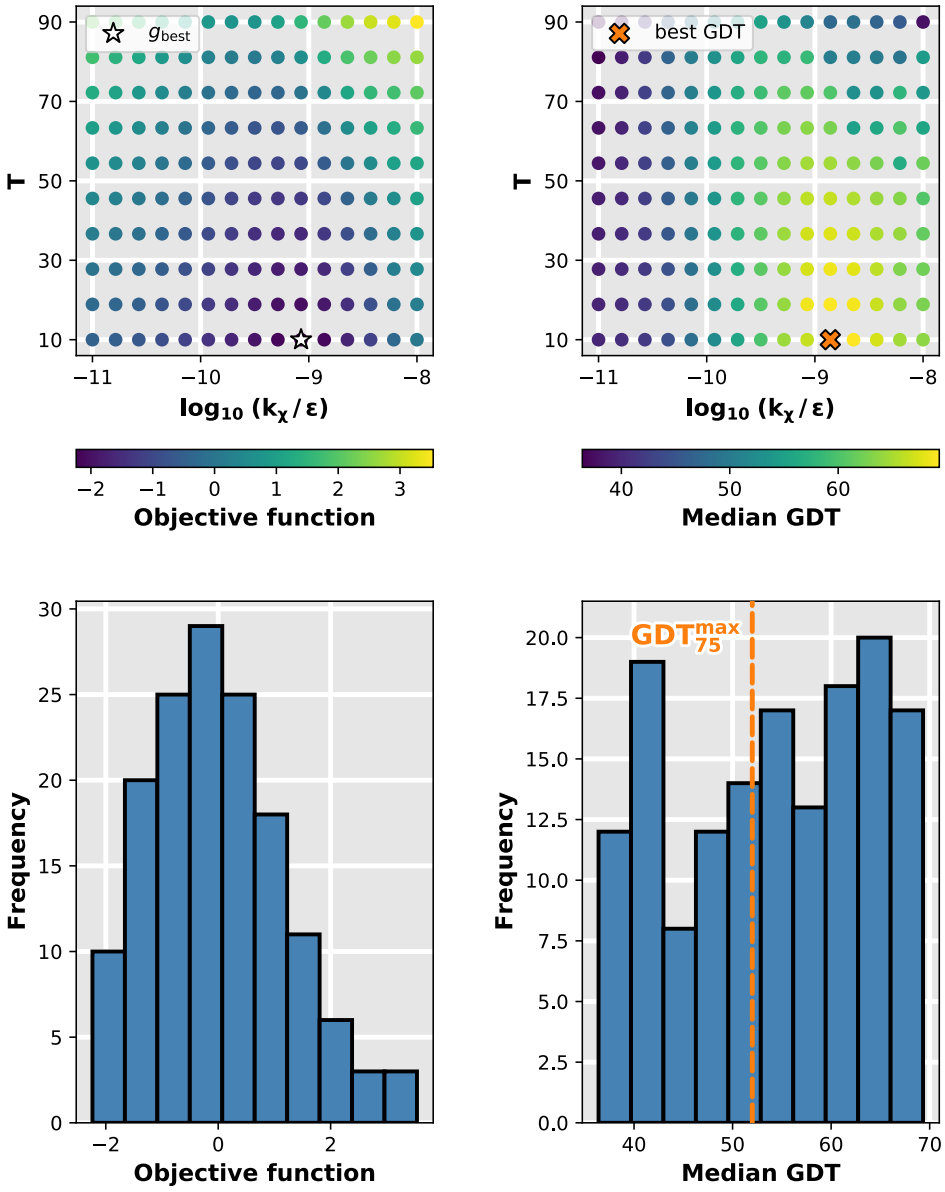


Figure E.11. Grid-search optimization for XSBM a → h transition of LAO protein. **Top:** Final topologies of OF (left) and median GDT (right). The global best position according to the OF, g_{best}, and the best position according to the median GDT are marked. **Bottom:** Frequency distributions of OF (left) and median GDT (right). 59 % of all simulations had a median GDT greater than $GDT_{75}^{max} = 0.75 \max\left(GDT^{med}\right)$. Reproduced from Ref.[173] under CC BY 4.0.

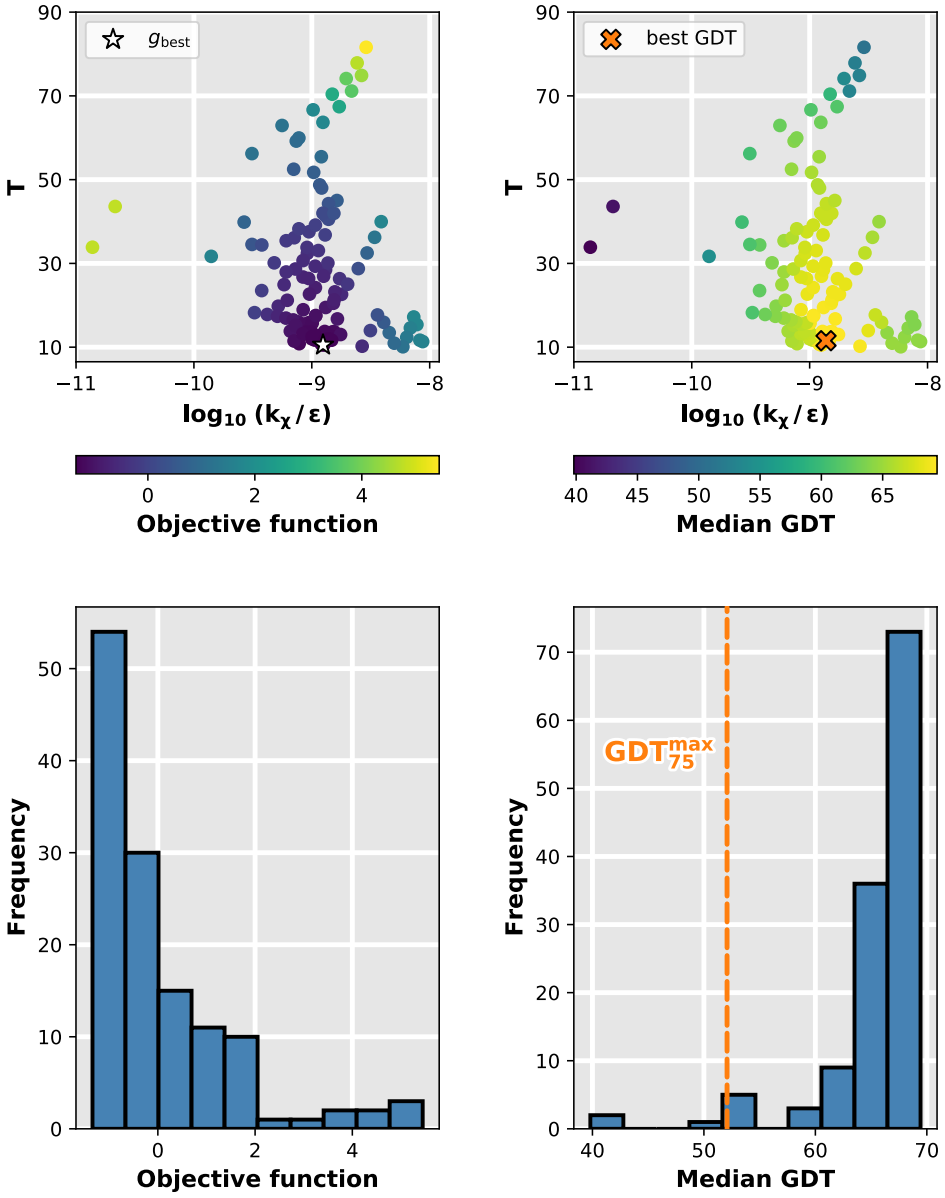


Figure E.12. `FLAPS` **optimization for** `XSBM` **a** $\rightarrow$ **h transition of LAO protein.** **Top:** Final topologies of OF (left) and median GDT (right). The global best position according to the OF, g_{best}, and the best position according to the median GDT are marked. **Bottom:** Frequency distributions of OF (left) and median GDT (right). 98 % of all simulations had a median GDT greater than $GDT_{75}^{max} = 0.75 \max\left(GDT^{med}\right)$. Reproduced from Ref.[173] under CC BY 4.0.

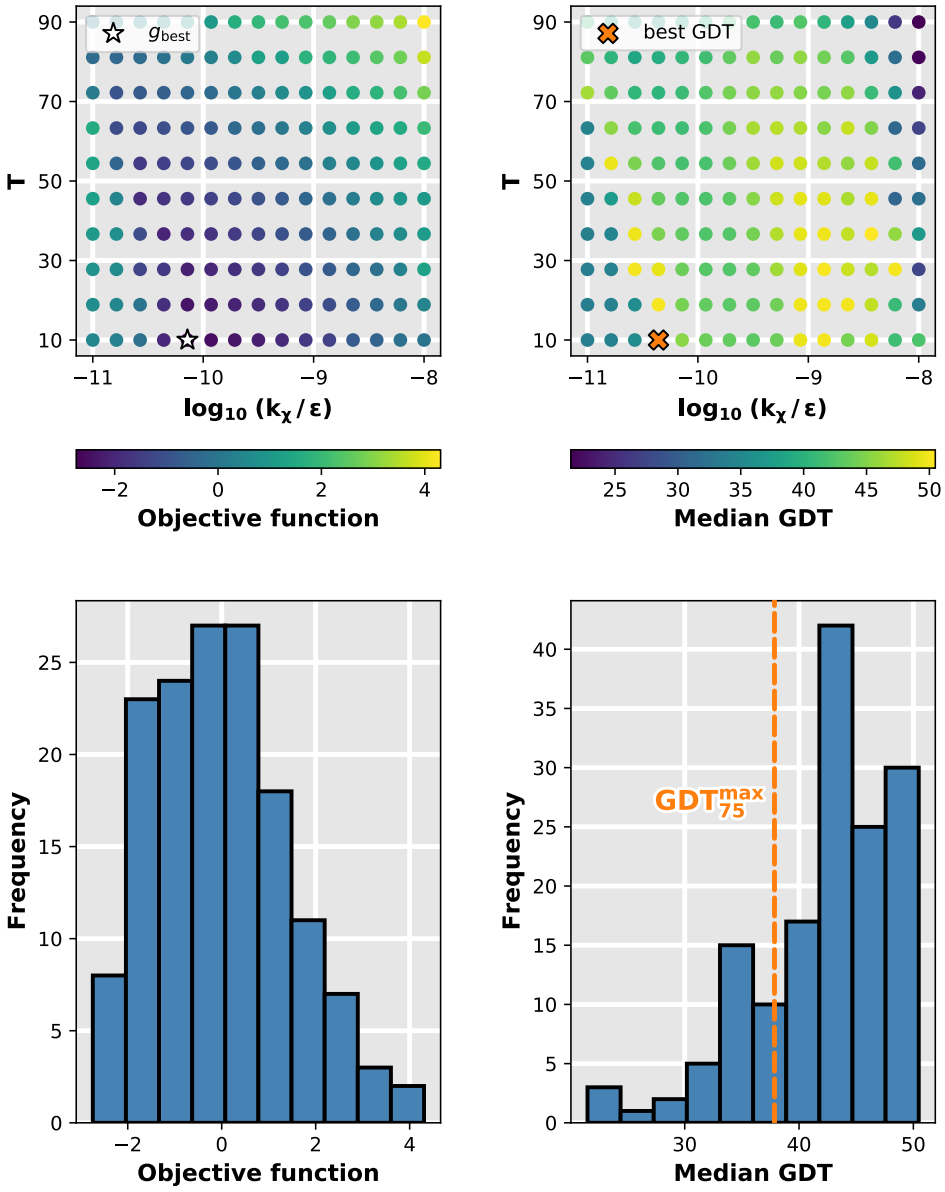


Figure E.13. Grid-search optimization for `XSBM` c → o transition of ADK. Top: Final topologies of OF (left) and median GDT (right). The global best position according to the OF, g_{best}, and the best position according to the median GDT are marked. **Bottom:** Frequency distributions of OF (left) and median GDT (right). 78 % of all simulations had a median GDT greater than $GDT_{75}^{max} = 0.75 \max\left(GDT^{med}\right)$. Reproduced from Ref.[173] under CC BY 4.0.

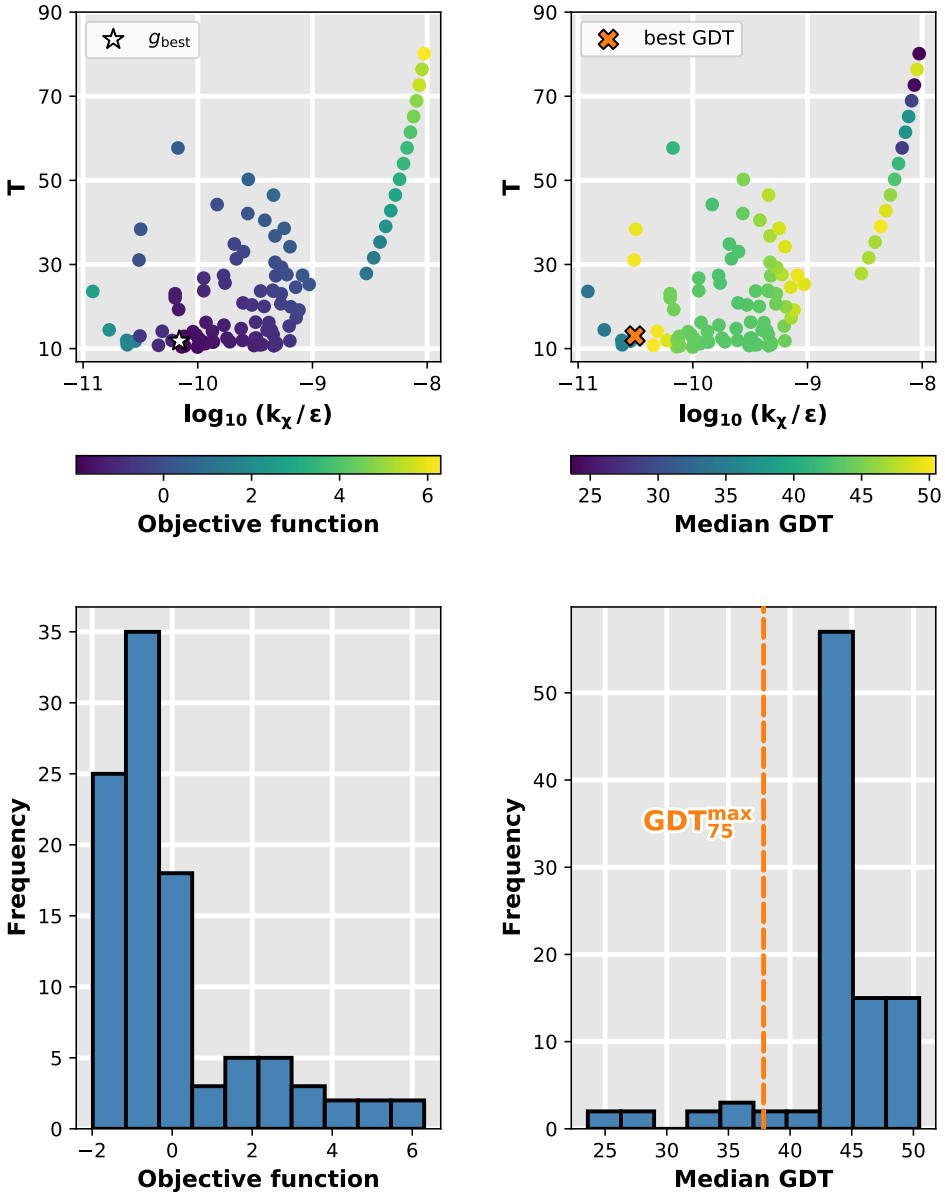


Figure E.14. `FLAPS` **optimization for** `XSBM` **c → o transition of ADK. Top:** Final topologies of OF (left) and median GDT (right). The global best position according to the OF, g_{best}, and the best position according to the median GDT are marked. **Bottom:** Frequency distributions of OF (left) and median GDT (right). 89 % of all simulations had a median GDT greater than $GDT_{75}^{max} = 0.75 \max\left(GDT^{med}\right)$. Reproduced from Ref.[173] under CC BY 4.0.

Table E.5. Grid-search optimization results[173]**.** OF, objective function f, k_χ, bias weight, T, temperature, RMSD^{med}, median RMSD, GDT^{med}, median GDT.

System	LAO h → a	ADK o → c	LAO a → h	ADK c → o
Best simulation in terms of OF				
$\min(f)$	-2.60	-2.14	-2.24	-2.74
k_χ / ε	4.394 e-11	1.390 e-09	8.483 e-10	7.197 e-11
T	10.00	10.00	10.00	10.00
RMSD^{med} / Å	2.2	3.0	2.2	4.3
GDT^{med}	69.70	62.62	66.49	45.44
Best simulation in terms of RMSD^{med}				
RMSD^{med} / Å	2.1	2.6	2.0	3.5
$f(\text{RMSD}^{\text{med}})$	-2.24	-1.66	-2.11	-1.36
k_χ / ε	3.162 e-10	3.728 e-09	1.390 e-09	1.390 e-09
T	10.00	10.00	10.00	10.00
GDT^{med}	70.91	63.20	69.33	50.00
Best simulation in terms of GDT^{med}				
GDT^{med}	70.91	63.78	69.33	50.47
$f(\text{GDT}^{\text{med}})$	-2.24	-2.03	-2.11	-1.97
k_χ / ε	3.162 e-10	2.276 e-09	1.390 e-09	4.394 e-11
T	10.00	10.00	10.00	10.00
RMSD^{med} / Å	2.1	2.7	2.0	4.9

E.4 Implementation

"The last good thing written in C was
Franz Schubert's Symphony Number 9."

Erwin Dieterich

`FLAPS` is implemented as a stand-alone solver in `Hyppopy`[1], a Python-based hyperparameter optimization package. `Hyppopy` provides various tools for black-box optimization. It has a simple, unified API which can be used to access a collection of solver libraries. My implementation of `FLAPS`[213,214] is available on GitHub[2,3]. I implemented an MPI-parallel version of the code using a sophisticated parallelization architecture described in Supplementary Fig. E.15. Available compute nodes comprising a given number of processors are divided into blocks. Each block corresponds to one particle in the swarm. Within one block, the simulation itself runs on a single core while all the others process the generated frames in the trajectory on the fly. This results in a massive reduction in runtime.

The experiments were run on the `ForHLR II` cluster system located at the Steinbuch Centre for Computing at Karlsruhe Institute of Technology. The system comprises 1152 purely CPU-based compute nodes. Each node is equipped with two 10-core Intel Xeon E5-2660 v3 Haswell CPUs at 3.3 GHz, 64 GB of DDR3 main memory and 4x Mellanox 100 Gbit EDR InfiniBand links. The software packages used were a RHEL Linux with kernel version 4.18.0 and Python 3.6.8. Each run used 51 compute nodes (overall 1020 cores). Due to the heavy I/O workload and many metadata operations, I used a private on-demand file system (BeeGFS On-Demand) with a stripe count of 1, where one node was reserved for its metadata server[215]. Each block in the underlying simulator-worker scheme consisted of five nodes, i.e., 100 cores (1 simulator, 99 workers). Each run cost approximately 40 000 CPU hours. For the

[1] https://github.com/MIC-DKFZ/Hyppopy
[2] https://github.com/FLAPS-NMI/FLAPS-Hyppopy/releases/tag/v1.0
[3] https://github.com/FLAPS-NMI/FLAPS-optunity/releases/tag/v1.0

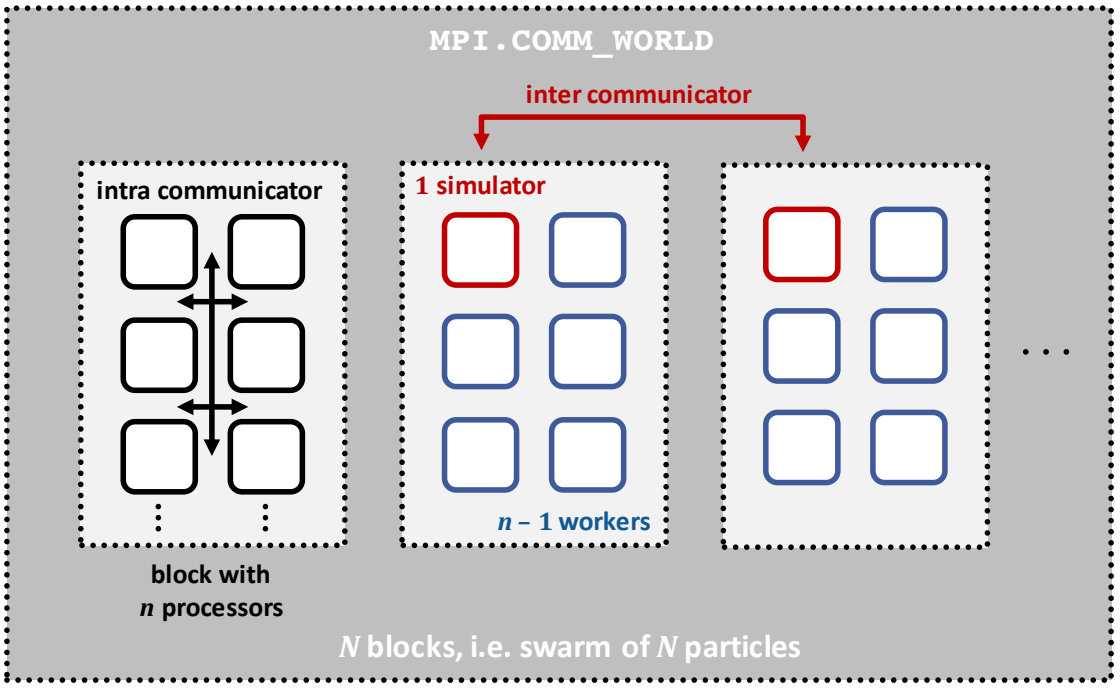


Figure E.15. Simulator-worker parallelization scheme used for `FLAPS` in `Hyppopy`. Reproduced from Ref.[173] under CC BY 4.0.

application of `FLAPS` to the optimization of `XSBM` parameters, I used a cognitive acceleration coefficient $\phi_1 = 2.0$ and a social acceleration coefficient $\phi_2 = 1.5$ in the particle update (see Algorithm 2). The complete setup[203] including all PSO hyperparameters used is available on Github[4].

[4] https://github.com/FLAPS-NMI/FLAPS-sim_setups/releases/tag/v1.0

Figure E.16. The sweet BeeGFS bee. Summ, summ, summ! Bienchen summ' herum ☺!

Bibliography

[1] Schrödinger, LLC, "The PyMOL Molecular Graphics System, Version 1.8," 2015.

[2] D. J. Selkoe, "Cell biology of protein misfolding: The examples of Alzheimer's and Parkinson's diseases," *Nature Cell Biology*, vol. 6, no. 11, pp. 1054–1061, 2004. doi: 10.1038/ncb1104-1054.

[3] C. Soto, "Unfolding the role of protein misfolding in neurodegenerative diseases," *Nature Reviews Neuroscience*, vol. 4, no. 1, pp. 49–60, 2003. doi: 10.1038/nrn1007.

[4] A. Mukherjee, D. Morales-Scheihing, P. C. Butler, and C. Soto, "Type 2 diabetes as a protein misfolding disease," *Trends in Molecular Medicine*, vol. 21, no. 7, pp. 439–449, 2015. doi: 10.1016/j.molmed.2015.04.005.

[5] C. M. Dobson, "Protein-misfolding diseases: Getting out of shape," *Nature*, vol. 418, no. 6899, pp. 729–730, 2002. doi: 10.1038/418729a.

[6] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," *Nucleic Acids Research*, vol. 28, no. 1, pp. 235–242, 2000. doi: 10.1093/nar/28.1.235.

[7] "The 2017 Nobel Prize in Chemistry - Press Release," 2017.

[8] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, K. Tunyasuvunakool, O. Ronneberger, R. Bates, A. Zidek, A. Bridgland, *et al.*, "High accuracy protein structure prediction using deep learning," *Fourteenth Critical Assessment of Techniques for Protein Structure Prediction (Abstract Book)*, vol. 22, p. 24, 2020.

[9] D. I. Svergun and M. H. J. Koch, "Small-angle scattering studies of biological macromolecules in solution," *Reports on Progress in Physics*, vol. 66, no. 10, pp. 1735–1782, 2003. doi: 10.1088/0034-4885/66/10/R05.

[10] L. Boldon, F. Laliberte, and L. Liu, "Review of the fundamental theories behind small angle X-ray scattering, molecular dynamics simulations, and relevant integrated application.," *Nano Reviews*, vol. 6, p. 25661, 2015. doi: 10.3402/nano.v6.25661.

[11] I. Reinartz, C. Sinner, D. Nettels, B. Stucki-Buchli, F. Stockmar, P. T. Panek, C. R. Jacob, G. U. Nienhaus, B. Schuler, and A. Schug, "Simulation of FRET dyes allows quantitative comparison against experimental data," *The Journal of Chemical Physics*, vol. 148, no. 12, p. 123321, 2018. doi: 10.1063/1.5010434.

[12] P.-c. Chen and J. S. Hub, "Interpretation of Solution X-Ray Scattering by Explicit-Solvent Molecular Dynamics," *Biophysical Journal*, vol. 108, no. 10, pp. 2573–2584, 2015. doi: 10.1016/j.bpj.2015.03.062.

[13] P. C. Whitford, A. Ahmed, Y. Yu, S. P. Hennelly, F. Tama, C. M. Spahn, J. N. Onuchic, and K. Y. Sanbonmatsu, "Excited states of ribosome translocation revealed through integrative molecular modeling," *Proceedings of the National Academy of Sciences*, vol. 108, no. 47, pp. 18943–18948, 2011. doi: 10.1073/pnas.1108363108.

[14] L. G. Trabuco, E. Villa, E. Schreiner, C. B. Harrison, and K. Schulten, "Molecular dynamics flexible fitting: A practical guide to combine cryo-electron microscopy and X-ray crystallography," *Methods*, vol. 49, no. 2, pp. 174–180, 2009. doi: 10.1016/j.ymeth.2009.04.005.

[15] A. T. Brünger, J. Kuriyan, and M. Karplus, "Crystallographic R Factor Refinement by Molecular Dynamics," *Science*, vol. 235, no. 4787, pp. 458–460, 1987. doi: 10.1126/science.235.4787.458.

[16] M. R. Hermann and J. S. Hub, "SAXS-Restrained Ensemble Simulations of Intrinsically Disordered Proteins with Commitment to the Principle of Maximum Entropy," *Journal of Chemical Theory and Computation*, vol. 15, no. 9, pp. 5103–5115, 2019. doi: 10.1021/acs.jctc.9b00338.

[17] P.-c. Chen, R. Shevchuk, F. M. Strnad, C. Lorenz, L. Karge, R. Gilles, A. M. Stadler, J. Hennig, and J. S. Hub, "Combined Small-Angle X-ray and Neutron Scattering Restraints in Molecular Dynamics Simulations," *Journal of Chemical Theory and Computation*, vol. 15, no. 8, pp. 4687–4698, 2019. doi: 10.1021/acs.jctc.9b00292.

[18] A. Björling, S. Niebling, M. Marcellini, D. van der Spoel, and S. Westenhoff, "Deciphering Solution Scattering Data with Experimentally Guided Molecular Dynamics Simulations," *Journal of Chemical Theory and Computation*, vol. 11, no. 2, pp. 780–787, 2015. doi: 10.1021/ct5009735.

[19] P. E. Leopold, M. Montal, and J. N. Onuchic, "Protein folding funnels: a kinetic approach to the sequence-structure relationship.," *Proceedings of the National Academy of Sciences*, vol. 89, no. 18, pp. 8721–8725, 1992. doi: 10.1073/pnas.89.18.8721.

[20] H. Frauenfelder, S. G. Sligar, and P. G. Wolynes, "The Energy Landscapes and Motions of Proteins," *Science*, vol. 254, no. 9, pp. 1598–1603, 1991. doi: 10.1126/science.1749933.

[21] J. K. Noel, P. C. Whitford, and J. N. Onuchic, "The shadow map: A general contact definition for capturing the dynamics of biomolecular folding and function," *The Journal of Physical Chemistry B*, vol. 116, no. 29, pp. 8692–8702, 2012. doi: 10.1021/jp300852d.

[22] P. C. Whitford, K. Y. Sanbonmatsu, and J. N. Onuchic, "Biomolecular dynamics: order-disorder transitions and energy landscapes," *Reports on Progress in Physics*, vol. 75, no. 7, p. 076601, 2012. doi: 10.1021/jp300852d.

[23] A. Schug and J. N. Onuchic, "From protein folding to protein function and biomolecular binding by energy landscape theory," *Current Opinion in Pharmacology*, vol. 10, no. 6, pp. 709–714, 2010. doi: 10.1016/j.coph.2010.09.012.

[24] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, "Funnels, pathways, and the energy landscape of protein folding: A synthesis," *Proteins: Structure, Function, and Bioinformatics*, vol. 21, no. 3, pp. 167–195, 1995. doi: 10.1002/prot.340210302.

[25] J. N. Onuchic and P. G. Wolynes, "Theory of protein folding," *Current Opinion in Structural Biology*, vol. 14, no. 1, pp. 70–75, 2004. doi: 10.1016/j.sbi.2004.01.009.

[26] J. D. Bryngelson and P. G. Wolynes, "Spin glasses and the statistical mechanics of protein folding," *Proceedings of the National Academy of Sciences*, vol. 84, no. 21, pp. 7524–7528, 1987. doi: 10.1073/pnas.84.21.7524.

[27] M. Weiel, I. Reinartz, and A. Schug, "Rapid interpretation of small-angle X-ray scattering data," *PLoS Computational Biology*, vol. 15, no. 3, p. e1006900, 2019. doi: 10.1371/journal.pcbi.1006900.

[28] A. Christiansen, M. Weiel, A. Winkler, A. Schug, and J. Reinstein, "The Trimeric Major Capsid Protein of Mavirus is stabilized by its Interlocked N-termini Enabling Core Flexibility for Capsid Assembly," *Journal of Molecular Biology*, vol. 433, no. 7, p. 166859, 2021. doi: 10.1016/j.jmb.2021.166859.

[29] I. Reinartz, M. Weiel, and A. Schug, "FRET Dyes Significantly Affect SAXS Intensities of Proteins," *Israel Journal of Chemistry*, vol. 60, no. 7, pp. 725–734, 2020. doi: 10.1002/ijch.202000007.

[30] C. B. Anfinsen, "Principles that Govern the Folding of Protein Chains," *Science*, vol. 181, no. 4096, pp. 223–230, 1973. url: https://www.jstor.org/stable/1736447.

[31] C. Levinthal, "How to fold graciously," *Mossbauer spectroscopy in biological systems*, pp. 22–24, 1969.

[32] B. Schuler and H. Hofmann, "Single-molecule spectroscopy of protein folding dynamics – expanding scope and timescales," *Current Opinion in Structural Biology*, vol. 23, no. 1, pp. 36–47, 2013. doi: 10.1016/j.sbi.2012.10.008.

[33] R. A. Goldbeck, Y. G. Thomas, E. Chen, R. M. Esquerra, and D. S. Kliger, "Multiple pathways on a protein-folding energy landscape: Kinetic evidence," *Proceedings of the National Academy of Sciences*, vol. 96, no. 6, pp. 2782–2787, 1999. doi: 10.1073/pnas.96.6.2782.

[34] J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes, "THEORY OF PROTEIN FOLDING: The Energy Landscape Perspective," *Annual Review of Physical Chemistry*, vol. 48, no. 1, pp. 545–600, 1997. doi: 10.1146/annurev.physchem.48.1.545.

[35] K. A. Dill and H. S. Chan, "From Levinthal to pathways to funnels," *Nature Structural Biology*, vol. 4, no. 1, pp. 10–19, 1997. doi: 10.1038/nsb0197-10.

[36] H. D. T. Mertens and D. I. Svergun, "Structural characterization of proteins and complexes using small-angle X-ray solution scattering," *Journal of Structural Biology*, vol. 172, no. 1, pp. 128–141, 2010. doi: 10.1016/j.jsb.2010.06.012.

[37] D. I. Svergun, "Restoring Low Resolution Structure of Biological Macromolecules from Solution Scattering Using Simulated Annealing," *Biophysical Journal*, vol. 76, no. 6, pp. 2879–2886, 1999. doi: 10.1016/S0006-3495(99)77443-6.

[38] C. D. Putnam, M. Hammel, G. L. Hura, and J. A. Tainer, "X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution," *Quarterly Reviews of Biophysics*, vol. 40, no. 03, pp. 191–285, 2007. doi: 10.1017/S0033583507004635.

[39] P. Debye, "Zerstreuung von Röntgenstrahlen," *Annalen der Physik*, vol. 351, no. 6, pp. 809–823, 1915. doi: 10.1002/andp.19153510606.

[40] A. Guinier, G. Fournet, and K. L. Yudowitch, "Small-angle scattering of X-rays," 1955.

[41] D. I. Svergun, C. Barberato, and M. H. J. Koch, "CRYSOL – a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates," *Journal of Applied Crystallography*, vol. 28, no. 6, pp. 768–773, 1995. doi: 10.1107/S0021889895007047.

[42] C. E. Shannon, "The mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.

[43] T. Förster, "Zwischenmolekulare Energiewanderung und Fluoreszenz," *Annalen der Physik*, vol. 437, pp. 55–75, 1948. doi: 10.1002/andp.19484370105.

[44] L. Stryer, "Fluorescence Energy Transfer as a Spectroscopic Ruler," *Annual Review of Biochemistry*, vol. 47, no. 1, pp. 819–846, 1978. doi: 10.1146/annurev.bi.47.070178.004131.

[45] R. Rieger, A. Kobitski, H. Sielaff, and G. U. Nienhaus, "Evidence of a Folding Intermediate in RNase H from Single-Molecule FRET Experiments," *ChemPhysChem*, vol. 12, no. 3, pp. 627–633, 2011. doi: 10.1002/cphc.201000693.

[46] R. Rieger and G. U. Nienhaus, "A combined single-molecule FRET and tryptophan fluorescence study of RNase H folding under acidic conditions," *Chemical Physics*, vol. 396, no. 1, pp. 3–9, 2012. doi: 10.1016/j.chemphys.2011.03.026.

[47] Y. Gambin, A. Schug, E. A. Lemke, J. J. Lavinder, A. C. Ferreon, T. J. Magliery, J. N. Onuchic, and A. A. Deniz, "Direct single-molecule observation of a protein living in two opposed native structures," *Proceedings of the National Academy of Sciences*, vol. 106, no. 25, pp. 10153–10158, 2009. doi: 10.1073/pnas.0904461106.

[48] E. Sisamakis, A. Valeri, S. Kalinin, P. J. Rothwell, and C. A. M. Seidel, "Chapter 18 - Accurate Single-Molecule FRET Studies Using Multiparameter Fluorescence Detection," in *Single Molecule Tools, Part B: Super-Resolution, Particle Tracking, Multiparameter, and Force Based Methods*, vol. 475 of *Methods in Enzymology*, pp. 455 – 514, Academic Press, 2010. doi: 10.1016/S0076-6879(10)75018-7.

[49] N. J. Greenfield, "Using circular dichroism spectra to estimate protein secondary structure," *Nature Protocols*, vol. 1, no. 6, p. 2876, 2006. doi: 10.1038/nprot.2006.202.

[50] L. Whitmore and B. A. Wallace, "Protein secondary structure analyses from circular dichroism spectroscopy: Methods and reference databases," *Biopolymers: Original Research on Biomolecules*, vol. 89, no. 5, pp. 392–400, 2008. doi: 10.1002/bip.20853.

[51] K. R. Karch, M. Coradin, L. Zandarashvili, Z.-Y. Kan, M. Gerace, S. W. Englander, B. E. Black, and B. A. Garcia, "Hydrogen-Deuterium Exchange Coupled to Top- and Middle-Down Mass Spectrometry Reveals Histone Tail Dynamics before and after Nucleosome Assembly," *Structure*, vol. 26, no. 12, pp. 1651–1663, 2018. doi: 10.1016/j.str.2018.08.006.

[52] M. Abraham, B. Hess, D. van der Spoel, and E. Lindahl, "GROMACS User Manual Version 5.0.7," 2015. doi: 10.1007/SpringerReference_28001.

[53] H. J. C. Berendsen, J. P. M. v. Postma, W. F. van Gunsteren, A. R. H. J. DiNola, and J. R. Haak, "Molecular dynamics with coupling to an external bath," *The Journal of Chemical Physics*, vol. 81, no. 8, pp. 3684–3690, 1984. doi: 10.1063/1.448118.

[54] H. Bekker, H. J. C. Berendsen, E. J. Dijkstra, S. Achertop, R. van Drumen, D. van der Spoel, A. Sijbers, H. Keegstra, B. Reitsma, and M. K. R. Renardus, "Gromacs: A parallel computer for molecular dynamics simulations," *Physics Computing*, vol. 92, 1993.

[55] H. J. C. Berendsen, D. van der Spoel, and R. van Drunen, "GROMACS: A message-passing parallel molecular dynamics implementation," *Computer Physics Communications*, vol. 91, no. 1-3, pp. 43–56, 1995. doi: 10.1016/0010-4655(95)00042-E.

[56] E. Lindahl, B. Hess, and D. van der Spoel, "GROMACS 3.0: a package for molecular simulation and trajectory analysis," *Molecular Modeling Annual*, vol. 7, no. 8, pp. 306–317, 2001. doi: 10.1007/s008940100045.

[57] D. van der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen, "GROMACS: Fast, flexible, and free," *Journal of Computational Chemistry*, vol. 26, no. 16, pp. 1701–1718, 2005. doi: 10.1002/jcc.20291.

[58] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, "GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation," *Journal of Chemical Theory and Computation*, vol. 4, no. 3, pp. 435–447, 2008. doi: 10.1021/ct700301q.

[59] S. Pronk, S. Páll, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. van der Spoel, B. Hess, and E. Lindahl, "GROMACS 4.5: A high-throughput and highly parallel open source molecular simulation toolkit," *Bioinformatics*, vol. 29, no. 7, pp. 845–854, 2013. doi: 10.1093/bioinformatics/btt055.

[60] S. Páll, M. J. Abraham, C. Kutzner, B. Hess, and E. Lindahl, *Tackling exascale software challenges in molecular dynamics simulations with GROMACS*, vol. 8759. 2015. doi: 10.1007/978-3-319-15976-8_1.

[61] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl, "Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers," *SoftwareX*, vol. 1-2, pp. 19–25, 2015. doi: 10.1016/j.softx.2015.06.001.

[62] M. Abraham, E. Apol, R. Apostolov, H. Berendsen, A. van Buuren, P. Bjelkmar, R. van Drunen, A. Feenstra, S. Fritsch, G. Groenhof, C. Junghans, J. Hub, P. Kasson, C. Kutzner, B. Lambeth, P. Larsson, J. Lemkul, E. Marklund, P. Meulenhoff, T. Murtola, S. Pall, S. Pronk, R. Schulz, M. Shirts, A. Sijbers, P. Tieleman, and M. Wolf, "GROMACS Manual," 2012. url: https://www.gromacs.org.

[63] W. F. van Gunsteren and H. J. C. Berendsen, "A leap-frog algorithm for stochastic dynamics," *Molecular Simulation*, vol. 1, no. 3, pp. 173–185, 1988. doi: 10.1080/08927028808080941.

[64] D. Frenkel and B. Smit, *Understanding molecular simulation: from algorithms to applications*, vol. 1. Elsevier, 2001.

[65] H. Lammert, A. Schug, and J. N. Onuchic, "Robustness and generalization of structure-based models for protein folding and function," *Proteins: Structure, Function, and Bioinformatics*, vol. 77, no. 4, pp. 881–891, 2009. doi: 10.1002/prot.22511.

[66] C. Clementi, H. Nymeyer, and J. N. Onuchic, "Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? an investigation for small globular proteins," *Journal of Molecular Biology*, vol. 298, no. 5, pp. 937–953, 2000. doi: 10.1006/jmbi.2000.3693.

[67] M. S. Cheung, J. M. Finke, B. Callahan, and J. N. Onuchic, "Exploring the Interplay between Topology and Secondary Structural Formation in the Protein Folding Problem," *The Journal of Physical Chemistry B*, vol. 107, no. 40, pp. 11193–11200, 2003. doi: 10.1021/jp034441r.

[68] L. L. Chavez, J. N. Onuchic, and C. Clementi, "Quantifying the roughness on the free energy landscape: entropic bottlenecks and protein folding rates.," *Journal of the American Chemical Society*, vol. 126, no. 27, pp. 8426–32, 2004. doi: 10.1021/ja049510+.

[69] M. F. Rey-Stolle, M. Enciso, and A. Rey, "Topology-based models and NMR structures in protein folding simulations," *Journal of Computational Chemistry*, vol. 30, no. 8, pp. 1212–1219, 2009. doi: 10.1002/jcc.21149.

[70] B. Lutz, C. Sinner, G. Heuermann, A. Verma, and A. Schug, "eSBMTools 1.0: Enhanced native structure-based modeling tools," *Bioinformatics*, vol. 29, no. 21, pp. 2795–2796, 2013. doi: 10.1093/bioinformatics/btt478.

[71] C. Sinner, B. Lutz, S. John, I. Reinartz, A. Verma, and A. Schug, "Simulating Biomolecular Folding and Function by Native-Structure-Based/Go-Type Models," *Israel Journal of Chemistry*, vol. 54, no. 8-9, pp. 1165–1175, 2014. doi: 10.1002/ijch.201400012.

[72] J. I. Sułkowska, J. K. Noel, and J. N. Onuchic, "Energy landscape of knotted protein folding," *Proceedings of the National Academy of Sciences*, vol. 109, no. 44, pp. 17783–17788, 2012. doi: 10.1073/pnas.1201804109.

[73] Y. Levy and J. N. Onuchic, "Mechanisms of Protein Assembly: Lessons from Minimalist Models," *Accounts of Chemical Research*, vol. 39, no. 2, pp. 135–142, 2006. doi: 10.1021/ar040204a.

[74] A. Schug, M. Weigt, J. N. Onuchic, T. Hwa, and H. Szurmant, "High-resolution protein complexes from integrating genomic information with molecular simulation," *Proceedings of the National Academy of Sciences*, vol. 106, no. 52, pp. 22124–22129, 2009. doi: 10.1073/pnas.0912100106.

[75] A. E. Dago, A. Schug, A. Procaccini, J. A. Hoch, M. Weigt, and H. Szurmant, "Structural basis of histidine kinase autophosphorylation deduced by integrating genomics, molecular dynamics, and mutagenesis," *Proceedings of the National Academy of Sciences*, vol. 109, no. 26, pp. E1733–E1742, 2012. doi: 10.1073/pnas.1201301109.

[76] E. De Leonardis, B. Lutz, S. Ratz, S. Cocco, R. Monasson, A. Schug, and M. Weigt, "Direct-Coupling Analysis of nucleotide coevolution facilitates RNA secondary and tertiary structure prediction," *Nucleic Acids Research*, vol. 43, no. 21, pp. 10444–10455, 2015. doi: 10.1093/nar/gkv932.

[77] J. I. Sułkowska, F. Morcos, M. Weigt, T. Hwa, and J. N. Onuchic, "Genomics-aided structure prediction," *Proceedings of the National Academy of Sciences*, vol. 109, no. 26, pp. 10340–10345, 2012. doi: 10.1073/pnas.1207864109.

[78] C. Clementi, P. A. Jennings, and J. N. Onuchic, "Prediction of folding mechanism for circular-permuted proteins," *Journal of Molecular Biology*, vol. 311, no. 4, pp. 879–890, 2001. doi: 10.1006/jmbi.2001.4871.

[79] P. C. Whitford, J. K. Noel, S. Gosavi, A. Schug, K. Y. Sanbonmatsu, and J. N. Onuchic, "An all-atom structure-based potential for proteins: Bridging minimal models with all-atom empirical forcefields," *Proteins*, vol. 23, no. 1, pp. 1–7, 2009. doi: 10.1002/prot.22253.

[80] J. K. Noel, P. C. Whitford, K. Y. Sanbonmatsu, and J. N. Onuchic, "SMOG@ctbp: Simplified deployment of structure-based models in GROMACS," *Nucleic Acids Research*, vol. 38, no. SUPPL. 2, pp. 657–661, 2010. doi: 10.1093/nar/gkq498.

[81] B.-H. Oh, J. Pandit, C.-H. Kang, K. Nikaido, S. Gokcen, G. F. Ames, and S.-H. Kim, "Three-dimensional structures of the periplasmic lysine/arginine/ornithine-binding protein with and without a ligand.," *Journal of Biological Chemistry*, vol. 268, no. 15, pp. 11348–11355, 1993. doi: 10.2210/PDB1LST/PDB.

[82] A. Zemla, "LGA: a method for finding 3D similarities in protein structures," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3370–3374, 2003. doi: 10.1093/nar/gkg571.

[83] I. Kufareva and R. Abagyan, "Methods of Protein Structure Comparison," in *Homology Modeling*, pp. 231–257, Springer, 2011. doi: 10.1007/978-1-61779-588-6_10.

[84] V. Modi, Q. Xu, S. Adhikari, and R. L. Dunbrack Jr, "Assessment of template-based modeling of protein structure in CASP11," *Proteins: Structure, Function, and Bioinformatics*, vol. 84, pp. 200–220, 2016. doi: 10.1002/prot.25049.

[85] J. Moult, K. Fidelis, A. Kryshtafovych, T. Schwede, and A. Tramontano, "Critical assessment of methods of protein structure prediction (CASP) – Round XII," *Proteins: Structure, Function, and Bioinformatics*, vol. 86, pp. 7–15, 2018. doi: 10.1002/prot.25415.

[86] M. G. Fischer and C. A. Suttle, "A Virophage at the Origin of Large DNA Transposons," *Science*, vol. 332, no. 6026, pp. 231–234, 2011. doi: 10.1126/science.1199412.

[87] D. L. Caspar, "Movement and self-control in protein assemblies. Quasi-equivalence revisited," *Biophysical Journal*, vol. 32, no. 1, pp. 103–138, 1980. doi: 10.1016/S0006-3495(80)84929-0.

[88] D. L. Caspar and A. Klug, "Physical principles in the construction of regular viruses," in *Cold Spring Harbor Symposia on Quantitative Biology*, vol. 27, pp. 1–24, Cold Spring Harbor Laboratory Press, 1962. doi: 10.1101/SQB.1962.027.001.005.

[89] H. Fukuhara, Y. Ino, and T. Todo, "Oncolytic virus therapy: a new era of cancer treatment at dawn," *Cancer Science*, vol. 107, no. 10, pp. 1373–1379, 2016. doi: 10.1111/cas.13027.

[90] A. Zlotnick, A. Lee, C. R. Bourne, J. M. Johnson, P. L. Domanico, and S. J. Stray, "In vitro screening for molecules that affect virus capsid assembly (and other protein association reactions)," *Nature Protocols*, vol. 2, no. 3, pp. 490–498, 2007. doi: 10.1038/nprot.2007.60.

[91] D. Born, L. Reuter, U. Mersdorf, M. Mueller, M. G. Fischer, A. Meinhart, and J. Reinstein, "Capsid protein structure, self-assembly, and processing reveal morphogenesis of the marine virophage mavirus," *Proceedings of the National Academy of Sciences*, vol. 115, no. 28, pp. 7332–7337, 2018. doi: 10.1073/pnas.1805376115.

[92] G. Chelvanayagam, J. Heringa, and P. Argos, "Anatomy and evolution of proteins displaying the viral capsid jellyroll topology," *Journal of Molecular Biology*, vol. 228, no. 1, pp. 220–242, 1992. doi: 10.1016/0022-2836(92)90502-B.

[93] S. Cheng and C. L. Brooks III, "Viral Capsid Proteins Are Segregated in Structural Fold Space," *PLoS Computational Biology*, vol. 9, no. 2, p. e1002905, 2013. doi: 10.1371/journal.pcbi.1002905.

[94] M. M. Roberts, J. L. White, M. G. Grutter, and R. M. Burnett, "Three-dimensional structure of the adenovirus major coat protein hexon," *Science*, vol. 232, no. 4754, pp. 1148–1151, 1986. doi: 10.1126/science.3704642.

[95] S. D. Benson, J. K. Bamford, D. H. Bamford, and R. M. Burnett, "Viral Evolution Revealed by Bacteriophage PRD1 and Human Adenovirus Coat Protein Structures," *Cell*, vol. 98, no. 6, pp. 825–833, 1999. doi: 10.1016/S0092-8674(00)81516-0.

[96] T. Klose, D. G. Reteno, S. Benamar, A. Hollerbach, P. Colson, B. La Scola, and M. G. Rossmann, "Structure of faustovirus, a large dsDNA virus," *Proceedings of the National Academy of Sciences*, vol. 113, no. 22, pp. 6206–6211, 2016. doi: 10.1073/pnas.1523999113.

[97] J. T. Vivian and P. R. Callis, "Mechanisms of Tryptophan Fluorescence Shifts in Proteins," *Biophysical Journal*, vol. 80, no. 5, pp. 2093–2109, 2001. doi: 10.1016/S0006-3495(01)76183-8.

[98] V. Nanda and L. Brand, "Aromatic interactions in homeodomains contribute to the low quantum yield of a conserved, buried tryptophan," *Proteins: Structure, Function, and Bioinformatics*, vol. 40, no. 1, pp. 112–125, 2000. doi: 10.1002/(SICI)1097-0134(20000701)40:1%3C112::AID-PROT130%3E3.0.CO;2-C.

[99] B. Lutz, C. Sinner, S. Bozic, I. Kondov, and A. Schug, "Native structure-based modeling and simulation of biomolecular systems per mouse click," *BMC Bioinformatics*, vol. 15, no. 1, p. 292, 2014. doi: 10.1186/1471-2105-15-292.

[100] M. J. Abraham, "Performance enhancements for GROMACS nonbonded interactions on BlueGene," *Journal of Computational Chemistry*, vol. 32, no. 9, pp. 2041–2046, 2011. doi: 10.1002/jcc.21766.

[101] C. A. Laughton, "A study of simulated annealing protocols for use with molecular dynamics in protein structure prediction," *Protein Engineering, Design and Selection*, vol. 7, no. 2, pp. 235–241, 1994. doi: 10.1093/protein/7.2.235.

[102] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw, "Improved side-chain torsion potentials for the Amber ff99SB protein force field," *Proteins: Structure, Function, and Bioinformatics*, vol. 78, no. 8, pp. 1950–1958, 2010. doi: 10.1002/prot.22711.

[103] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers: Original Research on Biomolecules*, vol. 22, no. 12, pp. 2577–2637, 1983. doi: 10.1002/bip.360221211.

[104] X. Zhang, S. Sun, Y. Xiang, J. Wong, T. Klose, D. Raoult, and M. G. Rossmann, "Structure of Sputnik, a virophage, at 3.5-Å resolution," *Proceedings of the National Academy of Sciences*, vol. 109, no. 45, pp. 18431–18436, 2012. doi: 10.1073/pnas.1211702109.

[105] M. J. Bennett, S. Choe, and D. Eisenberg, "Domain swapping: entangling alliances between proteins.," *Proceedings of the National Academy of Sciences*, vol. 91, no. 8, pp. 3127–3131, 1994. doi: 10.1073/pnas.91.8.3127.

[106] M. J. Bennett, M. P. Schlunegger, and D. Eisenberg, "3D domain swapping: a mechanism for oligomer assembly," *Protein Science*, vol. 4, no. 12, pp. 2455–2468, 1995. doi: 10.1002/pro.5560041202.

[107] S. Yang, S. S. Cho, Y. Levy, M. S. Cheung, H. Levine, P. G. Wolynes, and J. N. Onuchic, "Domain swapping is a consequence of minimal frustration," *Proceedings of the National Academy of Sciences*, vol. 101, no. 38, pp. 13786–13791, 2004. doi: 10.1073/pnas.0403724101.

[108] D. S. Goodsell and A. J. Olson, "Structural symmetry and protein function," *Annual Review of Biophysics and Biomolecular Structure*, vol. 29, no. 1, pp. 105–153, 2000. doi: 10.1146/annurev.biophys.29.1.105.

[109] F. H. C. Crick and J. D. Watson, "Structure of Small Viruses," *Nature*, vol. 177, no. 4506, pp. 473–475, 1956. doi: 10.1038/177473a0.

[110] A.-L. Hänninen, D. H. Bamford, and J. K. H. Bamford, "Assembly of Membrane-Containing Bacteriophage PRD1 is Dependent on GroEL and GroES," *Virology*, vol. 227, no. 1, pp. 207–210, 1997. doi: 10.1006/viro.1996.8308.

[111] C. L. Cepko and P. A. Sharp, "Assembly of adenovirus major capsid protein is mediated by a nonvirion protein," *Cell*, vol. 31, no. 2, pp. 407–415, 1982. doi: 10.1016/0092-8674(82)90134-9.

[112] J. J. Rux, P. R. Kuser, and R. M. Burnett, "Structural and Phylogenetic Analysis of Adenovirus Hexons by Use of High-Resolution X-Ray Crystallographic, Molecular Modeling, and Sequence-Based Methods," *Journal of Virology*, vol. 77, no. 17, pp. 9553–9566, 2003. doi: 10.1128/JVI.77.17.9553-9566.2003.

[113] L. Mindich, D. Bamford, T. McGraw, and G. Mackenzie, "Assembly of bacteriophage PRD1: particle formation with wild-type and mutant viruses.," *Journal of Virology*, vol. 44, no. 3, pp. 1021–1030, 1982.

[114] K. N. Parent, R. Khayat, L. H. Tu, M. M. Suhanovsky, J. R. Cortines, C. M. Teschke, J. E. Johnson, and T. S. Baker, "P22 Coat Protein Structures Reveal a Novel Mechanism for Capsid Maturation: Stability without Auxiliary Proteins or Chemical Crosslinks," *Structure*, vol. 18, no. 3, pp. 390–401, 2010. doi: 10.1016/j.str.2009.12.014.

[115] Y. P. Chuan, Y. Y. Fan, L. H. L. Lua, and A. P. J. Middelberg, "Virus assembly occurs following a pH- or Ca^{2+}-triggered switch in the thermodynamic attraction between structural protein capsomeres," *Journal of The Royal Society Interface*, vol. 7, no. 44, pp. 409–421, 2010. doi: 10.1098/rsif.2009.0175.

[116] A. Stortelder, J. Hendriks, J. B. Buijs, J. Bulthuis, C. Gooijer, S. M. van der Vies, and G. van der Zwan, "Hexamerization of the Bacteriophage T4 Capsid Protein gp23 and Its W13V Mutant Studied by Time-Resolved Tryptophan Fluorescence," *The Journal of Physical Chemistry B*, vol. 110, no. 49, pp. 25050–25058, 2006. doi: 10.1021/jp064881t.

[117] I. Gertsman, L. Gan, M. Guttman, K. Lee, J. A. Speir, R. L. Duda, R. W. Hendrix, E. A. Komives, and J. E. Johnson, "An unexpected twist in viral capsid maturation," *Nature*, vol. 458, no. 7238, pp. 646–650, 2009. doi: 10.1038/nature07686.

[118] N. G. Abrescia, J. J. Cockburn, J. M. Grimes, G. C. Sutton, J. M. Diprose, S. J. Butcher, S. D. Fuller, C. San Martín, R. M. Burnett, D. I. Stuart, *et al.*, "Insights into assembly from structural analysis of bacteriophage PRD1," *Nature*, vol. 432, no. 7013, pp. 68–74, 2004. doi: 10.1038/nature03056.

[119] A. G. Kikhney and D. I. Svergun, "A practical guide to small angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins," *FEBS Letters*, vol. 589, no. 19, pp. 2570–2577, 2015. doi: 10.1016/j.febslet.2015.08.027.

[120] B. Schuler, A. Soranno, H. Hofmann, and D. Nettels, "Single-Molecule FRET Spectroscopy and the Polymer Physics of Unfolded and Intrinsically Disordered Proteins," *Annual Review of Biophysics*, vol. 45, no. 1, pp. 207–231, 2016. doi: 10.1146/annurev-biophys-062215-010915.

[121] D. Thirumalai, H. S. Samanta, H. Maity, and G. Reddy, "Universal nature of collapsibility in the context of protein folding and evolution," *Trends in Biochemical Sciences*, 2019. doi: 10.1016/j.tibs.2019.04.003.

[122] J. A. Riback, M. A. Bowman, A. M. Zmyslowski, C. R. Knoverek, J. M. Jumper, J. R. Hinshaw, E. B. Kaye, K. F. Freed, P. L. Clark, and T. R. Sosnick, "Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water," *Science*, vol. 358, no. 6360, pp. 238–241, 2017. doi: 10.1126/science.aan5774.

[123] J. A. Riback, M. A. Bowman, A. M. Zmyslowski, C. R. Knoverek, J. M. Jumper, E. B. Kaye, K. F. Freed, P. L. Clark, and T. R. Sosnick, "Response to Comment on "Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water"," *Science*, vol. 361, no. 6405, p. eaar7949, 2018. doi: 10.1126/science.aar7949.

[124] H. Maity and G. Reddy, "Folding of protein L with implications for collapse in the denatured state ensemble," *Journal of the American Chemical Society*, vol. 138, no. 8, pp. 2609–2616, 2016. doi: 10.1021/jacs.5b11300.

[125] G. Fuertes, N. Banterle, K. M. Ruff, A. Chowdhury, D. Mercadante, C. Koehler, M. Kachala, G. Estrada Girona, S. Milles, A. Mishra, P. R. Onck, F. Gräter, S. Esteban-Martín, R. V. Pappu, D. I. Svergun, and E. A. Lemke, "Decoupling of size and shape fluctuations in heteropolymeric sequences reconciles discrepancies in SAXS vs. FRET measurements," *Proceedings of the National Academy of Sciences*, vol. 114, no. 31, pp. E6342–E6351, 2017. doi: 10.1073/pnas.1704692114.

[126] G. Fuertes, N. Banterle, K. M. Ruff, A. Chowdhury, R. V. Pappu, D. I. Svergun, and E. A. Lemke, "Comment on "Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water"," *Science*, vol. 361, no. 6405, p. eaau8230, 2018. doi: 10.1126/science.aau8230.

[127] R. B. Best, W. Zheng, A. Borgia, K. Buholzer, M. B. Borgia, H. Hofmann, A. Soranno, D. Nettels, K. Gast, A. Grishaev, *et al.*, "Comment on "Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water"," *Science*, vol. 361, no. 6405, p. eaar7101, 2018. doi: 10.1126/science.aar7101.

[128] A. Borgia, W. Zheng, K. Buholzer, M. B. Borgia, A. Schüler, H. Hofmann, A. Soranno, D. Nettels, K. Gast, A. Grishaev, R. B. Best, and B. Schuler, "Consistent View of Polypeptide Chain Expansion in Chemical Denaturants from Multiple Experimental Methods," *Journal of the American Chemical Society*, vol. 138, no. 36, pp. 11714–11726, 2016. doi: 10.1021/jacs.6b05917.

[129] A. L. Main, T. S. Harvey, M. Baron, J. Boyd, and I. D. Campbell, "The three-dimensional structure of the tenth type III module of fibronectin: An insight into RGD-mediated interactions," *Cell*, vol. 71, no. 4, pp. 671–678, 1992. doi: 10.1016/0092-8674(92)90600-H.

[130] R. Pankov and K. M. Yamada, "Fibronectin at a glance," *Journal of Cell Science*, vol. 115, no. 20, pp. 3861–3863, 2002. doi: 10.1242/jcs.00059.

[131] C. M. Williams, A. J. Engler, R. D. Slone, L. L. Galante, and J. E. Schwarzbauer, "Fibronectin expression modulates mammary epithelial cell proliferation during acinar differentiation," *Cancer Research*, vol. 68, no. 9, pp. 3185–3192, 2008. doi: 10.1158/0008-5472.CAN-07-2673.

[132] C. A. McPhalen and M. N. G. James, "Crystal and molecular structure of the serine proteinase inhibitor CI-2 from barley seeds," *Biochemistry*, vol. 26, no. 1, pp. 261–269, 1987. doi: 10.1021/bi00375a036.

[133] S. E. Jackson, N. elMasry, and A. R. Fersht, "Structure of the hydrophobic core in the transition state for folding of chymotrypsin inhibitor 2: A critical test of the protein engineering method of analysis," *Biochemistry*, vol. 32, no. 42, pp. 11270–11278, 1993. doi: 10.1021/bi00093a002.

[134] L. S. Itzhaki, D. E. Otzen, and A. R. Fersht, "The Structure of the Transition State for Folding of Chymotrypsin Inhibitor 2 Analysed by Protein Engineering Methods: Evidence for a Nucleation-Condensation Mechanism for Protein Folding," *Journal of Molecular Biology*, vol. 254, no. 2, pp. 260–288, 1995. doi: 10.1006/jmbi.1995.0616.

[135] T. R. Killick, S. M. V. Freund, and A. R. Fersht, "Real-time NMR studies on folding of mutants of barnase and chymotrypsin inhibitor 2," *FEBS Letters*, vol. 423, no. 1, pp. 110–112, 1998. doi: 10.1016/S0014-5793(98)00075-1.

[136] J. L. Neira, L. S. Itzhaki, D. E. Otzen, B. Davis, and A. R. Fersht, "Hydrogen exchange in chymotrypsin inhibitor 2 probed by mutagenesis," *Journal of Molecular Biology*, vol. 270, no. 1, pp. 99–110, 1997. doi: 10.1006/jmbi.1997.1088.

[137] S. L. Kazmirski, K.-B. Wong, S. M. V. Freund, Y.-J. Tan, A. R. Fersht, and V. Daggett, "Protein folding from a highly disordered denatured state: the folding pathway of chymotrypsin inhibitor 2 at atomic resolution," *Proceedings of the National Academy of Sciences*, vol. 98, no. 8, pp. 4349–4354, 2001. doi: 10.1073/pnas.071054398.

[138] W. Kremer, B. Schuler, S. Harrieder, M. Geyer, W. Gronwald, C. Welker, R. Jaenicke, and H. R. Kalbitzer, "Solution NMR structure of the cold-shock protein from the hyperthermophilic bacterium Thermotoga maritima," *European Journal of Biochemistry*, vol. 268, no. 9, pp. 2527–2539, 2001. doi: 10.1046/j.1432-1327.2001.02127.x.

[139] M. Mueller, U. Grauschopf, T. Maier, R. Glockshuber, and N. Ban, "The structure of a cytolytic α-helical toxin pore reveals its assembly mechanism," *Nature*, vol. 459, no. 7247, pp. 726–730, 2009. doi: 10.1038/nature08026.

[140] A. J. Wallace, T. J. Stillman, A. Atkins, S. J. Jamieson, P. A. Bullough, J. Green, and P. J. Artymiuk, "E. coli hemolysin E (HlyE, ClyA, SheA): X-ray crystal structure of the toxin and observation of membrane pores by electron microscopy.," *Cell*, vol. 100, no. 2, pp. 265–76, 2000. doi: 10.1016/S0092-8674(00)81564-0.

[141] url: https://www.thermofisher.com.

[142] url: https://biotium.com.

[143] D. van der Spoel, E. Lindahl, B. Hess, A. R. van Buuren, E. Apol, P. J. Meulenhoff, D. P. Tieleman, A. L. T. M. Sijbers, K. A. Feenstra, R. van Drunen, and H. J. C. Berendsen, *Gromacs User Manual Version 4.5.6*, 2010. url: http://www.gromacs.org.

[144] J. Henriques, L. Arleth, K. Lindorff-Larsen, and M. Skepö, "On the calculation of SAXS profiles of folded and intrinsically disordered proteins from computer simulations," *Journal of Molecular Biology*, vol. 430, no. 16, pp. 2521–2539, 2018. doi: 10.1016/j.jmb.2018.03.002.

[145] P. J. Flory, "The Configuration of Real Polymer Chains," *The Journal of Chemical Physics*, vol. 17, pp. 303–310, 1949. doi: 10.1063/1.1747243.

[146] S. Ahn, K. H. Kim, Y. Kim, J. Kim, and H. Ihee, "Protein Tertiary Structural Changes Visualized by Time-Resolved X-ray Solution Scattering," *The Journal of Physical Chemistry B*, vol. 113, no. 40, pp. 13131–13133, 2009. doi: 10.1021/jp906983v.

[147] M. Andersson, E. Malmerberg, S. Westenhoff, G. Katona, M. Cammarata, A. B. Wöhri, L. C. Johansson, F. Ewald, M. Eklund, M. Wulff, J. Davidsson, and R. Neutze, "Structural Dynamics of Light-Driven Proton Pumps," *Structure*, vol. 17, pp. 1265–1275, 2009. doi: 10.1016/j.str.2009.07.007.

[148] J. S. Hub, "Interpreting solution X-ray scattering data using molecular simulations," *Current Opinion in Structural Biology*, vol. 49, pp. 18–26, 2018. doi: 10.1016/j.sbi.2017.11.002.

[149] E. Karaca, J. P. G. L. M. Rodrigues, A. Graziadei, A. M. J. J. Bonvin, and T. Carlomagno, "M3: an integrative framework for structure determination of molecular machines," *Nature Methods*, vol. 14, no. 9, pp. 897–902, 2017. doi: 10.1038/nmeth.4392.

[150] R. Shevchuk and J. S. Hub, "Bayesian refinement of protein structures and ensembles against SAXS data using molecular dynamics," *PLoS Computational Biology*, vol. 13, no. 10, p. e1005800, 2017. doi: 10.1371/journal.pcbi.1005800.

[151] G. Hummer and J. Köfinger, "Bayesian ensemble refinement by replica simulations and reweighting," *The Journal of Chemical Physics*, vol. 143, p. 243150, 2015. doi: 10.1063/1.4937786.

[152] P.-c. Chen and J. S. Hub, "Validating Solution Ensembles from Molecular Dynamics Simulation by Wide-Angle X-ray Scattering Data," *Biophysical Journal*, vol. 107, no. 2, pp. 435–447, 2014. doi: 10.1016/j.bpj.2014.06.006.

[153] W. Rieping, M. Habeck, and M. Nilges, "Inferential Structure Determination," *Science*, vol. 309, no. 5732, pp. 303–306, 2005. doi: 10.1126/science.1110428.

[154] V. V. Volkov and D. I. Svergun, "Uniqueness of ab initio shape determination in small-angle scattering," *Journal of Applied Crystallography*, vol. 36, no. 3 I, pp. 860–864, 2003. doi: 10.1107/S0021889803000268.

[155] D. Franke and D. I. Svergun, "DAMMIF, a program for rapid ab-initio shape determination in small-angle scattering," *Journal of Applied Crystallography*, vol. 42, no. 2, pp. 342–346, 2009. doi: 10.1107/S0021889809000338.

[156] A. A. Gupta, I. Reinartz, G. Karunanithy, A. Spilotros, V. R. Jonna, A. Hofer, D. I. Svergun, A. J. Baldwin, A. Schug, and M. Wolf-Watz, "Formation of a Secretion-Competent Protein Complex by a Dynamic Wrap-around Binding Mechanism," *Journal of Molecular Biology*, vol. 430, no. 18, pp. 3157–3169, 2018. doi: 10.1016/j.jmb.2018.07.014.

[157] H. Takala, A. Björling, O. Berntsson, H. Lehtivuori, S. Niebling, M. Hoernke, I. Kosheleva, R. Henning, A. Menzel, J. A. Ihalainen, and S. Westenhoff, "Signal amplification and transduction in phytochrome photosensors," *Nature*, vol. 509, pp. 245–248, 2014. doi: 10.1038/nature13310.

[158] D. Arnlund, L. C. Johansson, C. Wickstrand, A. Barty, G. J. Williams, E. Malmerberg, J. Davidsson, D. Milathianaki, D. P. DePonte, R. L. Shoeman, *et al.*, "Visualizing a protein quake with time-resolved X-ray scattering at a free-electron laser," *Nature Methods*, vol. 11, no. 9, pp. 923–926, 2014. doi: 10.1038/nmeth.3067.

[159] O. Berntsson, R. P. Diensthuber, M. R. Panman, A. Björling, E. Gustavsson, M. Hoernke, A. J. Hughes, L. Henry, S. Niebling, H. Takala, J. A. Ihalainen, G. Newby, S. Kerruth, J. Heberle, M. Liebi, A. Menzel, R. Henning, I. Kosheleva, A. Möglich, and S. Westenhoff, "Sequential conformational transitions and α-helical supercoiling regulate a sensor histidine kinase," *Nature Communications*, vol. 8, no. 284, pp. 1–8, 2017. doi: 10.1038/s41467-017-00300-5.

[160] M. R. Panman, E. Biasin, O. Berntsson, M. Hermann, S. Niebling, A. J. Hughes, J. Kübel, K. Atkovska, E. Gustavsson, A. Nimmrich, *et al.*, "Observing the Structural Evolution in the Photodissociation of Diiodomethane with Femtosecond Solution X-Ray Scattering," *Physical Review Letters*, vol. 125, no. 22, p. 226001, 2020. doi: 10.1103/PhysRevLett.125.226001.

[161] P. A. Doyle and P. S. Turner, "Relativistic Hartree-Fock X-ray and electron scattering factors," *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, vol. 24, no. 3, pp. 390–397, 1968. doi: 10.1107/S0567739468000756.

[162] R. D. B. Fraser, T. P. MacRae, and E. Suzuki, "An Improved Method for Calculating the Contribution of Solvent to the X-Ray Diffraction Pattern of Biological Molecules," *Journal of Applied Crystallography*, vol. 11, pp. 693–694, 1978. doi: 10.1107/S0021889878014296.

[163] S. Yang, S. Park, L. Makowski, and B. Roux, "A Rapid Coarse Residue-Based Computational Method for X-Ray Solution Scattering Characterization of Protein Folds and Multiple Conformational States of Large Protein Complexes," *Biophysical Journal*, vol. 96, no. 11, pp. 4449–4463, 2009. doi: 10.1016/j.bpj.2009.03.036.

[164] S. Niebling, A. Björling, and S. Westenhoff, "MARTINI bead form factors for the analysis of time-resolved X-ray scattering of proteins," *Journal of Applied Crystallography*, vol. 47, no. 4, pp. 1190–1198, 2014. doi: 10.1107/S1600576714009959.

[165] A. Schug, C. Hyeon, and J. N. Onuchic, "Coarse-Grained Structure-Based Simulations of Proteins and RNA," in *Coarse-Graining of Condensed Phase and Biomolecular Systems* (G. Voth, ed.), ch. 9, pp. 123–140, Boca Raton: CRC Press Taylor & Francis, 1 ed., 2008.

[166] D. Schneidman-Duhovny, M. Hammel, J. A. Tainer, and A. Sali, "Accurate SAXS profile computation and its assessment by contrast variation experiments," *Biophysical Journal*, vol. 105, no. 4, pp. 962–974, 2013. doi: 10.1016/j.bpj.2013.07.020.

[167] C. J. McKnight, P. T. Matsudaira, and P. S. Kim, "NMR structure of the 35-residue villin headpiece subdomain," *Nature Structural Biology*, vol. 4, no. 3, pp. 180–184, 1997. doi: 10.1038/nsb0397-180.

[168] G. F.-L. Ames, "Bacterial Periplasmic Transport Systems: Structure, Mechanism, and Evolution," *Annual Review of Biochemistry*, vol. 55, no. 1, pp. 397–425, 1986. doi: 10.1146/annurev.bi.55.070186.002145.

[169] C. W. Müller, G. J. Schlauderer, J. Reinstein, and G. E. Schulz, "Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding," *Structure*, vol. 4, no. 2, pp. 147–156, 1996. doi: 10.1016/S0969-2126(96)00018-4.

[170] C. W. Müller and G. E. Schulz, "Structure of the complex between adenylate kinase from *Escherichia coli* and the inhibitor Ap_5A refined at 1.9 Å resolution: A model for a catalytic transition state," *Journal of Molecular Biology*, vol. 224, no. 1, pp. 159–177, 1992. doi: 10.1016/0022-2836(92)90582-5.

[171] P. C. Whitford, O. Miyashita, Y. Levy, and J. N. Onuchic, "Conformational Transitions of Adenylate Kinase: Switching by Cracking," *Journal of Molecular Biology*, vol. 366, no. 5, pp. 1661–1671, 2007. doi: 10.1016/j.jmb.2006.11.085.

[172] A. Marina, C. D. Waldburger, and W. A. Hendrickson, "Structure of the entire cytoplasmic portion of a sensor histidine-kinase protein.," *The EMBO Journal*, vol. 24, no. 24, pp. 4247–4259, 2005. doi: 10.1038/sj.emboj.7600886.

[173] M. Weiel, M. Götz, A. Klein, D. Coquelin, R. Floca, and A. Schug, "Dynamic particle swarm optimization of biomolecular simulation parameters with flexible objective functions," *Nature Machine Intelligence*, pp. 1–8, 2021. doi: 10.1038/s42256-021-00366-3.

[174] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of ICNN'95 - International Conference on Neural Networks*, vol. 4, pp. 1942–1948, IEEE, 1995. doi: 10.1109/ICNN.1995.488968.

[175] J. Kennedy, "The particle swarm: social adaptation of knowledge," in *Proceedings of 1997 IEEE International Conference on Evolutionary Computation (ICEC'97)*, pp. 303–308, IEEE, 1997. doi: 10.1109/ICEC.1997.592326.

[176] Y. Shi and R. Eberhart, "A Modified Particle Swarm Optimizer," in *1998 IEEE International Conference on Evolutionary Computation Proceedings. IEEE World Congress on Computational Intelligence (Cat. No. 98TH8360)*, pp. 69–73, IEEE, 1998. doi: 10.1109/ICEC.1998.699146.

[177] M. Clerc and J. Kennedy, "The Particle Swarm – Explosion, Stability, and Convergence in a Multidimensional Complex Space," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 1, pp. 58–73, 2002. doi: 10.1109/4235.985692.

[178] K. E. Parsopoulos and M. N. Vrahatis, "Recent approaches to global optimization problems through particle swarm optimization," *Natural Computing*, vol. 1, no. 2-3, pp. 235–306, 2002. doi: 10.1023/A:1016568309421.

[179] T. Blackwell, "Particle Swarm Optimization in Dynamic Environments," in *Evolutionary Computation in Dynamic and Uncertain Environments*, pp. 29–49, Springer, 2007. doi: 10.1007/978-3-540-49774-5_2.

[180] X. Cui, C. T. Hardin, R. K. Ragade, T. E. Potok, and A. S. Elmaghraby, "Tracking non-stationary optimal solution by particle swarm optimizer," in *Sixth International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing and First ACIS International Workshop on Self-Assembling Wireless Network*, pp. 133–138, IEEE, 2005. doi: 10.1109/SNPD-SAWN.2005.77.

[181] K. O. Stanley, J. Clune, J. Lehman, and R. Miikkulainen, "Designing neural networks through neuroevolution," *Nature Machine Intelligence*, vol. 1, no. 1, pp. 24–35, 2019. doi: 10.1038/s42256-018-0006-z.

[182] M. Taherkhani and R. Safabakhsh, "A novel stability-based adaptive inertia weight for particle swarm optimization," *Applied Soft Computing*, vol. 38, pp. 281–295, 2016. doi: 10.1016/j.asoc.2015.10.004.

[183] R. C. Eberhart and Y. Shi, "Comparing inertia weights and constriction factors in particle swarm optimization," in *Proceedings of the 2000 Congress on Evolutionary Computation. CEC00 (Cat. No. 00TH8512)*, vol. 1, pp. 84–88, IEEE, 2000. doi: 10.1109/CEC.2000.870279.

[184] M. E. H. Pedersen and A. J. Chipperfield, "Simplifying particle swarm optimization," *Applied Soft Computing*, vol. 10, no. 2, pp. 618–628, 2010. doi: 10.1016/j.asoc.2009.08.029.

[185] M. Meissner, M. Schmuker, and G. Schneider, "Optimized Particle Swarm Optimization (OPSO) and its application to artificial neural network training," *BMC Bioinformatics*, vol. 7, no. 1, p. 125, 2006. doi: 10.1186/1471-2105-7-125.

[186] M. S. Nobile, P. Cazzaniga, D. Besozzi, R. Colombo, G. Mauri, and G. Pasi, "Fuzzy Self-Tuning PSO: A settings-free algorithm for global optimization," *Swarm and Evolutionary Computation*, vol. 39, pp. 70–85, 2018. doi: 10.1016/j.swevo.2017.09.001.

[187] R. Poli, "Analysis of the Publications on the Applications of Particle Swarm Optimisation," *Journal of Artificial Evolution and Applications*, vol. 2008, 2008. doi: 10.1155/2008/685175.

[188] S. Sengupta, S. Basak, and R. A. Peters, "Particle Swarm Optimization: A Survey of Historical and Recent Developments with Hybridization Perspectives," *Machine Learning and Knowledge Extraction*, vol. 1, no. 1, pp. 157–191, 2019. doi: 10.3390/make1010010.

[189] M. N. A. Wahab, S. Nefti-Meziani, and A. Atyabi, "A Comprehensive Review of Swarm Optimization Algorithms," *PLoS ONE*, vol. 10, no. 5, pp. 1–36, 2015. doi: 10.1371/journal.pone.0122827.

[190] J. H. Holland, "Genetic Algorithms," *Scientific American*, vol. 267, no. 1, pp. 66–73, 1992. url: https://www.jstor.org/stable/24939139.

[191] R. Storn and K. Price, "Differential Evolution – A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces," *Journal of Global Optimization*, vol. 11, no. 4, pp. 341–359, 1997. doi: 10.1023/A:1008202821328.

[192] M. Dorigo, "Optimization, learning and natural algorithms," *Ph.D. Thesis, Politecnico di Milano, Italy*, 1992. url: https://ci.nii.ac.jp/naid/10016599043/en/.

[193] M. Dorigo, M. Birattari, and T. Stutzle, "Ant colony optimization," *IEEE Computational Intelligence Magazine*, vol. 1, no. 4, pp. 28–39, 2006. doi: 10.1109/MCI.2006.329691.

[194] D. Karaboga, "An idea based on honey bee swarm for numerical optimization," tech. rep., Citeseer, 2005.

[195] D. Karaboga and B. Basturk, "A powerful and efficient algorithm for numerical function optimization: artificial bee colony (abc) algorithm," *Journal of Global Optimization*, vol. 39, no. 3, pp. 459–471, 2007. doi: 10.1007/s10898-007-9149-x.

[196] T. Navalertporn and N. V. Afzulpurkar, "Optimization of tile manufacturing process using particle swarm optimization," *Swarm and Evolutionary Computation*, vol. 1, no. 2, pp. 97–109, 2011. doi: 10.1016/j.swevo.2011.05.003.

[197] P. J. Pawar, R. V. Rao, and J. P. Davim, "Multiobjective Optimization of Grinding Process Parameters Using Particle Swarm Optimization Algorithm," *Materials and Manufacturing Processes*, vol. 25, no. 6, pp. 424–431, 2010. doi: 10.1080/10426910903124860.

[198] C. Ma and L. Qu, "Multiobjective Optimization of Switched Reluctance Motors Based on Design of Experiments and Particle Swarm Optimization," *IEEE Transactions on Energy Conversion*, vol. 30, no. 3, pp. 1144–1153, 2015. doi: 10.1109/TEC.2015.2411677.

[199] C. Zhang, Z. Chen, Q. Mei, and J. Duan, "Application of Particle Swarm Optimization Combined with Response Surface Methodology to Transverse Flux Permanent Magnet Motor Optimization," *IEEE Transactions on Magnetics*, vol. 53, no. 12, pp. 1–7, 2017. doi: 10.1109/TMAG.2017.2749565.

[200] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," *CoRR*, 2015. url: http://arxiv.org/abs/1502.03167.

[201] R. F. Alford, A. Leaver-Fay, J. R. Jeliazkov, M. J. O'Meara, F. P. DiMaio, H. Park, M. V. Shapovalov, P. D. Renfrew, V. K. Mulligan, K. Kappel, *et al.*, "The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design," *Journal of Chemical Theory and Computation*, vol. 13, no. 6, pp. 3031–3048, 2017. doi: 10.1021/acs.jctc.7b00125.

[202] J. K. Leman, B. D. Weitzner, S. M. Lewis, J. Adolf-Bryfogle, N. Alam, R. F. Alford, M. Aprahamian, D. Baker, K. A. Barlow, P. Barth, *et al.*, "Macromolecular modeling and design in Rosetta: recent methods and frameworks," *Nature Methods*, pp. 1–14, 2020. doi: 10.1038/s41592-020-0848-2.

[203] M. Weiel, M. Götz, A. Klein, D. Coquelin, R. Floca, and A. Schug, "Minimal dataset repository for reproduction of presented results." FLAPS-NMI@Github (repository: FLAPS-sim_setups), 2021. https://github.com/FLAPS-NMI/FLAPS-sim_setups/releases/tag/v1.0, doi: 10.5281/zenodo.4773999.

[204] S. M. Sedlak, L. K. Bruetzel, and J. Lipfert, "Quantitative evaluation of statistical errors in small-angle X-ray scattering measurements," *Journal of Applied Crystallography*, vol. 50, no. 2, pp. 621–630, 2017. doi: 10.1107/S1600576717003077.

[205] D. I. Svergun, "Determination of the regularization parameter in indirect-transform methods using perceptual criteria," *Journal of Applied Crystallography*, vol. 25, no. pt 4, pp. 495–503, 1992. doi: 10.1107/S0021889892001663.

[206] H. B. Stuhrmann, "Interpretation of small-angle scattering functions of dilute solutions and gases. A representation of the structures related to a one-particle scattering function," *Acta Crystallographica Section A*, pp. 297–306, 1970. doi: 10.1107/S0567739470000748.

[207] D. Franke, M. V. Petoukhov, P. V. Konarev, A. Panjkovich, A. Tuukkanen, H. D. T. Mertens, A. Kikhney, N. R. Hajizadeh, J. M. Franklin, C. M. Jeffries, *et al.*, "ATSAS 2.8: a comprehensive data analysis suite for small-angle scattering from macromolecular solutions," *Journal of Applied Crystallography*, vol. 50, no. 4, pp. 1212–1225, 2017. doi: 10.1107/S1600576717007786.

[208] M. V. Petoukhov, D. Franke, A. V. Shkumatov, G. Tria, A. G. Kikhney, M. Gajda, C. Gorba, H. D. T. Mertens, P. V. Konarev, and D. I. Svergun, "New developments in the ATSAS program package for small-angle scattering data analysis," *Journal of Applied Crystallography*, vol. 45, no. 2, pp. 342–350, 2012. doi: 10.1107/S0021889812007662.

[209] A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, and M. Karplus, "All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins," *The Journal of Physical Chemistry B*, vol. 102, no. 18, pp. 3586–3616, 1998. doi: 10.1021/jp973084f.

[210] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, "Comparison of simple potential functions for simulating liquid water," *The Journal of Chemical Physics*, vol. 79, no. 2, pp. 926–935, 1983. doi: 10.1063/1.445869.

[211] D. Homouz, M. Perham, A. Samiotakis, M. Cheung, and P. Wittung-Stafshede, "Crowded, cell-like environment induces shape changes in aspherical protein," *Proceedings of the National Academy of Sciences*, vol. 105, no. 33, pp. 11754–11759, 2008. doi: 10.1073/pnas.0803672105.

[212] D. Bratton and T. Blackwell, "A simplified recombinant PSO," *Journal of Artificial Evolution and Applications*, vol. 2008, 2008. doi: 10.1155/2008/654184.

[213] M. Weiel, M. Götz, A. Klein, D. Coquelin, R. Floca, and A. Schug, "FLAPS Hyppopy code repository." FLAPS-NMI@Github (repository: FLAPS-Hyppopy), 2021. https://github.com/FLAPS-NMI/FLAPS-Hyppopy/releases/tag/v1.0, doi: 10.5281/zenodo.4773970.

[214] M. Weiel, M. Götz, A. Klein, D. Coquelin, R. Floca, and A. Schug, "FLAPS Optunity code repository." FLAPS-NMI@Github (repository: FLAPS-optunity), 2021. https://github.com/FLAPS-NMI/FLAPS-optunity/releases/tag/v1.0, doi: 10.5281/zenodo.4773992.

[215] M. Soysal, M. Berghoff, T. Zirwes, M.-A. Vef, S. Oeste, A. Brinkmann, W. E. Nagel, and A. Streit, "Using On-Demand File Systems in HPC Environments," in *2019 International Conference on High Performance Computing & Simulation (HPCS)*, pp. 390–398, IEEE, 2019. doi: 10.1109/HPCS48598.2019.9188216.