Technische Universität München

Fakultät für Chemie

Lehrstuhl für Computergestützte Biokatalyse

Integrative multi-scale simulations of complex biomolecular systems

Alexander Jussupow, M. Sc.

Vollständiger Abdruck der von der Fakultät für Chemie der Technischen Universität München zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften (Dr. rer. nat.) genehmigten Dissertation.

Vorsitzende/-r: Prof. Dr. Job Boekhoven

Prüfende/-r der Dissertation:

1. Prof. Dr. Ville R. I. Kaila
2. Prof. Dr. Michael Sattler
3. Prof. Dr. Carlo Camilloni

Die Dissertation wurde am 12.07.2021 bei der Technischen Universität München eingereicht und durch die promotionsführende Einrichtung 16.08.2021 angenommen.

Table of Contents

II

Abstract

In-silico methods facilitate the investigation of the properties of complex biomolecular systems. In this thesis, we study the dynamics of Complex I, linear polyubiquitin, and Hsp90 using a combination of quantum mechanical and QM/MM methods, atomistic and coarse-grained molecular dynamics simulation, as well as enhanced sampling and integration of experimental data in a Bayesian framework. By applying such computational methods at different scales, we provide insights complementary to experimental results. In this thesis, we characterize the quinone binding site in Complex I, and validate the findings by cryo-EM, as well as additional functional studies. We also link the quinone dynamics with large-scale conformational changes in Complex I and how these are regulated by lipids such as cardiolipin. At a methodological level, we introduce a novel approach to integrate SAXS data into molecular dynamics simulations in the metadynamics & metainference framework, and apply these to the characterize highly dynamical linear polyubiquitin chains. Moreover, through the combination of computational and experimental methods, our studies also suggest that fully extended conformational states are pivotal for the dynamics and function of Hsp90.

IV

Zusammenfassung

In-silico Methoden ermöglichen die Untersuchung der Eigenschaften von komplexen biomolekularen Systemen. In dieser Dissertation untersuchen wir die Dynamik von Komplex I, linearem Polyubiquitin und Hsp90 mit einer Kombination aus quantenmechanischen und QM/MM Methoden, atomistischer und grobkörniger Molekulardynamik, erweiterten Sampling-Methoden und der Integration von experimentellen SAXS-Daten in einem bayesschen Ansatzes. Durch die Verwendung von Berechnungsmethoden unterschiedlicher Skalen liefern wir Erkenntnisse, welche die experimentellen Befunde komplementieren. Hierdurch charakterisieren wir die Chinon-Bindungsstelle im Komplex I und validieren diese durch eine neue Kryo-EM-Struktur und zusätzlichen funktionellen Studien. Außerdem verknüpfen wir die Chinon-Dynamik mit großräumigen Konformationsänderungen im Komplex I, reguliert durch Cardiolipin. Zusätzlich führen wir eine neuartige Strategie zur Integration von SAXS-Daten in Molekulardynamik-Simulationen mittels Metadynamics & Metainference ein und wenden diese auf die hochdynamischen linearen Polyubiquitin-Ketten an. Darüber hinaus deuten unsere Resultate aus einer Kombination von rechnerischen und experimentellen Methoden darauf hin, dass vollständig erweiterte Konformatione, für die Dynamik und Funktion von Hsp90 entscheidend sind.

VI

List of publications and author contribution

Respiratory complex I (C-I – C-IV)

C-I <u>**A. Jussupow***</u>, A. Di Luca*, V. R. I. Kaila, How cardiolipin modulates the dynamics of respiratory complex I. *Sci Adv* **5**, eaav1850 (2019).

C-II A. P. Gamiz-Hernandez, <u>**A. Jussupow**</u>, M. P. Johansson, V. R. I. Kaila, Terminal Electron-Proton Transfer Dynamics in the Quinone Reduction of Respiratory Complex I. *J Am Chem Soc* **139**, 16282-16288 (2017).

C-III H. R. Bridges, J. G. Fedor, J. N. Blaza, A. Di Luca, **A. Jussupow**, O. D. Jarman, J. J. Wright, A. A. Agip, A. P. Gamiz-Hernandez, M. M. Roessler, V. R. I. Kaila, J. Hirst, Structure of inhibitor-bound mammalian complex I. *Nat Commun* **11**, 5261 (2020).

C-IV M. Röpke, D. Riepl, P. Saura, A. Di Luca, M. E. Mühlbauer, **A. Jussupow**, A. P. Gamiz-Hernandez, V. R. I. Kaila, *PNAS* (2021) *accepted*

Own contribution: In publication C-I, I performed all cgMD simulations, was involved in the project design, analysis and interpretation of the computational results, and manuscript creation. In paper C-II, I performed QM cluster model calculations and QM/MM simulations on the ubiquinone setups. In paper C-III and C-IV, I performed cgMD simulations. In both papers, I was involved in analyzing and interpreting the computational results and manuscript creation.

Integration of SAXS data into MD simulations (S-I – S-IV)

S-I <u>**A. Jussupow**</u>, A. C. Messias, R. Stehle, A. Geerlof, S. M. O. Solbak, C. Paissoni, A. Bach, M. Sattler, C. Camilloni, The dynamics of linear polyubiquitin. *Sci Adv* **6**, eabc3786 (2020).

S-II C. Paissoni, <u>**A. Jussupow**</u>, C. Camilloni, Martini bead form factors for nucleic acids and their application in the refinement of protein–nucleic acid complexes against SAXS data. *Journal of Applied Crystallography* **52**, 394-402 (2019).

S-III C. Paissoni, <u>**A. Jussupow**</u>, C. Camilloni, Determination of Protein Structural Ensembles by Hybrid-Resolution SAXS Restrained Molecular Dynamics. *J Chem Theory Comput* **16**, 2825-2834 (2020).

S-IV T. Lohr, **A. Jussupow**, C. Camilloni, Metadynamic metainference: Convergence towards force field independent structural ensembles of a disordered peptide. *J Chem Phys* **146**, 165102 (2017).

Own contribution: In publication S-I, I performed all cgMD simulations, was involved in the project design, analysis and interpretation of the experimental and computational results, and manuscript creation. In publications SII and SIII, I developed the used SAXS module in Plumed and the initial hySAXS approach. In publication SIV, I was involved in the expansion of the M&M approach. In all three publications, I was involved in the manuscript creation.

The dynamics of Hsp90 (H-I)

H-I **A. Jussupow***, A. Lopez*, M. Baumgart*, S. L. Mader, M. Sattler, V. R. I. Kaila, Extended conformational dynamics is crucial for the function of Hsp90. (2021) *submitted*

Own contribution: In publication H-I, I performed and analyzed all aMD and cgMD simulations, was involved in the project design, analysis and interpretation of the experimental and computational results, and manuscript creation.

*Authors contributed equally.

Other papers (not part of the thesis)

1.   M. C. Ahmed, L. K. Skaanning, **A. Jussupow**, E. A. Newcombe, B. B. Kragelund, C. Camilloni, A. E. Langkilde, K. Lindorff-Larsen, Refinement of α-Synuclein Ensembles Against SAXS Data: Comparison of Force Fields and Methods. *Frontiers in Molecular Biosciences* **8**, (2021).

2.   B. A. K. Kriebisch, **A. Jussupow**, A. M. Bergmann, F. Kohler, H. Dietz, V. R. I. Kaila, J. Boekhoven, Reciprocal Coupling in Chemically Fueled Assembly: A Reaction Cycle Regulates Self-Assembly and Vice Versa. *J Am Chem Soc* **142**, 20837-20844 (2020).

3.   Z. Niu, E. Prade, E. Malideli, K. Hille, **A. Jussupow**, Y. G. Mideksa, L. M. Yan, C. Qian, M. Fleisch, A. C. Messias, R. Sarkar, M. Sattler, D. C. Lamb, M. J. Feige, C. Camilloni, A. Kapurniotu, B. Reif, Structural Insight into IAPP-Derived Amyloid Inhibitors and Their Mechanism of Action. *Angew Chem Int Ed Engl* **59**, 5771-5781 (2020).

4.   C. Wanzke, **A. Jussupow**, F. Kohler, H. Dietz, V. R. I. Kaila, J. Boekhoven, Dynamic Vesicles Formed By Dissipative Self-Assembly. *ChemSystemsChem* **2**, e1900044 (2019).

5.   M. Fottner, A. D. Brunner, V. Bittl, D. Horn-Ghetko, **A. Jussupow**, V. R. I. Kaila, A. Bremm, K. Lang, Site-specific ubiquitylation and SUMOylation using genetic-code expansion and sortase. *Nat Chem Biol* **15**, 276-284 (2019).

6.   P. consortium, Promoting transparency and reproducibility in enhanced molecular simulations. *Nat Methods* **16**, 670-673 (2019).

7.   H. Kooshapur, N. R. Choudhury, B. Simon, M. Muhlbauer, **A. Jussupow**, N. Fernandez, A. N. Jones, A. Dallmann, F. Gabel, C. Camilloni, G. Michlewski, J. F. Caceres, M. Sattler, Structural basis for terminal loop recognition and stimulation of pri-miRNA-18a processing by hnRNP A1. *Nat Commun* **9**, 2479 (2018).

8.   D. C. Rodriguez Camargo, K. J. Korshavn, **A. Jussupow**, K. Raltchev, D. Goricanec, M. Fleisch, R. Sarkar, K. Xue, M. Aichler, G. Mettenleiter, A. K. Walch, C. Camilloni, F. Hagn, B. Reif, A. Ramamoorthy, Stabilization and structural analysis of a membrane-associated hIAPP aggregation intermediate. *Elife* **6** (e31226), (2017).

x

1 Introduction

1.1 The dilemma of biomolecular simulation

In recent decades, computational or *in-silico* methods have emerged as powerful tools to gain insights into the behavior and properties of individual biomolecules as well as their interactions with each other. This is due to the increasing availability of computational resources and improvements in theoretical techniques, specifically in the field of quantum mechanics (QM) (*1*) and molecular dynamics (MD) (*2*). Nevertheless, the choice of an appropriate computational method for a scientific question depends on factors such as the available computational resources, system size, required accuracy, and time scale, as each approach involves an inevitable trade-off.

In principle, methods that handle more degrees of freedom can generate more detailed results for local quantities and address questions inaccessible to simpler models. However, this comes at the sacrifice of sampling and system size. For example, quantum mechanical methods (see section 2.1), involving an explicit description of electrons, allow the characterization of the formation and breaking of chemical bonds, as well as complex quantum effects, which are not accessible to classical molecular dynamics simulations (see section 2.2). The latter assumes fixed chemical bonds and approximates the complex behavior of electrons with simple point charges localized at atomic positions. The accurate description of all electrons is, nevertheless, extraordinarily computational demanding, limiting QM methods to small system sizes of a few hundred atoms and sampling in ps time scale (*3*). As such, it is not possible to sample a high number of conformations, which is often crucial for understanding biochemical processes.

A similar trade-off also exists between atomistic MD (aMD) and coarse-grained (cg) MD (*4*) simulations (see section 2.4), where a group of individual atoms is represented by larger-sized beads. While atomistic MD simulations are usually limited to a µs timescale for a system of up to a million atoms (*5*), cgMD force fields like MARTINI (*6-8*) can be used to simulate systems of up to a hundred million atoms or extend the sampling to a millisecond timescale, but at the expense of accuracy due to the loss of isotropic interactions (*4*). Therefore, cgMD simulations are often used in areas where directional interactions (*e.g.*, hydrogen bonds) are not critical. A notable example would be membrane systems, which are primarily stabilized by relatively unspecific van-der-Waals interactions.

At the same time, the intrinsic conformational dynamics of similarly sized proteins (under native conditions) can vary dramatically, while the dynamics of single-domain proteins can often be adequately expressed as fluctuations around

a well-defined structure. In these cases, a long sampling time might not be necessary to fully characterize the properties of a protein. On the other side, even tiny, disordered proteins may necessitate a detailed conformational ensemble because rare states or transitions might be critical for their functions (*9*). Moreover, multi-domain proteins can show substantial conformational flexibility between individual (rigid) domains.

Different strategies have been developed to extend the capabilities of pure QM or MD simulations. A common approach is the usage of hybrid methods like QM/MM (see section 2.5) (*10*): Hereby, a small part, *e.g.*, the catalytic center, is treated at the QM level, while the surrounding environment is modeled at a classical level. As such, a much larger system can be simulated but without significantly increasing the computational cost. A similar strategy exists by combining atomistic with coarse-grained MD simulations. This type of approach is often referred to as "multi-scale".

A completely different approach is taken by enhanced sampling methods (*11, 12*). Two commonly used examples are metadynamics (*13, 14*) and umbrella sampling (*15, 16*) (see sections 2.7). In both cases, the goal is to speed up the sampling of collective variables (CVs) of special interest. As such, these methods require some form of prior knowledge about the dynamics of the system to determine the CVs, while a few new approaches are trying to determine the CVs automatically (*17, 18*). Effectively, these methods can increase the time scales of processes accessible by QM or MD methods. However, the major drawback of these approaches is that the simulation trajectories do not correctly represent the dynamics of the system, but rather reproduce the conformational space after statistical re-weighting.

At last, introducing experimental data into simulations can increase their accuracy, either posterior using reweighting approaches (*19, 20*) or during the simulation by restraining them to reproduce an experimental value. For the latter, a Bayesian approach such as metainference (see section 2.8) (*21*) can be used to account for different sources of experimental and methodical errors in the biomolecular force fields. Suitable experimental data for such simulations are often obtained by nuclear magnetic resonance (NMR) spectroscopy (such as chemical shifts (*22*), or paramagnetic relaxation enhancement (PRE) (*23*)) (*24*), small-angle X-ray scattering (SAXS) (see section 2.9) (*25*), or cryogenic electron microscopy (cryo-EM) (*26*). Depending on the type of experimental data, the additional computational cost of estimating the experimental restraints may introduce an additional computational cost in the simulations and requires prior experimental knowledge of the system. Nevertheless, combining enhanced sampling methods with the integration of experimental data can increase the sampling and accuracy of a simulation.

In this thesis, a wide range of computational methods was employed to explore the dynamics of challenging biomolecular systems. Through the combined use of QM calculations, QM/MM, aMD, and cgMD simulations (publications C-I to IV), we established how the coupling between Complex I and the surrounding membrane modulates its functionality by controlling the quinone dynamics within the quinone cavity of Complex I. In publication (S-I to IV), we introduced a new (hybrid) approach to incorporate SAXS data into aMD and cgMD simulations, allowing us to characterize the properties of linear polyubiquitin chains, with implications for the behavior of all types of polyubiquitin chains. In publication H-I, we utilized aMD and cgMD simulation together with SAXS and PRE data to highlight the importance of extended conformation for the functionality of Hsp90.

Table 1.1: Overview of computational methods

	QM models	QM/MM	aMD	cgMD	Enhanced sampling	Integrating experimental data
Level	Electronic structure	Electronic and atomistic detail	Atomistic detail	Large-scale details	Free energy surface rather than dynamics	Increased accuracy
System size	< 1000	mixed	a few million	up to 100 million	(Increased)	
Sampling	fs-ps	ps-ns	μs	ms	Free energy landscape	

1.2 Systems

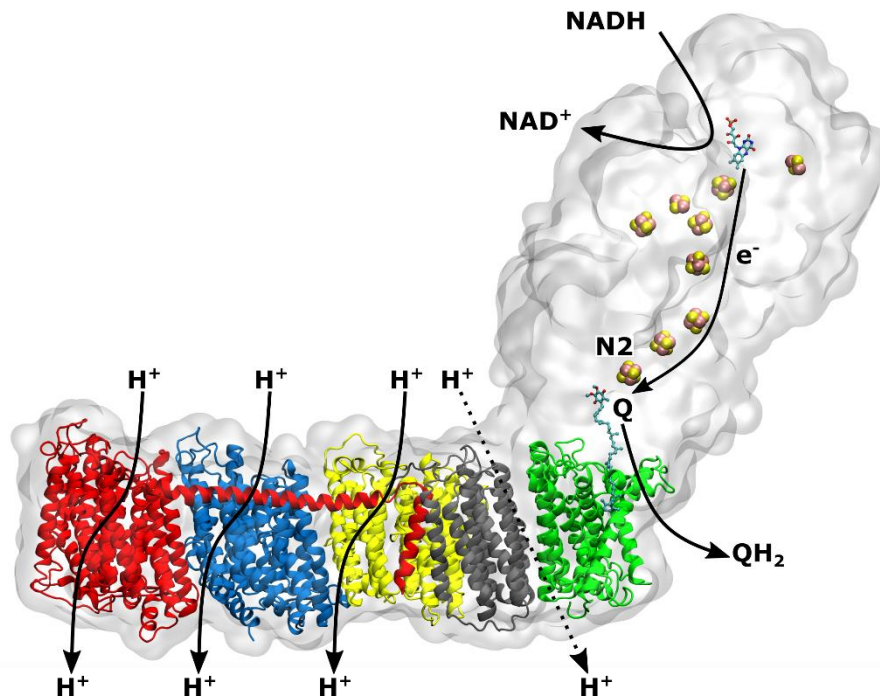### 1.2.1 Respiratory Complex I

Figure 1.1: Structure of respiratory Complex I of T. thermophiles. The hydrophobic domain is explicitly shown as a cartoon representation (Nqo12 red, Nqo13 blue, Nqo14 yellow, Nqo7/10/11 gray, Nqo12 green).

The respiratory chain is responsible for the energy conversion in the mitochondria and bacteria. Its main entry point is Complex I (**Figure** 1.1) (*27-29*), with up to 45 subunits and a molecular mass of up to 1 MDa (*30-33*), and the largest and least understood enzyme of the respiratory chain. Complex I acts as a redox-driven proton pump by catalyzing the electron transfer between nicotinamide adenine dinucleotide (NADH) and quinone (Q) via a chain of iron-sulfur centers (ISC). The transduced energy is harnessed to transfer protons from the negative (n) to the positive (p) site of the inner mitochondrial membrane or the cytoplasmic membrane in bacteria (*34, 35*). The overall reaction is:

$$NADH + H^+ + Q + 4H_n^+ \longrightarrow NAD^+ + QH_2 + 4H_p^+ \qquad 1.2.1$$
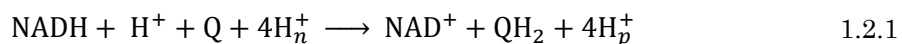
This process establishes an electrochemical proton gradient, often referred to as proton motive force (pmf), which is subsequently utilized by F_oF_1-ATP synthase to synthesize adenosine triphosphate (ATP) (*36, 37*). Remarkably, this reaction is reversible, allowing Complex I to employ the pmf to produce NADH (*38*). Additionally, depending on the organism, the quinone, either in the form of ubiquinone (Q) or coenzyme Q_{10} (in humans) (*39*), menaquinone (*40*), or plastoquinone (*41*), and the reduced quinol (QH_2), act as a two-electron carrier which is also used by other enzymes of the respiratory chain. Q_{10} is composed of a

benzoquinone headgroup and a long aliphatic tail. The Q reduction process acts as a coupling element between the electron and proton transfer which are spatially separated by up to 200 Å across different domains of the enzyme (*42*). However, the mechanism of this remarkable coupling remains elusive. Thus, substances with a similar structure to the benzoquinone headgroup, but without redox activity can inhibit Complex I. One such example is piericidin A (*43-45*), which resembles the short-chain ubiquinone (Q_3) but has a pyridone instead of benzoquinone headgroup (**Figure** 1.2).
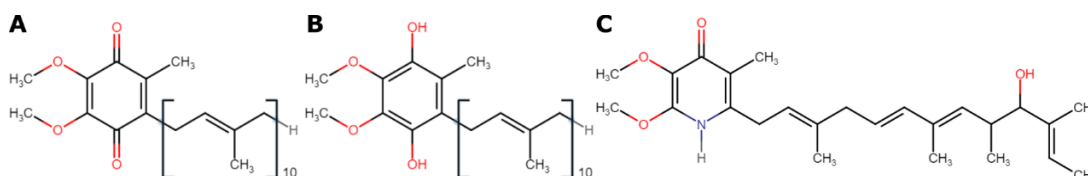

Figure 1.2: Structure of ubiquinone (A), ubiquinol (B), and piericidin A (C)

Overall, Complex I consists of 14 conserved core subunits(*46*), forming an L-shape (**Figure** 1.1). The nomenclature of the subunits depends on the specific enzyme, but for consistency, the nomenclature of *T. thermophiles* will be used here if not otherwise mentioned. The mammalian Complex I has 31 additional supernumerary subunits, which are likely contribute to assembly, stability, and regulation (*31, 47, 48*). The core subunits can be categorized into the solvent-exposed hydrophilic domain and the membrane-bound hydrophobic domain. The hydrophilic domain of Complex I is primarily responsible for the electron transfer from NADH to Q, with the NADH binding site at the top edge of the hydrophilic domain and the Q binding site close to the membrane plane (*28, 42, 49*). The electron transfer process takes around 100 µs (*49*) and progresses through a "wire" formed by flavin mononucleotide (FMN) and multiple iron-sulfur centers (ISC). The terminal N2 ISC acts as the electron donor for the Q. Since the ISCs act as one electron carriers, the transfer of two separate electrons is required to reduce Q to QH_2.

The hydrophobic domain consists of seven core subunits located in the membrane plane, containing the proton pumping subunits of Complex I. The transfer of three (of the four) protons is performed by the antiporter-like subunits Nqo12, Nqo13, and Nqo14 (*50, 51*). The pathway for the fourth proton is still unknown, but it is likely located either between the Nqo11/Nq14 or Nqo8/Nqo10 subunit (*30, 52*). The pumping units are coupled by a chain of charged amino acids connecting Nqo8 with Nqo12 (*52-54*). The Q binding pocket is located at the interface between the Nqo4 and Nqo6 subunits of hydrophilic and the Nqo8 subunit of the membrane domain (*52, 55-58*). The binding pocket consists of a hydrophobic upper and lower part as well as a kink region in the middle containing a significant amount of polar and charged residues. The quinone headgroup binds at the top of the cavity

~20-30 Å above the membrane plane and around 12 Å away from the N2 ISC (*30, 52*). His38 and Tyr87 of Nqo4 coordinate the Q-headgroup at the active site suggested to act as proton donors (*52, 56, 59*), but so far, no experimental resolved structure with a Q_{10} inside the quinone pocket have been determined. Nevertheless, EPR data (*60, 61*), recent computational studies (*62*), as well as cryoEM structures (*63, 64*), suggest possible additional quinone binding sites in the cavity.

Recently resolved Cryo-EM structures reveal that Complex I can undergo large conformational changes, which have been associated with the transition from the active (A) to deactive (D) state of the mammalian enzyme (*31, 33, 65, 66*). In the deactive state, Complex I exhibits a low turnover rate that may be linked to bending and twisting motions around the hydrophilic and membrane domain during this transition. These motions also occur in the bacterial Complex I although this variant lack the A/D transition states (*66*) suggesting that global dynamics is conserved across different species and might be functionally important for Complex I.

As a membrane protein, Complex I is impacted by the composition of the embedding membrane. In particular, the anionic cardiolipin (1,3-bis(sn-3'-phosphatidyl)-sn-glycerol, or diphosphatidylglycerol) has been found to affect the turnover of Complex I (*67-69*), as well as other respiratory enzymes (*70, 71*). Cardiolipin constitutes about 20% of the inner mitochondrial membrane and is composed of two phosphatidic acid groups linked by a glycerol backbone carrying two aliphatic chains each. In contrast, lipids such as POPE (palmitoyl-oleoyl phosphatidylethanolamine) or POPC (palmitoyl-oleoyl phosphatidylcholine) comprise only one phosphatic acid group carrying two aliphatic chains. Recent Cryo-EM structures (*32, 33*) show some bound cardiolipins around Complex I, but kinetic experiments suggest that up to ten cardiolipin molecules are bound to Complex I (*72*). However, the functional role of these lipids still remains unclear.

In chapter 3 of this thesis, the results of publication CI to CIV are presented, where we characterize Q-binding in the active site, validated by a new cryo-EM structure of Complex I with bound piericidin. We also highlight the quinone dynamics and show its coupling with the previously mentioned large-scale motions of Complex I and the presence of cardiolipin.
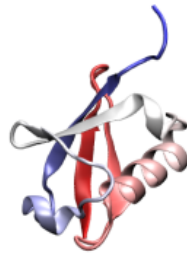
1.2.2 Polyubiquitin

Figure 1.3: Crystal structure of ubiquitin. The colors indicate the residues from C-terminus (blue) to N-terminus (red).

Ubiquitin is a highly conserved 76 amino acids long regulatory protein. For example, the amino acid sequence between yeast and mammalian isoforms deviates only by three amino acids (*73*). The defining feature of this compact 8.6 kDa large protein is a ß-grasp fold(*74*) with five antiparallel ß-strands interacting with an α-helix (**Figure** 1.3). The hydrophobic patch around the residues Ile44, Leu8, and Val70 serves as the main interaction site with other proteins (*75*), but other spots like the Ile36 (*76*), the Phe4 (*77*), and the Asp58 (*78*) patch potentially assist in protein recognition.

Ubiquitination, the process of adding ubiquitin as a post-transcriptional modification, regulates various physiological processes in the cell, including protein degradation, DNA repair, apoptosis, cell cycle, and signal transduction (*79-81*). This process typically occurs through an iso-peptide bond on the ε-amino group of a substrate lysine or through a regular peptide bond with the N-terminus of the substate. Alternative sides such as the sulfhydryl group on cysteine or the hydroxyl group on threonine and serine have also been observed (*82*). Subsequent to this modification, the process can be repeated on the ubiquitin by linking the C-terminus of the distal with Lys6/11/27/29/33/48/63 or Met1 of the proximal ubiquitin. Thus, chains of diverse topologies can be formed, where both the linking site (*83, 84*) as well as the chain length (*85*) regulate the recognition of the chains by different binding partners and, therefore, their function. However, the role of the chain length and the dynamics of polyubiquitin is poorly understood (*84-86*). Due to the flexibility of the chains, these systems are a major challenge for molecular simulations.

In this thesis, the Met1-linked, linear polyubiquitin chains were of primary interest. These linear chains are involved in inflammation and immune responses (*87-89*), specifically through their engagement in the NF-κB signaling (*87, 90-95*). Hereby, the IKK complex, composed of the catalytic IKKα, IKKß subunits, and the regulatory NF-κB essential modulator (NEMO), is activated by the linear ubiquitin chain assembly complex (LUBAC) (*96*), which catalyzes the assembly of linear polyubiquitin. In turn, linear polyubiquitin can bind to the helical coiled

dimeric UBAN motif in NEMO (*95, 97*). Interestingly, although the isolated NEMO$_{UBAN}$ domain recognizes diubiquitin, the full-length NEMO can only bind Ub$_4$ or longer (*91*). Moreover, chains comprising more than ten ubiquitins appear to induce a conformational change in NEMO (*90*). Thus, the activation mechanism might be length dependent.

Isolated linear polyubiquitin chains display a highly dynamic behavior (*98*), and consequently, a comprehensive ensemble is essential to characterize their dynamics. In Chapter 5 and publication S-I, we present a novel method for incorporating SAXS data into cgMD simulations, allowing us to accurately characterize the dynamics of polyubiquitin. Our results indicate that a simple self-avoiding polymer model might be enough to describe the properties of all polyubiquitin chains. In publication S-II to S-IV, we establish the foundation for utilizing SAXS data in the metadynamics and metainference (M&M) (*99*) framework, and we introduce a hybrid approach for the integration of SAXS into aMD simulations.
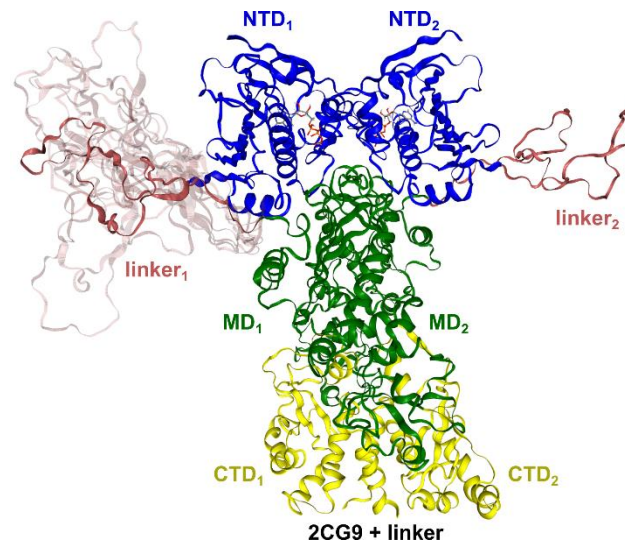
1.2.3 Hsp90

Figure 1.4: Hsp90 structure based on the PDB ID: 2CG9 crystal structure. The image is obtained modified from the publication HI.

The heat shock protein 90 (Hsp90) is a 90 kDa well-conserved ATP-dependent molecular chaperone that regulates the folding of client proteins in the eukaryotic cell (*100-103*). As one of the most abounded protein, accounting for 1-2% of all proteins in cells (*104*), Hsp90 interacts with hundreds of different client proteins (*105*), and its dysfunction is implicated in the development of cancer (*106, 107*).

Hsp90 is a homodimer, where the monomers alone are not functional (*108*). At the N-terminus is the 25 kDa large ATP-binding N-terminal domain (NTD). A ~60 amino acid long highly charged linker connects the NTD to the 35 kDa large middle domain (MD) (*109-113*), whereas the 30 kDa large C-terminal domain (CTD) is essential for the dimerization of Hsp90 (*114*). An Hsp90 construct containing only NTD and MD primarily forms a non-functional monomer but has been used here to study the functional dynamics of this protein.

During its catalytic cycle, Hsp90 undergoes a conformational change from an open form to a closed conformation, where the NTDs and MDs of individual monomers are in contact with each other (*115-118*), as shown in **Figure** 1.4. In the *apo*-state, the Hsp90 dimer is in an open state that could allow for the recognition of misfolded proteins through the NTD and MD domains (*119-121*). However, the conformational ensemble of open conformations is currently unknown. SAXS experiments (*111, 113*) suggest that the open conformation has an average radius of gyration of (R_g) >60 Å, compared to R_g ~ 41 Å in the compact nucleotide-bound state(*122-124*).

In publication H-I, we show, based on experimental data, mutagenesis experiments, and *in-silico* results, that a separation between the NTD and MD domains is crucial for the dynamics of Hsp90.

2 Methods

In this chapter, a brief overview is shown of all the different theoretical methods used in publications presented in chapter 3, containing sections about quantum mechanical and classical mechanical methods, as well as theory behind the used enhanced sampling methods and the incorporating experimental data into simulations. Table 2.1 gives an overview of the relevant sections for each paper.

Table 2.1: Overview of relevant sections for individual papers

Paper	Sections
C-I	2.2, 2.3, 2.4, 2.6
C-II	2.2, 2.3, 2.1, 2.5
C-III	2.2, 2.3, 2.4
C-IV	2.2, 2.3, 2.4
S-I	2.2, 2.4, 2.7, 2.8, 2.9
S-II	2.2, 2.3, 2.9
S-III	2.2, 2.3, 2.7, 2.8, 2.9
S-IV	2.2, 2.3, 2.7, 2.8
H-I	2.2, 2.3, 2.4

2.1 Density Functional Theory

Quantum mechanical (QM) methods are applied to study the electronic structure of molecules in order to investigate processes like chemical reactions or electron transfer events. These methods can reach a high level of accuracy but are only applicable for small system sizes (typically up to a few hundred atoms). Originally, QM methods were based on the description of the wave function with (post-)Hartree-Fock methods (125). Nevertheless, density functional theory (DFT) is now widely adopted as it provides a well-balanced compromise between accuracy and computational cost.

In DFT, the energy calculation is carried out using an electron density instead of wave functions. The theoretical basis is provided by the two Hohenberg-Kohn theorems (126). The first one proves that the ground state density uniquely defines the external potential (and thus also the Hamiltonian). Therefore, the properties of the electronic structure can be expressed in terms of the three spatial coordinates of the density instead of the high-dimensional wave function. Thus, the energy of the system is a functional of electron density:

$$E = E[\rho(\boldsymbol{r})] \qquad\qquad 2.1.1$$

The second theorem states that the electron density of the ground state corresponds to the minimum energy. Therefore, it is possible to optimize the electron density with the variational principle.

However, an exact form for energy functional is not known. Typically, the Kohn-Sham approximation (127) is adopted to circumvent this problem. Thereby, a (fictional) system of non-interactive electrons residing in an external potential is assumed, which generates the same density as for a set of interactive particles. As such, the density can be expressed based on single-electron Kohn–Sham orbitals φ:

$$\rho(\boldsymbol{r}) = \sum_{i=1}^{N} |\varphi(\boldsymbol{r})|^2 \qquad\qquad 2.1.2$$

The associated ground state energy E_{KS} can be divided into the known term for kinetic energy T_s, a term generated by the external potential V_{ext} (at the minimum, a potential caused by the nuclei), the average repulsion of the electron density E_{H} and into the unknown exchange-correlation term E_{xc}:

$$E_{\mathrm{KS}}[\rho(\boldsymbol{r})] = T_{\mathrm{S}}[\rho(\boldsymbol{r})] + E_{ext}[\rho(\boldsymbol{r})] + E_{\mathrm{H}}[\rho(\boldsymbol{r})] + E_{xc}[\rho(\boldsymbol{r})] =$$

$$= \frac{1}{2}\sum_{i=1}^{N}\langle\varphi_i|\nabla_i|\varphi_i\rangle + \int \rho(\boldsymbol{r})\, V_{\mathrm{ext}}(\boldsymbol{r})\, d\boldsymbol{r} + \frac{1}{2}\int \frac{\rho(\boldsymbol{r})\rho(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|}\, d\boldsymbol{r}d\boldsymbol{r}' \qquad 2.1.3$$

$$+ E_{xc}[\rho(\boldsymbol{r})]$$

2 Methods

Like in the wave-function based Hartree Fock method, the Kohn-Sham orbitals are optimized in a self-consistent field approach since the energy functional depends on the density itself.

As the choice of the exchange-correlation functional is crucial for an accurate description of the electronic structure, various approaches have been proposed to express this term. The simplest one is local density approximation (LDA) (*127*), representing the electron density as a homogeneous electron gas. This approach works decently well for metals but not for organic molecules. Hence, the generalized gradient approximation (GGA) (*128*) was introduced, where the energy also depends on the electron density gradient. These functionals can be further improved by adding higher-order derivatives (meta-GGA) (*129*) and/or replacing a fraction of the local density functional with the exact exchange term from the Hartree Fock approach.

Hybrid functionals, such as B3LYP (*130, 131*), which is of the most common functionals in chemistry, combines local (LDA and VWN (*132*)), gradient corrected functionals (B88x (*133*) and LYP (*130*)), and the Hartree Fock exchange with three parameters $\alpha_0 = 0.2$, $\alpha_X = 0.72$ and $\alpha_C = 0.81$:

$$E_{XC}^{B3LYP} = (1 - \alpha_0)E_X^{LDA} + \alpha_0 E_X^{HF} + \alpha_X E_X^{B88} + \alpha_C E_C^{LYP} + (1 - \alpha_C)E_C^{LDA} \qquad 2.1.4$$

The Kohn–Sham orbitals molecular orbitals are expressed as a linear combination of atomic orbitals (LCAO), which in turn are often represented as a linear combination of Gauss type orbitals (GTOs)(*134*) due to their analytical integral and easy multiplication. The general form of a GTO is often expressed in spherical coordinates:

$$\varphi_{l,m}(r,\theta,\Phi) = Y_{l,m}(\theta,\Phi) \cdot r^l \cdot e^{-\alpha r^2} \qquad 2.1.5$$

where $Y_{l,m}$ is the angular momentum part depending on the quantum number l and m, r is the distance of the electrons from the nuclei, and alpha is the orbital exponent related to the (effective) charge of the nucleus. In this work, the basis functions from the Ahlrichs def2 family (*135*) were used.

In this thesis, multiple complementary methods were used to overcome certain limitations of DFT. Firstly, dispersion effects are usually not well reproduced in DFT. Thus, the empirical Grimme's D3 dispersion correction term (*136*) is widely added to improve the accuracy of the functional. Additionally, the electrostatic interaction with the surrounding solvent or protein can play a significant role in the local electronic structure. Specifically, a solvent with a high relative permittivity (*e.g.*, water) will lead to more stable charged residues than a medium with a low dielectric constant (*e.g.*, the interior of a protein). Here, the environment is either embedded with the Conductor-like Screening Model (COSMO) (*137*) approximation or explicitly with QM/MM (see section 2.5). Lastly,

we use the broken symmetry approach to model the complicated spin behavior of iron-sulfur clusters (*138*).

2.2 Classical Molecular Dynamics

Detailed electronic structure calculations cannot characterize the conformational dynamics of large biomolecular systems. Instead, using an approach based on classical mechanics is often more appropriate for such problems. This approach is warranted when chemical reactions or quantum effects can be neglected. In molecular dynamics (MD), the forces caused by interacting particles are propagated by numerically solving Newton's second law of motion. This approach has been implemented in various software packages like NAMD (*139*) or GROMACS (*140*). The interactions between particles are usually represented by a *force field*, which describes the potential energy as a function of the relative positions of the particles (see section 2.3). This approach results in a trajectory of frames that represents the motions of particles over time.

2.2.1 Integration of Newton's second law of motion

The acting force **F** acting upon the particle i at position x on time t, can be derived from the gradient of the potential energy:

$$\mathbf{F}\big(\boldsymbol{x}_i(t)\big) = -\nabla \mathbf{V}\big(\boldsymbol{x}_i(t)\big) \qquad\qquad 2.2.1$$

According to Newton's second law of motion, this force is equal to the product of the mass m and the acceleration $\boldsymbol{a}$ of the particle. It can be rewritten as a differential equation since the acceleration $\boldsymbol{a}$ is the first derivative of the velocity $\boldsymbol{v}$, which in turn is the first derivative of the position of a particle:

$$\mathbf{F}\big(x_i(t)\big) = m_i \cdot \boldsymbol{a}_i(t) = m_i \cdot \frac{\partial \boldsymbol{v}_i(t)}{\partial t} = m \cdot \frac{\partial^2 \boldsymbol{x}_i(t)}{\partial t^2} \qquad\qquad 2.2.2$$

Due to the typically large system sizes in MD simulations, equation 2.2.2 is repeatedly solved numerically (rather than analytically) at each subsequent time step Δt, resulting in a trajectory. Common integrators are based on the Verlet algorithm (e.g., used by default in NAMD (*139*)) and the leapfrog algorithm (e.g., used by default in GROMACS (*140*)) (*141*). Both algorithms are time-reversible and symplectic, ensuring the conservation of the total energy of the system, i.e., the sum of kinetic and potential energy. In the leapfrog integrator, the updated position and velocity are calculated by:

$$\boldsymbol{a}_i(t) = -\frac{\nabla \mathbf{V}\big(\boldsymbol{x}_i(t)\big)}{m_i} \qquad\qquad 2.2.3$$

$$\boldsymbol{v}_i\left(t + \frac{1}{2}\Delta t\right) = \boldsymbol{v}_i\left(t - \frac{1}{2}\Delta t\right) + \boldsymbol{a}_i(t) \cdot \Delta t \qquad\qquad 2.2.4$$

$$\boldsymbol{x}_i(t + \Delta t) = \boldsymbol{x}_i(t) + \boldsymbol{v}_i\left(t + \frac{1}{2}\Delta t\right) \cdot \Delta t \qquad \text{2.2.5}$$

Since the short-term accuracy is limited, Δt must be significantly smaller than the time required for the shortest events of the system. For example, for atomistic MD simulations of biomolecules, the fastest process is usually the C-H bond vibration, leading to a required time step between 1-2 fs. Nevertheless, using a strategy of restraining hydrogen bonds and utilizing virtual sites, it is possible to extend it to 4-5 fs (*142*).

The leapfrog algorithm requires initial velocities for $t = t_0 - \frac{1}{2}\Delta t$ to start a new simulation. These are usually generated by assigning random (atomistic) velocities adherent to the Maxwell-Boltzmann distribution. The probability p for the velocity of a particle in a given direction can be calculated as:

$$p(v_{x,i}) = \sqrt{\frac{m_i}{2\pi \cdot k_\mathrm{B}T}} \cdot \exp\left(-\frac{m_i v_{x,i}}{k_\mathrm{B}T}\right) \qquad \text{2.2.6}$$

where k_B is the Boltzmann constant, and T is the (absolute) temperature of the system.

2.2.2   Periodic boundary condition

To avoid effects caused by the edges of the biomolecular systems, MD simulations are generally performed under periodic boundary conditions(143). Hereby, the particles of a simulated system are inside a box with the dimension vectors $\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{z}$ which act as a unit cell, and is repeated in all three dimensions. Based on the minimum image convention, the box size R should be at least twice the cut-off for non-bonded interaction:

$$R = \frac{1}{2}\min(\|\mathbf{x}\|, \|\mathbf{y}\|, \|\mathbf{z}\|) \qquad \text{2.2.7}$$

This ensures that there are only interactions with the nearest image of each particle. The optimal shape of the box depends on the shape of the biomolecules of interest. For example, for globular proteins (e.g., ubiquitin), using a dodecahedron instead of a cubic box requires a smaller volume for a given minimum distance between proteins of two boxes. Therefore, a simulation with a dodecahedron box would require a smaller number of water molecules.

2.2.3   Temperature and pressure coupling

Conventional MD simulations result in a microcanonical or NVE ensemble where the volume of the system, as well as the number of particles, are both constant, and the MD integrator preserves the total energy. However, the balance between

potential and kinetic energy changes during the simulation. The average kinetic energy $\langle E_{\text{kin}} \rangle$ is directly associated with the temperature of a system:

$$\langle E_{\text{kin}} \rangle = \frac{3}{2} N \cdot k_{\text{B}} T \qquad \text{2.2.8}$$

Temperature fluctuation is generally not desired since, typically, this parameter remains constant under experimental conditions. To produce a canonical or *NVT* ensemble, an additional thermostat is added to the MD simulation to keep the temperature around the desired value. One popular approach is the *velocity rescaling(144)* algorithm, where a canonical distribution of the kinetic energy E_{kin} is imposed. For this, a target value for the kinetic energy $E_{\text{kin},t}$ is selected based on the probability distribution for E_{kin}:

$$p\left(E_{\text{kin},t}\right) \propto E_{\text{kin},t}^{\left(\frac{N_{\text{f}}}{2}-1\right)} \cdot \exp\left(-\frac{E_{\text{kin},t}}{k_{\text{B}}T}\right) \qquad \text{2.2.9}$$

with N_{f} as the number of degrees of freedom. The velocity rescale factor α can then be calculated as:

$$\alpha = \sqrt{\frac{E_{\text{kin},t}}{E_{\text{kin}}}} \qquad \text{2.2.10}$$

As long as the canonical distribution in equation 2.2.9 is satisfied, $E_{\text{kin},t}$ can be chosen not only randomly, but also based on its previous value $E_{\text{kin},t-1}$. As such, in the *velocity rescaling* algorithm, the change of kinetic energy dE_{kin} is defined with a coupling parameter τ as:

$$dE_{\text{kin}} = (\langle E_{\text{kin}} \rangle - E_{\text{kin}}) \frac{dt}{\tau} + 2\sqrt{\frac{\langle E_{\text{kin}} \rangle \cdot E_{\text{kin}}}{N_f}} \cdot \frac{dW}{\sqrt{\tau}} \qquad \text{2.2.11}$$

where dW is a Wiener noise.

Instead of performing simulations at constant volume, MD simulations can also be done under constant pressure to achieve an isothermal-isobaric or *NPT* ensemble by adding a barostat. In particular, the Parrinello-Rahman barostat(*145, 146*) is widely adopted, allowing the simulation box to change the volume and shape. The latter is vital for membrane systems, where anisotropic pressure coupling is necessary to ensure the correct properties of the membrane.

2.2.4 Adequate sampling

A pivotal issue in MD is the assessment of the necessary sampling time and quantification of the uncertainties since even large-scale high-performance resources do not guarantee convergence of target metrics. The first point depends on the importance of various states and the activation barriers separating them.

Assuming a simple two-state model, the probability of finding the system in the given states P_1 and P_2 corresponds to their Gibbs free energy difference ΔG:

$$\Delta G = -RT \cdot \ln\left(\frac{P_1}{P_2}\right) \qquad\qquad 2.2.12$$

where R is the universal gas constant.

Thus, an important but rare state might be only represented in a few frames of a simulation trajectory which may not be sufficient to draw conclusions about that particular configuration. Moreover, based on the transition state theory, the two states are separated by an activation barrier $\Delta G^{\ddagger}$, which determines the transition rate between them. The rate constant k can be estimated based on the Eyring equation (*147*):

$$k = \kappa \frac{k_{\mathrm{B}} T}{h} e^{-\frac{\Delta G^{\ddagger}}{RT}} \qquad\qquad 2.2.13$$

with κ as the transmission coefficient (often assumed as 1) and h as Planck's constant. Based on this, the half-life time $t_{1/2}$ of a state can be calculated as:

$$t_{1/2} = \frac{\ln(2)}{k} \qquad\qquad 2.2.14$$

Table 2.2 presents the lifetime $t_{1/2}$ of a state as a function of the activation barrier $\Delta G^{\ddagger}$. To obtain sufficient sampling for the transition between P_1 and P_2, $t_{1/2}$ of each states must be much smaller than the simulation time. Since typically, all-atom MD simulation can achieve a sampling time in the µs range, and $\Delta G^{\ddagger}$ for the process of interest must thus be below 10 kcal/mol to achieve a statistically significant number of transitions between the states.

As an alternative to performing one long simulation, multiple shorter once can be conducted, since according to the ergodic hypothesis (*148*), the time average of an observable $\overline{X}$ equals its average across a statistical ensemble $\langle X \rangle$:

$$\overline{X} = \lim_{t \to \infty} \frac{1}{t} \int X(t) \cdot dt = \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} X_i = \langle X \rangle \qquad\qquad 2.2.15$$

However, since the simulation length or the number of conformations in a trajectory is finite, the MD results have an averaging error which can be expressed as the standard deviation for the average value (or standard error) $\sigma_{\langle X \rangle_{\mathrm{sim}}}$. Based on the central-limit theory, $\sigma_{\langle X \rangle_{\mathrm{sim}}}$ depend on the number of configurations in the trajectory/trajectories N and the standard deviation of the observable σ_X:

$$\sigma_{\langle A \rangle_{\mathrm{sim}}} = \frac{\sigma_A}{\sqrt{N}} \qquad\qquad 2.2.16$$

$$\sigma_X = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \langle X \rangle_{sim})^2}{N - 1}} \qquad 2.2.17$$

Equation 2.2.17 is only valid for independent data points. In MD simulations, subsequent snapshots are generally correlated depending on the relaxation time of the specific property. A common approach to estimate the standard error of correlated data points is the block average method (*149, 150*). Hereby, a trajectory with $N = M \cdot n$ frames is grouped into a series of M blocks with a block length of n. Instead of using the individual data points to calculate the standard deviation, the set of block averages are used in equation 2.2.17. When the block length n is significantly longer than the correlation time, the individual block averages become independent from each other. Hence, the standard error can be estimated by dividing the standard deviation by the square root of the number of blocks.

Table 2.2 Lifetime depending on the activation barrier calculated based on the Eyring equation (with T=310 K).

$\Delta G^{\ddagger}$ [kcal/mol]	$t_{1/2}$
1.0	0.5 ps
2.5	6.2 ps
5.0	0.4 ns
7.5	20 ns
10.0	1.2 μs
12.5	70 μs
15.0	4.0 ms
17.5	0.2 s
20.0	14 s

2.3   Atomistic force fields

The force field is the backbone of the molecular dynamics simulations that defines the potential energy surface of the particles. The individual interactions are usually additive; hence, each energy component and the associated force can be calculated independently. In general, a force field is built based on a set of classical analytical expressions tuned by a group of adjustable parameters. These expressions can be divided into bonded and non-bonded terms. In an all-atom MD simulation, every atom has an assigned atom type with associated bonded and

non-bonded parameters. In the most common representation, bonded terms contain bond, angle, proper and improper dihedral angle potentials between covalently bound particles, while the non-bonded terms describe the Coulomb and Lennard-Jones interaction between two non-bonded particles.

CHARMM (*151*) and AMBER (*152*) are two widely used atomistic force field families. While using similar terms for their interactions, the specific parameters are not compatible. The most significant difference is that CHARMM force field and extensions are typically parameterized based on a three-atom water model TIP3P (*153*), while current AMBER force fields (*154*) use a four-atom model TIP4P-D (*155*), which include an additional virtual atom to reproduce the properties of water better.

2.3.1 Bonded interactions

A standard approximation of a chemical bond is via a harmonic potential. Hereby, the potential energy of a covalent bond, V_b, depends on the distance r_{ij}, between two covalently bound atoms i and j the force constant k_{ij}^b, and a reference distance r_{ij}^0, as parameters:

$$V_b(r_{ij}) = \frac{1}{2} k_{ij}^b (r_{ij} - r_{ij}^0)^2$$

2.3.1

The potential energy of an angle V_θ is formed by three covalently bound atoms i, j, and k and is also approximated by a harmonic oscillator given a force constant k_{ijk}^θ and an equilibrium angle θ_{ijk}^0 as parameters:

$$V_\theta(\theta_{ijk}) = \frac{1}{2} k_{ijk}^\theta (\theta_{ijk} - \theta_{ijk}^0)^2$$

2.3.2

Improper dihedrals are used to model planar groups (*e.g.*, aromatic rings) to restrict plane bending and similar effects. Like the bond and angle potential, the improper dihedral potential V_{id} between atoms i, j, k, and l is also often represented by a harmonic oscillator with k_{ijkl}^ξ as force constant and a reference dihedral ξ_{ijkl}^0:

$$V_{id}(\xi_{ijkl}) = \frac{1}{2} k_{ijkl}^\xi (\xi_{ijkl} - \xi_{ijkl}^0)^2$$

2.3.3

The dihedral angle potential V_d for atoms i, j, k, and l is generally periodic. Thus, they are described with a cosine function where k_{ijkl}^φ is the force constant and φ_{ijkl}^0 the phase shift:

$$V_{id}(\varphi_{ijkl}) = k_{ijkl}^\varphi (1 + \cos(n\varphi - \varphi_{ijkl}^0))$$

2.3.4

where n is the periodicity (corresponding to the number of minima).

Since bonded terms describe the behavior of atoms that are up to three bonds away from each other, the non-bonded interactions are not applied (or with reduced interaction strength) between these atoms.

2.3.2 Non-bonded interactions

The non-bonded interactions are generally defined by pairwise interaction as a function of the distance between the atoms/particles. The Lennard-Jones potential contains an attractive component, the Van der Waals force (r^{-6} term), and a repulsive (r^{-12} term) counterpart. The Van der Waals force results from the weak non-covalent interaction between atoms and molecules due to an induced polarization of the electron density. As such, this is a part of dynamic correlation effects. The strength of the interaction falls off with the sixth power of distance. The repulsion term represents the Pauli repulsion at short distances between overlapping electron orbitals, approximated by a decay with the twelfth power. The Lennard-Jones potential V_{LJ} between atom i and j is often written as:

$$V_{LJ}(r_{ij}) = 4\varepsilon_{ij}\left(\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right)$$

<div align="right">2.3.5</div>

where ε_{ij} is the depth of the Lennard-Jones potentials and σ_{ij} is the distance at which the $V_{\mathrm{LJ}}(\sigma_{ij}) = 0$. An alternative representation is written with $C^{(6)}$ and $C^{(12)}$ parameters:

$$V_{LJ}(r_{ij}) = \frac{C_{ij}^{(12)}}{r_{ij}^{12}} - \frac{C_{ij}^{(6)}}{r_{ij}^{6}}$$

<div align="right">2.3.6</div>

The parameters ε_{ij} and σ_{ij} can be either predefined for every individual interaction between two atom types or constructed from combination rules of atom type specific parameters. The advantage of the first approach is the high flexibility in representing the attractive Van der Waals interactions, but it involves more parameters. This strategy is used for the (coarse-grained) MARTINI force field (see section 2.4). CHARMM and AMBER force fields employ the second method. Here, the σ_{ij} is calculated by an arithmetic average while ε_{ij} is constructed through the geometric average from the atom type specific counterparts:

$$\sigma_{ij} = \frac{1}{2}(\sigma_i + \sigma_j)$$

<div align="right">2.3.7</div>

$$\varepsilon_{ij} = (\varepsilon_i + \varepsilon_j)^2$$

<div align="right">2.3.8</div>

The electrostatic interactions V_{el} between two atoms i and j depend on their assigned partial charge q:

$$V_{el}(r_{ij}) = \frac{1}{4\pi\varepsilon_0}\frac{q_iq_j}{\varepsilon_r r_{ij}} \qquad\qquad 2.3.9$$

with ε_0 as vacuum permittivity and ε_r as the relative permittivity. The relative permittivity is commonly set to 1 for atomistic simulations.

The non-bonded interactions are the calculation-intensive part of an MD simulation since the number of atom pairs scales quadratically with the number of atoms N ($O(N^2)$). Consequently, various approximations have been developed to decrease the scaling of the problem. The most straightforward method employs a cut-off distance r_c and discard all atom pairs with a distance above the cut-off threshold. Typical r_c values are between 10 to 15 Å for (atomistic) force fields. To avoid an abrupt change of energy and force at the cut-off, the potential can be shifted by a constant setting $V_{el/LJ}(r_c) = 0$. This approach is adequate for the Lennard-Jones potential due to the rapid decay of its individual terms. In sharp opposition, the contribution of the electrostatic interaction beyond a typical cut-off is still substantial.

A possible alternative ansatz is the reaction field approximation where a constant dielectric environment with a relative dielectric constant ε_{rf} beyond the cut-off threshold is used. For that, equation 2.3.9 is modified to:

$$V_{rf}(r_{ij}) = \frac{1}{4\pi\varepsilon_0}\frac{q_iq_j}{\varepsilon_r r_{ij}}\left(1 + \frac{\varepsilon_{rf}-\varepsilon_r}{2\varepsilon_{rf}+\varepsilon_r}\frac{r_{ij}^3}{r_c^3}\right) - \frac{1}{4\pi\varepsilon_0}\frac{q_iq_j}{\varepsilon_r r_{ij}}\frac{3\varepsilon_{rf}}{2\varepsilon_{rf}+\varepsilon_r} \qquad 2.3.10$$

For $\varepsilon_{rf} = \varepsilon_r$, equation 2.3.10 is reduced to equation 2.3.9. The right side of equation 2.3.10 ensures that the potential energy is zero at the cut-off distance point.

More precise but computationally more expensive is to incorporate the long-range electrostatic interaction explicitly with the Particle Mesh Ewald approximation (*156*) (PME). This strategy reduces the scaling from $O(N^2)$ to $O(N \cdot \log(N))$. Hereby, the long-range electrostatic potential is calculated in the reciprocal space on a discrete mesh grid. The short-range potential is modified with the complementary error function:

$$V_{el}(r_{ij}) = \frac{1}{4\pi\varepsilon_0}\frac{q_iq_j}{\varepsilon_r r_{ij}}\cdot\frac{2}{\sqrt{\pi}}\int_{r_{ij}}^{\infty}\exp(-u^2)\,du \qquad 2.3.11$$

This approach is widely utilized together with atomistic force fields. Despite the better scaling, the PME and electrostatic interaction remain the most expensive part of atomistic MD simulation.

2.4 MARTINI coarse-grained force field

While all-atom MD simulations are nowadays routinely used for million atom systems and able to characterize the dynamics at the microsecond time scale, these simulations are inherently computationally demanding. For instances where more sampling or larger system sizes are required, and a full atomistic view is not needed to address the scientific question, using a coarse-grained molecular dynamics (CG) representation is a practical alternative. In this approach, groups of atoms are clustered into beads to be used as input for MD simulations. In principle, this approach works well for groups characterized by unspecific, isotropic interactions.

The MARTINI (*8, 157*) force field has been widely adopted due to its support for various types of biomolecules and its simple application in GROMACS (*140*). Originally, it was developed for lipids, an easy target for coarse-graining due to their high ratio of unspecific van der Waals interactions. Later, it was expanded to describe proteins (*7*) as well as DNA and RNA (*158*). MARTINI maps four heavy (non-hydrogen) atoms onto one bead, except for ring systems. Specifically, the MARTINI water bead represents four water molecules with a single interaction center, which interacts solely through Lennard-Jones interactions with other beads. Hence, it is not able to capture a variety property of actual water.

Overall, the number of interaction centers is reduced by an approximate factor of 10, resulting in a much smaller number of non-bonded interactions. Less obvious is that the removal of the hydrogen atoms enables the use of much larger time steps since the fastest motions of the coarse-grained systems are much slower. Therefore, time steps between 20 to 40 fs are used in MARTINI, compared to 2 fs in the all-atom MD (see section 2.2.1).
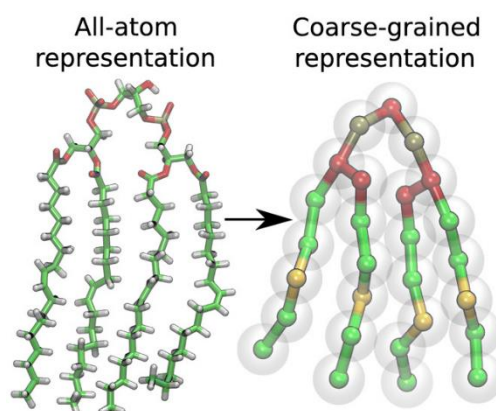

Figure 2.1 Atomistic and coarse-grained representation of cardiolipin. The figure is adapted from publication CI.

MARTINI beads do not have partial charges. Thus, there are no explicit descriptions of many electrostatic effects, such as the dipole of an amino acid

backbone. Only the ions and ionic groups carry a charge. Long-range electrostatic interactions are ordinarily ignored in MARTINI, using either a simple cut-off or the reaction field approximation (*159*). Additionally, a relative dielectric constant of 15 is used to screen charges, making explicit electrostatic terms even less significant. Consequently, most non-bonded interactions are parameterized via the Lennard-Jones potential term, whose parameters implicitly include the short-range electrostatic interactions.

In contrast to atomistic force fields, where the Lennard Jones parameters are constructed with combination rules based on the parameters associated with the respective atom types, the Lennard-Jones interactions in MARTINI are divided into ten levels. These represent different interaction strengths determined by the involved bead types with an ε_{ij} ranging from 2.0 to 5.2 kJ/mol and a σ_{ij} of 0.47 nm. The bead types are separated into polar, intermediate, apolar, and charged types based on the importance of the local electrostatic interactions. For computational efficiency, the particles have the same mass of 72 u. The exception is ring particles, which use a smaller σ_{ij} of 0.42 nm, a smaller mass of 45 u, and 25% lower interaction strength than regular bead types due to their different mapping.

The MARTINI beads are parameterized to reproduce their partitioning between water and apolar organic solvents. Consequently, the parameterization strategy is based on the reproduction of free energies. However, the grouping of atoms results in the loss of internal degrees of freedom, leading to a reduction of the system's entropy, which has to be compensated by the enthalpy term (mainly due to the strength of interactions). Therefore, the entropy-enthalpy balance is not correct, which primarily affects temperature-sensitive properties.

The bonded interaction terms are similar in structure to those in the atomistic force fields. The notable exception is the recommended use of a restricted bending potential (*160*) V_{ReB} as a substitute for a standard harmonic angular potential (see equation). The goal is to prevent the bending angle from reaching 180°, avoiding numerical instability and inaccuracy:

$$V_{ReB}(\theta_{ijk}) = \frac{1}{2} k_{ijk}^{\theta} \frac{\left(cos\,\theta_{ijk} - cos\,\theta_{ijk}^0\right)^2}{sin^2\,\theta_{ijk}} \qquad 2.4.1$$

Protein residues are represented by one bead for the backbone and one to three beads for the side chain (zero for GLY and ALA). The bead type of the protein backbone bead and the bonded interactions with the neighboring amino acids depend on the secondary structure. Therefore, the MARTINI force field cannot describe transitions in secondary structure and cannot be used to study protein folding or intrinsically disordered proteins.

Additional steps are often required to stabilize the tertiary structure of proteins. The simplest option is the use of an elastic network (*161*). Thereby, the secondary and tertiary structure is stabilized by introducing additional harmonic restraints between backbone beads based on the initial conformation:

$$V_{el}(\boldsymbol{r}) = \frac{1}{2}k_{el} \sum_{r_{ij}^0 < r_c} \left(r_{ij} - r_{ij}^0\right)^2$$

<div align="right">2.4.2</div>

where V_{el} is the total potential energy of the elastic network, k_{el} is the force constant for the harmonic bond, and r_c is the cut-off radius of the elastic network.

A pervasive problem associated with the MARTINI representation of proteins results from an overestimation of protein-protein interaction versus protein-solvent (and protein-membrane) interactions (*162-164*), leading to excessive formation of compact conformation and aggregation simulations with multiple proteins. For publication CI and HI, we used a separate bead type for water molecules to tune the protein-water Lennard-Jones interaction without modifying other variables.

Recently, MARTINI3 (6) was released, enhances MARTINI in several areas by expanding the amount and reparametrizing of bead types and improving the molecular packing, resulting in better protein-protein interactions.

2.5 QM/MM

Electronic structure methods and molecular dynamics simulation have contrasting application profiles. While QM methods allow the characterization of chemical reactions and quantum effects, their high computational cost limits their utility to small system sizes. In contrast, MD simulations can characterize the dynamics of large systems but are intrinsically limited to a fixed chemical state. The hybrid QM/MM method combines both approaches by treating the active site with QM methods and the surrounding as classical particles.

Hereby, the potential energies V_{tot} can be expressed as the sum of the potential energies of the QM system, V_{QM}, the MM energy, V_{MM}, and a third term resulting from the interaction between the QM and MM regions $V_{QM/MM}$:

$$V_{tot} = V_{QM} + V_{MM} + V_{QM/MM}$$

<div align="right">2.5.1</div>

The $V_{QM/MM}$ potential can be split in a Lennard-Jones and electrostatic interaction between the QM and MM regions. As such, the first part can be expressed similarly to the equation 2.3.5. In contrast, the electrostatic interactions are a more complicated issue since the surrounding charges are able to polarize the QM electron density (and vice-versa). Several approaches can deal with this problem,

but the electrostatic embedding approach (*165*), as used in the CHARMM/TURBOMOLE interface (*166*), is the most common. Thereby, the MM atomic point charges interaction is explicitly incorporated in the QM Hamiltonian.

Optimally, there are no covalent bonds between the QM and the MM region. However, this cannot be avoided, for example, in protein systems. A common approach is the usage of link atoms (*167, 168*). Herby, a virtual hydrogen atom, is added to the bond between QM and MM region, and the forces are distributed between both regions. As a result, the QM region becomes saturated. Nevertheless, the local electronic structure in the boundary region might still deviate from reality. Additionally, the QM region might be overpolarized by the point charges close to the QM region. This can be avoided by redistributing them among the closest atoms (*165, 169*).

2.6 Principal Component Analysis

The analysis of high-dimensional data, such as simulation trajectory, can be extremely challenging. While it is relatively easy to characterize the motion of specific side chains with geometrical functions or basic collective variables, more sophisticated techniques are required to describe the collective motions of a protein against a background of random thermal fluctuations. To this end, the *principal component analysis* (PCA), a multivariable statistical approach used for dimension reduction, is a powerful tool for extracting the components with the highest variance (*170, 171*). For this purpose, the covariance matrix C defined based on the variable vector $\boldsymbol{X}$ is used:

$$\mathsf{C}(\boldsymbol{X}) = \; \langle (\boldsymbol{X} - \langle \boldsymbol{X} \rangle) \cdot (\boldsymbol{X} - \langle \boldsymbol{X} \rangle)^T \rangle \qquad\qquad 2.6.1$$

where the brackets $\langle \dots \rangle$ mark the average.

The covariance matrix is then diagonalized to extract the eigenvalues and eigenvectors. The latter ones are the principal components representing the dimensions with the highest variance. The eigenvalues correspond to the proportion of the total variance represented by the individual principal components. Generally, the eigenvalues decrease drastically after the initial eigenvectors. If the protein coordinates are used as an input, the PC components represent the dimensions with the largest conformational changes in the protein (*170*).

2.7 Metadynamics

Atomistic molecular dynamics (aMD) simulation are suited to describe processes in the ns-µs timescale, while coarse-grained molecular dynamics (cgMD) simulation can reach µs-ms timescales. Based on the Eyring equation, these timescales can describe processes with an activation barrier of up to ~10 kcal/mol for aMD and up to ~14 kcal/mol for cgMD simulations. However, many biological processes occur on a longer time scale or exhibit a higher activation barrier. Therefore, methods to improve the sampling of MD simulations are often necessary.

Metadynamics (*13, 14, 172*) is a well-established enhanced sampling method that accelerates the exploration of conformational space (e.g., for proteins). It involves applying an additional history-dependent bias potential term, consisting of a sum of Gaussians for a chosen collective variable (CV), to the free energy surface defined by the force field. For a given height w and a given width σ of the used Gaussians, the bias potential B, for a given time t and the value of the CV s, is defined as:

$$B_{t+1}(s) = B_t(s) + w \cdot exp\left(-\frac{(s_t - s)^2}{2\sigma^2}\right)$$
2.7.1

A constant Gaussian width may not be optimal, especially for CVs, where the fluctuation can vary depending on the specific state of the system. In these cases, an *adaptive Gaussians* (*173*) scheme can be used, where the Gaussians width is adjusted on the fly to match the local properties of the free-energy surface. Thus, broader or narrower Gaussian potentials are applied depending on the dynamics of the system.

The bias potential accelerates sampling of the conformational space by discouraging regions that the simulation has previously explored. The deposited biases B allow for a reconstruction of the underlying free energy profile F (+ an additive constant C) given by:

$$\lim_{t\to\infty} B_t(s) = F(s) + C$$
2.7.2

However, due to the deposition of Gaussians with constant height, the free energy does not converge in a finite time, but rather constantly fluctuates around the actual value. In addition, the steadily increasing bias potential can push the system into physically irrelevant regions of conformational space. This limitation can be overcome by using the *well-tempered metadynamics* (*174*) approach. Here, the height of the newly added Gaussians is scaled as a function of the previous bias. As such, equation 2.7.1 is modified to:

$$B_{t+1}(s) = B_t(s) + w \cdot exp\left(-\frac{B_t(s)}{\Delta T}\right) exp\left(-\frac{(s_t - s)^2}{2\sigma^2}\right)$$
2.7.3

ΔT is often integrated into the bias factor ɣ depending on the temperature *T*:

$$\gamma = \frac{T + \Delta T}{T} \qquad\qquad 2.7.4$$

Hence, ΔT or the bias factor controls the decay of the Gaussian height and guarantees that the bias potential converges to a fraction of the free energy $-\frac{\Delta T}{T + \Delta T} F(s)$ *(174)*.

The quality of the sampling depends on the choice of the set of CVs for the simulation. Ideally, a small number of CVs captures the slowest modes of the system, as it becomes computationally more demanding to reconstruct the free energy surface in more dimensions. Nevertheless, it can be challenging to select a small number of CVs that still accurately describe the essential degrees of freedom of the system (*172*). Different methods have been proposed to deal with a larger number of CVs. In *bias-exchange metadynamics* (*175*), a simulation with multiple replicas is performed where each replica biases a single CV with regular exchanges between the bias potential of the replicas. An alternative approach is *Parallel Bias metadynamics* (*176*) (PBMetaD), where multiple low-dimensional bias potentials are applied simultaneously. Here, the magnitude of the individual Gaussians, as well as the total bias, depends on biases of all CV:

$$w_{i,t} = w_{i,0} \cdot \frac{\exp\left(-\frac{B_{i,t}(s)}{k_B T}\right)}{\sum_{j=1}^{N} \exp\left(-\frac{B_{j,t}(s)}{k_B T}\right)} \qquad\qquad 2.7.5$$

$$B_t(s) = -k_B T \cdot \log \sum_{i=1}^{N} \exp\left(-\frac{B_t(s_i)}{k_B T}\right) \qquad\qquad 2.7.6$$

with k_B as Boltzmann constant.

An additional speed-up can be achieved using the *multiple walkers* (*177*) approach where multiple simulations (walkers) contribute to a global history-dependent potential and, in turn, used to explore and reconstruct the free energy surface.

PBmetaD with *multiple walkers* and *adaptive Gaussians* was used for publication SI and SIII to study linear and Lys63-linked polyubiquitin.

2.8 Metainference

2.8.1 Incorporation of experimental data into simulations

Determining accurate conformational ensembles for dynamic systems, e.g., multi-domain or intrinsically disordered proteins, requires an accurate model combined with extensive sampling. In many cases, atomistic force fields may not provide an accurate enough representation to describe such systems quantitatively. This

limitation can be overcome by integrating experimental data. For this to be possible, several sources of error affecting the experimental data and their interpretation must be accounted for: First, the experimental results can be susceptible to random noise and thus to statistical and systematic errors. Second, the computational method used to calculate the experimental observable from a structure ("forward model") is typically an approximation. Finally, the use of a finite number of structures to compute the measured quantity introduces an averaging error.

Metainference (21, 99, 178) is a replica-based Bayesian framework to reconcile a prior knowledge (the force field used) with auxiliary information (the experimental data) and associated errors to form a modified posterior probability distribution (*P*). The energy E_{MI} related to posterior distribution is defined as:

$$E_{MI} = -k_B T \cdot ln(P) \qquad 2.8.1$$

with k_B as the Boltzmann constant and T as the temperature of the system.

For a finite number of replicas, N_r, P can generally be expressed as:

$$P\left(\tilde{\boldsymbol{f}}, \boldsymbol{\sigma}^B, \boldsymbol{X}, \boldsymbol{\sigma}^{SEM}\middle|\boldsymbol{d}\right) =$$

$$\prod_{r=1}^{N_r} p(X_r) \prod_{i=1}^{N_d} p\left(\sigma_{r,i}^B\right) \cdot p\left(d_i\middle|\tilde{f}_{r,i}, \sigma_{r,i}^B\right) \cdot p\left(\sigma_i^{SEM}\right) \cdot p\left(\tilde{f}_{r,i}\middle|\langle f_i(X_r)\rangle, \sigma_i^{SEM}\right) \qquad 2.8.2$$

where $p(X_r)$ is the prior (the force field) depending on the conformation X_r (e.g., of a protein), and N_d is the number of experimental data points. $p\left(\sigma_{r,i}^B\right)$ is the prior for the experimental and forward model error depending on the uncertainty parameter $\sigma_{r,i}^B$. $P\left(d_i\middle|\tilde{f}_I, \sigma_{r,i}^B\right)$ is the conditional probability of the experimental data d_i depending on the uncertainty parameter $\sigma_{r,i}^B$ and the (estimated) average of the forward model $\tilde{f}$ over an infinite number of replicas. $p\left(\sigma_i^{SEM}\right)$ describes the scaling of the standard error of the mean with σ_i^{SEM} as the uncertainty parameter, which is proportional to $1/\sqrt{N_r}$. $p\left(\tilde{f}_i\middle|\langle f_i(X_r)\rangle_{N_r}, \sigma_i^{SEM}\right)$ is the probability of observing a specific average $\tilde{f}$ depending on the finite average $\langle f_i(X_r)\rangle_{N_r}$ and the uncertainty parameter σ_i^{SEM}.

Based on the central limit theorem, $p\left(\tilde{f}_i\middle|\langle f_i(X_r)\rangle_{N_r}, \sigma_i^{SEM}\right)$ can be represented as a gaussian distribution, which allows simplifying equation 2.8.2 under the assumption of the Jeffreys prior for $p\left(\sigma_{r,i}^B\right)$ and $p\left(\sigma_i^{SEM}\right)$:

$$E_{MI}(\boldsymbol{X}, \boldsymbol{\sigma}|\boldsymbol{d}) = \sum_{r=1}^{N_r}\left\{E_{ff}(X_r) + k_B T \sum_{i=1}^{N_d}\left[\frac{(\langle f_i(X_r)\rangle - d_i)^2}{2\sigma_{r,i}^2} + \frac{1}{2}ln(2\pi\sigma_{r,i}^2) + \frac{1}{2}ln\left(\frac{\sigma_{r,i}^2}{2}\right)\right]\right\} \quad 2.8.3$$

with $E_{ff}(X_r)$ as the force field energy and $\sigma_{r,i}$ as the total uncertainty defined as:

2 Methods

$$\sigma_{r,i} = \sqrt{\left(\sigma_{r,i}^B\right)^2 + \left(\sigma_i^{SEM}\right)^2}$$ 2.8.4

2.8.2 Modification in publication S-IV

In publication SIV (*178*), the *Metainference* approach was expanded to include an additional scaling factor λ since specific experimental measurements, such as residual dipolar coupling (RDC) or small-angle X-ray scattering (SAXS), are proportional to the values calculated by the forward model. Consequently, the experimental and computed values cannot be directly compared. Hence, equation 2.8.3 and 2.8.4 can be modified to:

$$E_{MI}(\boldsymbol{X}, \boldsymbol{\sigma}|\boldsymbol{d}) =$$

$$\sum_{r=1}^{N_r} \left\{ E_{ff}(X_r) + k_B T \sum_{i=1}^{N_d} \left[\frac{\left(\lambda \langle f_i(X_r)\rangle_{N_r} - d_i\right)^2}{2\sigma_{r,i}^2} + \frac{1}{2} ln(2\pi\sigma_{r,i}^2) + \frac{1}{2} ln\left(\frac{\sigma_{r,i}^2}{2}\right) \right] \right\}$$ 2.8.5

$$\sigma_{r,i} = \sqrt{\left(\sigma_{r,i}^B\right)^2 + \lambda^2 \left(\sigma_i^{SEM}\right)^2}$$ 2.8.6

Theoretically, σ_i^{SEM} should be different for each experimental data point used. However, it can be practical to estimate a shared value estimate based on the maximum variance of the forward model:

$$\sigma_i^{SEM} = \sqrt{max\left(Var[f_i](t)\right)/N_r}$$ 2.8.7

2.8.3 Metadynamics Metainference (M&M)

To facilitate enhanced sampling, Metainference can be used together with PBmetaD. In the combined metadynamic metainference (99) (M&M) approach, an ensemble of replicas is simulated with the Metainference energy function jointly with the bias potential from metadynamics, with each replica using the same set of CVs. As such, the total M&M energy ($E_{M\&M}$) can be written as:

$$E_{M\&M}(\mathbf{X}, \boldsymbol{\sigma}, t) = E_{MI}(\mathbf{X}, \boldsymbol{\sigma}) + \sum_{r=1}^{N_r} B_t(\mathbf{s}(X_r))$$ 2.8.8

Adding a bias term to the force field results in a combined potential corresponding to an ensemble that is less similar to the right ensemble. A way to avoid this is to compute the unbiased average for Metainference on the fly. Hence, a weighted average instead of an arithmetic one is used to compute the mean observables across all replicas. The weight (w) of each replica is determined by the bias potential obtained from PBmetaD:

$$w(X_r) = exp\left(\frac{B_t(s_i(X_r))}{k_B T}\right)$$

<div align="right">2.8.9</div>

The implication is that a replica located in a well-sampled region of the conformational space has a higher weight than a replica in an energetically unfavorable conformation.

M&M was used for publication SI to integrate SAXS data in coarse-grained simulations, for publication SIII to integrate SAXS data in atomistic simulations, and for SIV to study the ensemble generated by different force fields integrated with a set of experimental NMR data.

2.9 SAXS

2.9.1 Computing scattering intensities

Small-angle X-ray scattering (SAXS) is an experimental technique that allows the study of large biomolecules in solution. Although SAXS data do not provide atomistic resolution, they can shed light on the size and shape, as well as the global dynamics and interactions of a system. The calculation of SAXS intensities for structures is computationally expensive since it depends on all inter-atom distances in a system. Therefore, for a structure with N atoms, the calculation of SAXS intensities is an O(N^2) problem.

The total scattering amplitude A, from a collection of N point-like atomic scatterers, at the scattering vector $\mathbf{q}$ is calculated by (*179*):

$$A(\mathbf{q}) = \sum_{i=1}^{N} f_i \cdot e^{i\mathbf{q}\cdot\mathbf{r_i}}$$

<div align="right">2.9.1</div>

where $\mathbf{r}_i$ is the atomic position and f_i is the atomic scattering factor (or form factor). Typically, biomolecules in SAXS experiments do not have a fixed orientation but are randomly oriented in solution. In this case, equation 2.9.1 can be simplified by multiplying 2.9.1 by its complex conjugate to obtain the scattering intensity (I):

$$I(\boldsymbol{q}) = \sum_{i=1}^{N} f_i \cdot e^{i\boldsymbol{q}\cdot r_i} \sum_{j=1}^{N} f_j \cdot e^{-i\boldsymbol{q}\cdot r_i} = \sum_{i=1}^{N}\sum_{j=1}^{N} f_i \cdot f_j \cdot e^{-i\boldsymbol{q}\cdot r_i}$$

<div align="right">2.9.2</div>

This equation can be transformed to the Debye equation(*180*) by simply taking the spherical average (marked as ⟨...⟩) with $q = |\boldsymbol{q}| = 4\pi \sin\theta/\lambda$, λ as the X-ray wavelength, $\mathbf{r}_{ij}$ as the vector between atom i and j and $r_{ij} = |\boldsymbol{r}_{ij}|$:

$$I(q) = \langle I(\boldsymbol{q}) \rangle = \langle \sum_{i=1}^{N} \sum_{j=1}^{N} f_i \cdot f_j \cdot e^{i\boldsymbol{q} \cdot \boldsymbol{r}_{ij}} \rangle = \sum_{i=1}^{N} \sum_{j=1}^{N} f_i(q) \cdot f_j(q) \frac{sin(qr_{ij})}{qr_{ij}} \qquad \text{2.9.3}$$

The Debye equation is valid for a single molecule in a vacuum. The maximum value is reached for $q = 0$ Å^{-1} where the scattering intensity depends solely on the atomic form factors. In turn, these scattering factors depend on the specific atoms and represent the Fourier transformation of the electron density around them. For $q = 0$ Å^{-1}, the form factors are equal to the atomic number (or the number of electrons for form factors of ions). Because of that, SAXS measurements can be used to determine the molecular weight of a biomolecule. In general, the atomic scattering factors are derived by the Cromer-Mann analytic function, which fits the result of the Fourier transformation as a sum of multiple Gaussians:

$$f_i(q) = \sum_{k=1}^{4} a_k e^{-b_k(q/4\pi)^2} + c \qquad \text{2.9.4}$$

with a_k, b_k, and c as Cromer-Mann coefficients as fitting coefficients (*181, 182*).

Since SAXS measurements on biomolecules are usually performed in a buffer solution, the resulting unwanted scattering intensities of the buffer are subtracted from the sample scattering. Therefore, the forward model must account for the displaced solvent. A commonly used approximation uses corrected form factors that assume the solvent has a uniform electron density (*183*). These reduced scattering factors $f_i'(q)$ can be expressed as:

$$f_i'(q) = f_i(q) - V_i \rho_b e^{-q^2 v_i^{2/3}/4\pi} \qquad \text{2.9.5}$$

with V_i as the solvent volume of displaced by atom i and ρ_b as the electron density of the solvent (for bulk water, the value is $0.334e$ Å^{-3}). However, due to the solute-solvent interaction, the electron density in the hydration sphere may differ compared to bulk solvent. Several approaches have been proposed to address this issue: Either the solvation shell can be modeled implicitly as a homogeneous border layer (e.g., in CRYSOL (*184*)) or, more elaborately, with an explicit solvent approach (e.g., in Chen *et al.* (*185*)).

The experimental SAXS data can be further processed to extract the properties of biomolecules. An often-used representation is the *Kratky plot (186)*, $I(q) \cdot q^2$ vs q, which can be used to obtain information about the shape of the biomolecule. In addition, the *Guinier plot (187)* is often applied to determine the radius of gyration (R_g), which gives an estimation of the size of the biomolecule. For small q values ($qR_g < 1.1$), the following equation connects the R_g with scattering intensity I(q):

$$I(q) \simeq I(0) \cdot \exp\left(-\frac{q^2 R_g^2}{3}\right) \qquad \text{2.9.6}$$

By plotting $\ln[I(q)]$ vs. q^2, equation 2.9.6 becomes linearized, and the R_g can be extracted over the slope.

Furthermore, the scattering intensity profile can be transformed into the pair distance distribution ($P(r)$), which contains the distribution of all intermolecular distances of a specific biomolecule (as well as the maximum distance of the system). Hence, the $P(r)$ allows for a more straightforward interpretation of the results of the scattering experiments. $I(q)$ and $P(r)$ are closely related:

$$I(q) = \int_0^{d_{max}} P(r) \frac{sin(qr)}{qr} dr \qquad 2.9.7$$

where d_{max} is the maximum distance of the system. Via the inverse transformation, $P(r)$ can be calculated by:

$$P(r) = \frac{r}{2\pi^2} \int_0^\infty q \cdot I(q) \cdot \sin(qr) \, dq \qquad 2.9.8$$

However, since experimentally measured I(q) is prone to statistical errors, the range for q is limited, and q-values are only available at discrete points, the obtained $P(r)$ might be vulnerable to errors (*188*).

2.9.2 Form factors for a Martini coarse-grained representation

Since the solution of the Debye equation is an O(N^2) problem, the computational cost of the problem becomes high for larger systems. Several strategies have therefore been employed to reduce the scaling of the problem. One commonly implemented strategy uses expansions of spherical harmonics (as adopted in CRYSOL (*184*)); other approaches involve a hierarchical algorithm (*189*) or the particle mesh Ewald approach (*190*). Alternatively, the Debye formula can be employed with a coarse-grained representation of the biomolecule (*191-193*) since reducing the number of scattering centers can lead to a substantial speed-up. This approach is also justified since SAXS contains low-resolution information predominantly.

Such an approach requires combining atomic form factors to coarse grained scattering factors $F_i(q)$ which can replace the atomic form factors in the Debye equation 2.9.3. As boundary condition for $q = 0 \, \text{Å}^{-1}$, the sum of the (reduced) atomic scatting factors must be equal to the combined (coarse-grained) form factor:

$$F_i(0) = \sum_{k \in M} f_k'(0) \qquad 2.9.9$$

One of the most straightforward approaches to construct $F_i(q)$ is the single bead approximation (*191*). Here, the combined scattering factors $F_I(q)$ are calculated as the square root of the scattering intensity defined by the M atoms incorporated by the coarse grain bead:

$$F_i(q) = \sqrt{\sum_{k \in M} \sum_{l \in M} f_k'(q) f_l'(q) \frac{sin(qr_{kl})}{qr_{kl}}}$$

2.9.10

with $f'(q)$ as reduced atomic scattering intensity as defined in equation 2.9.5. The result of equation 2.9.10 is always positive. However, suppose a bead represents atoms with a lower electron density than the solvent (e.g., due to a high number of hydrogen atoms), then the combined scattering factors should be negative for $q = 0$ Å^{-1}, according to equation 2.9.9. To correct the affected form factors, the use of a sixth-order polynomial fit for the positive high-q data point constraint at $q = 0$ Å^{-1} was suggested in Niebling *et al.* (*192*).

In publication SII, we utilize this approach to calculate combined form factors for DNA and RNA for the Martini force field, building on the Martini protein form factors in Niebling *et al.* (*192*). We employed SAXS constraints with M&M in S-I and SIII to increase the accuracy of our cgMD and aMD simulations of polyubiquitin chains. In publication H1, we used coarse grained and atomistic simulations to rationalize the SAXS scattering data and $P(r)$ profiles of Hsp90.

3 Results

3.1 Complex I (C-I – C-IV)

Own contribution: In publication C-I, I performed all cgMD simulations, was involved in the project design, analysis and interpretation of the computational results, and manuscript creation. In paper C-II, I performed QM cluster model calculations and QM/MM simulations on the ubiquinone setups. In paper C-III and C-IV, I performed cgMD simulations. In both papers, I was involved in analyzing and interpreting the computational results and manuscript creation.

3.1.1 How cardiolipin modulates the dynamics of respiratory Complex I

Respiratory Complex I (NADH:ubiquinone oxidoreductase) is a prominent member of the respiratory chain responsible for energy conversion in mitochondria and bacteria (*27-29*). This enzyme harnesses the released energy from electron transfer between nicotinamide adenine dinucleotide (NADH) and quinone (Q) to pump protons across membranes (*34, 35*). Cardiolipin, an anionic lipid accounting for about 20% of the mitochondrial membrane, has been reported to regulate the activity of Complex I (and other respiratory enzymes) (*67-71*). However, the molecular mechanism of this process has remained unclear. In publication C-I, we used coarse-grained molecular dynamics (cgMD) simulations on millisecond time scales to determine how cardiolipin binding modulates the global conformational dynamics of Complex I. In turn, these global motions control the accessibility of quinone to the catalytic active site.

We performed cgMD simulations with the bacterial Complex I embedded in a membrane composed equally of POPC (palmitoyl-oleoyl-phosphatidylcholine) and POPE (palmitoyl-oleoyl-phosphatidylethanolamine) with or without 20% cardiolipin. Our results suggest that Complex I is surrounded by approximately 150 lipid molecules, of which ~45 are cardiolipin that interacted with positively charged residues of the enzyme. The cardiolipin lipids are concentrated around Nqo8, the quinone channel-containing subunit that connects the membrane and the hydrophilic domain. In addition, we observe binding sites at the interface between the antiporter-like subunits Nqo12/13 and Nqo13/14 near the proton channels. In contrast to POPE and POPC, cardiolipin forms much stronger interactions with Complex I. While POPC/POPE has an average retention time of about 80 ns, the retention time is ~600 ns for cardiolipin and specifically 1.2 µs for cardiolipin around Nqo8. The predicted cardiolipin binding sites are observed in recent structures cryo-EM of Complex I, validating our findings (**Figure** 3.1).
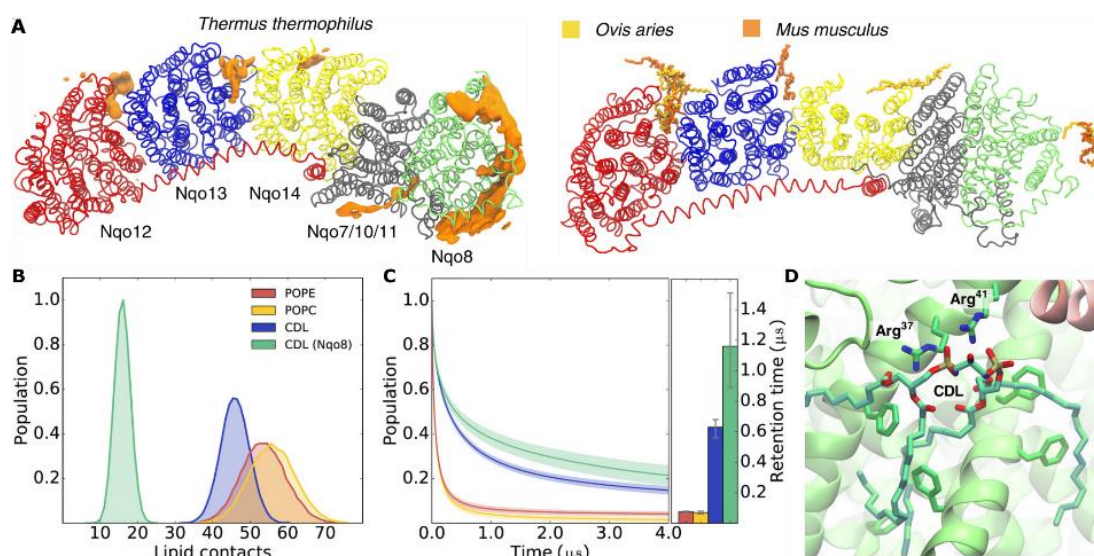
Figure 3.1 **A)** Left: Cardiolipin (CDL) binding sites identified by cgMD simulations of Complex I from *Thermus thermophilus*. Right: Experimentally refined CDL molecules around cryo-EM structures of Complex I from *Ovis aries* (PDB ID: 5LNK) and *Mus musculus* (PDB ID: 6G2J). **(B)** Lipid contact statistics for the membrane domain of Complex I (from cgMD simulations). **(C)** lipid-binding autocorrelation. **(D)** CDL binding sites in Nqo8, obtained from atomistic simulations. Figure modified from publication C-I.

Using principal component (PC) analysis, we analyze the large-scale dynamics of Complex I based on the cgMD simulations. We observe large-scale bending and twisting motion in Complex I, with Nqo8 acting as the hinge region. Similar motions have also been captured in the cryoEM classification and associated with the *active* to *deactive* transition of mammalian Complex I (*31, 33, 65, 66*). Our results indicate that the bending-twisting movement is associated with quinone dynamics in its substrate tunnel and strongly affected by the presence of cardiolipin. Our cgMD simulations identify three stable quinone binding sites within the cavity, the active site at the top (d ~ 0.5 to 0.7 nm away from Tyr87 of Nqo8), a binding site slightly below (d ~ 0.7 to 0.9 nm), and one in the lower part (d ~ 3.2 to 3.7 nm). These regions are similar to sites observed in previous atomistic studies and later also validated experimentally (see below, publication C-III). In the presence of cardiolipin, the quinone motion along the substrate tunnel is favored by the twisting and bending modes observed in the PCA. Interestingly, Complex I shows similar conformations in the *apo* state as well as when the quinone is bound at the lower part of the substrate tunnel, suggesting that no large-scale conformational changes occur upon substrate entry. However, when the Q passes a highly bent kink region (d ~ 2 nm of Tyr87), the hydrophilic domain rotates relative to the membrane domain, followed by a bending motion (see **Figure** 3.2). However, in simulations without cardiolipin, we did not observe similar coordinated conformational changes in Complex I, suggesting that the lipid could modulate the large-scale dynamics of the enzyme.

We also observe that cardiolipin stabilizes, in addition to the quinone tunnel, also two channels suitable for the transfer of protons via water molecules. One of the tunnels leads from the matrix side to the active site of quinone reduction and thus may be responsible for the reprotonation of residues His38 and Tyr87 of Nqo4, which act as proton donors during the quinone reduction to QH_2. A second channel forms at the Nqo8/Nqo9 interface and leads to the kink region of the quinone tunnel. Both channels are open in around 20-30% of the simulations with cardiolipin and quinone in the active site, compared to only 0.4%-3% without cardiolipin.
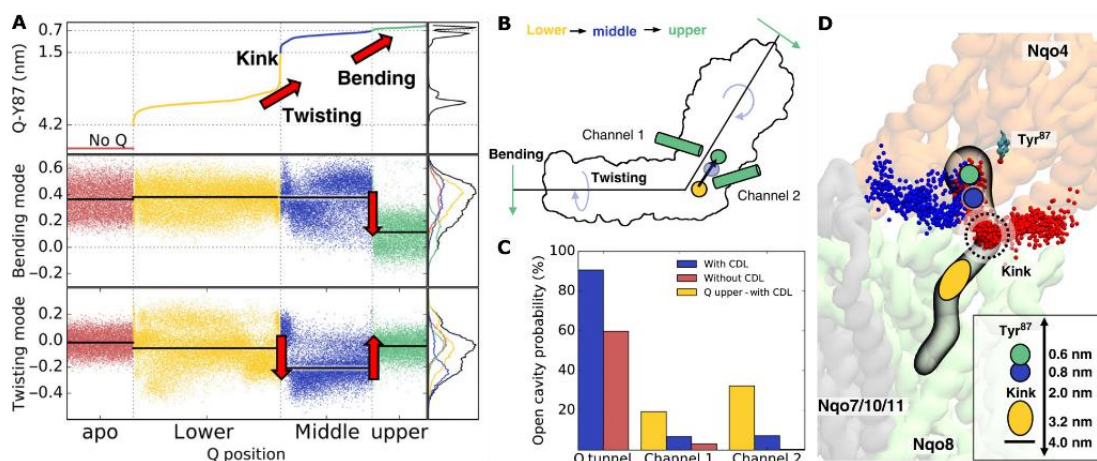
Figure 3.2 (**A**) Projection of complex I dynamics with CDL along the bending and twisting modes obtained by PCA sorted by the quinone (**Q**) position along the tunnel. (**B**) Schematic representation of structural changes in Complex I that couple to the quinone motion. (**C**) Probability of finding open channels in Complex I. (**D**) Quinone binding sites and channels linking the quinone cavity with the N-side of the membrane. Figure modified from publication C-I.

3.1.2   Quinone dynamics

The reduction of (ubi)quinone (Q) is a pivotal step in the catalytic cycle of complex I, a proton pump in the respiratory chain. The active site of Q-reduction is centered between a conserved tyrosine (Tyr-87 of Nqo4) and histidine (His-38 of Nqo4) about 20 Å above the membrane plane and 10 Å away from the terminal N2 iron-sulfur cluster (ISC) (*28, 42, 49*). The latter is responsible for electron transfer (eT) to a quinone. In publication C-II, we used density functional theory (DFT) cluster models, hybrid QM/MM simulations, and atomistic MD (aMD) simulations based on the crystal structure of bacterial Complex I from *Thermus thermophilus* to characterize the binding modes of ubiquinone. In addition, we examine the conformational changes upon Q-reduction.

We performed DFT calculations based on cluster models comprising the quinone and oxidized/reduced N2 with surrounding residues. Our findings show that Q can adopt two different binding modes, defined by either a hydrogen bond or a

stacked conformation between Q and His-38. For the oxidized Q (with reduced N2), the stacked conformation is ~2 kcal/mol more stable than the hydrogen-bonded one. In contrast, the reduction of Q to semiquinone (SQ) leads to a more favorable hydrogen-bonded conformation. While the hydrogen-bonded SQ (with oxidized N2) is stabilized by -9 kcal/mol (relative to the oxidized Q / reduced N2 state), the stacked bond form of SQ is favored by only -4 kcal/mol. Complementary continuum Poisson-Boltzmann (PB) electrostatic calculations indicate that the Q/SQ redox couple has a redox potential of -260 mV in the hydrogen bonding conformation, compared to -380 mV for the stacked conformation. Our QM/MM simulations are consistent with these outcomes: We observe a fast eT transfer between a (reduced) N2 ISC cluster and an (oxidized) Q in the hydrogen bond conformation, while the SQ formation is slow in the stacked conformation. Moreover, our QM/MM simulations suggest that the second eT to SQ is coupled with proton transfer from Tyr-87 and His-38, leading to the formation of ubiquinol (QH2) (**Figure** 3.3).
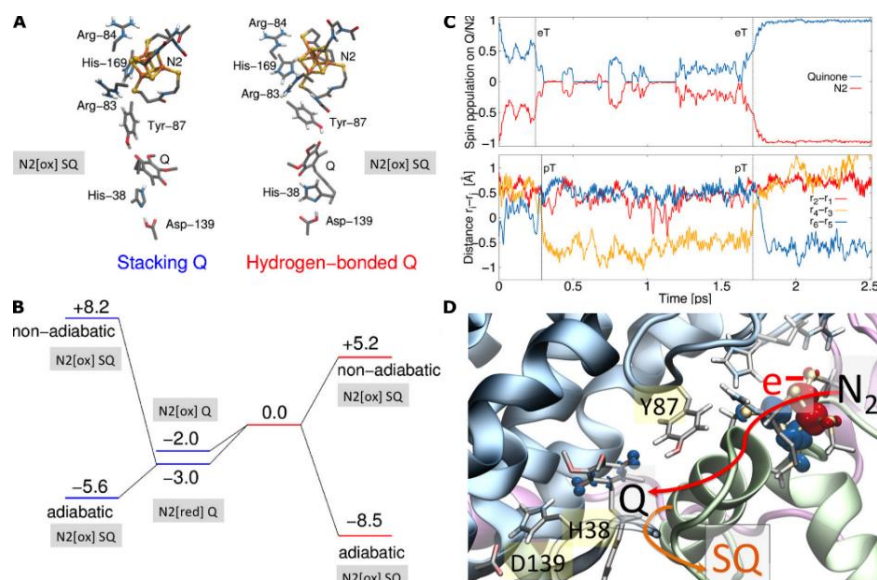
Figure 3.3: (**A**) Structure and (**B**) energetics (in kcal/mol) of hydrogen-bonded (right) and stacked (left) configuration of UQ in complex I. (**C**) Dynamics of the second eT from N2 to SQ (top) upon deprotonation of His-169 and Tyr-87 (partial deprotonation at 1 ps, below). (**D**) First eT from N2 to UQ in the hydrogen-bonded conformation forming SQ. Figure modified from publication C-II.

While there is solid computational support for Q-binding, little structural data is available for (ubi)quinone or quinone-like inhibitors bound to Complex I. In publication C-III, we present a new 3.0 Å resolution cryo-EM structure of the mammalian complex I with bound piericidin A (P), a quinone-like inhibitor of Complex I, and use experimental and computational techniques to rationalize its binding within the ubiquinone cavity.

Our cryo-EM map shows a density suitable for piericidin at the top of the ubiquinone binding cavity. The head group is in proximity to residues Tyr108 and

His58 of NDUFS2 (Tyr87 and His38 of Nqo4 are the equivalent residues in *Thermus thermophilus* used as a model for publication C-I and C-II). Interestingly, kinetic experiments suggest that the mammalian Complex I likely binds a second piericidin in a two-site competitive inhibition model. Indeed, we observe a noisy cryoEM density in the lower part of the ubiquinone cavity, implying that two piericidins may occupy the same channel. To obtain a more detailed picture of piericidin and quinone binding, we performed atomistic MD, cgMD, and QM/MM simulations. We find that piericidin forms a stable hydrogen bond with Tyr108, together with a van der Waals (VdW) stabilized interaction with the His59 ring, similar to the binding mode described in publication CII. Additional coarse-grained MD simulations show that piericidin remains in the binding pocket for at least ten microseconds. Our aMD and cgMD simulations also suggest that a second piericidin may indeed be responsible for density in the lower part of the ubiquinone cavity. The second piericidin might be able to form a transient hydrogen bond with the upper piericidin, and our simulations further suggest that the piericidin in the second binding site is indeed much more flexible than the first, consistent with the noisy cryoEM density (**Figure** 3.4). For publication C-IV, we performed additional aMD and cgMD simulations on the mammalian complex I using Q instead of, where we observed a similar binding sites for both substrates. The binding regions are also consistent with the one observed in *Thermus thermophilus* (publication C-I).
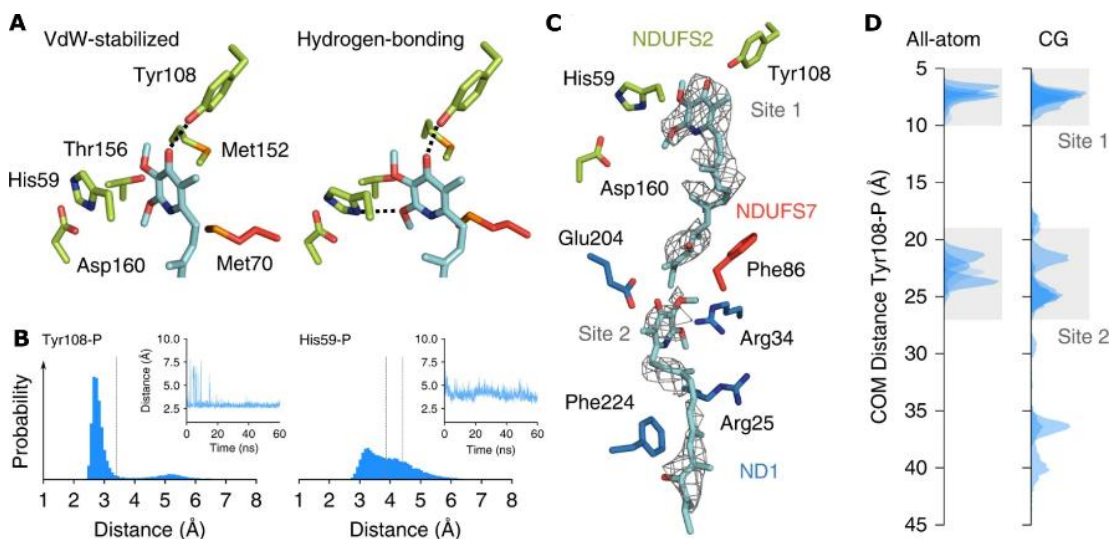
Figure 3.4: (**A**) Snapshots from the MD simulations of van der Waals-stabilized and hydrogen-bonded piericidin conformations. (**B**) Distance distributions and time traces of distances between piericidin (P) and critical residues. (**D**) The two modeled piericidin molecules with the experimental cryo-EM and surrounding residues. (**E**) Distance distributions of piericidin in the substrate cavity obtained from aMD and cgMD simulations. Figure modified from publication C-III.

3.2 Integration of SAXS data (S-I – S-IV)

In publication S-I, I performed all cgMD simulations, was involved in the project design, analysis and interpretation of the experimental and computational results, and manuscript creation. In publications SII and SIII, I developed the used SAXS module in Plumed and the initial hySAXS approach. In publication SIV, I was involved in the expansion of the M&M approach. In all three publications, I was involved in the manuscript creation.

3.2.1 The dynamics of linear polyubiquitin

Ubiquitin (Ub) is a compact and highly conserved (8.6 kDa and 76 amino acids) protein ubiquitously found in eukaryotes. It can be attached as a post-transcriptional modification to regulate various physiological processes, such as in protein degradation, cell cycle, apoptosis, DNA repair, and signal transduction (*79-81*). In addition, ubiquitin itself can serve as a starting point for longer chains by binding the C-terminus to either lysine residues (Lys6/11/27/29/33/48/63) or the N-terminal methionine (Met1) of additional Ub. This allows the formation of polyubiquitin chains with different lengths and topologies, accompanied by a high degree of conformational variability (*83-85*). As such, characterizing the behavior of polyubiquitin chains is challenging from a computational and experimental perspective.

In publication S-I, we focus on the linear (Met1-linked) polyubiquitin that participates, among others (*87-89*), in the activation of the canonical nuclear factor κB (NF-κB) pathway (*87, 90-95*), which in turn regulates cellular processes such as inflammatory responses or cellular growth.

Here, we present a new computational method for integrating small-angle X-ray scattering (SAXS) data with coarse-grained simulation (cgMD). We apply it to accurately and efficiently characterize the dynamics of linear Ub_2, Ub_3, and Ub_4 chains. Furthermore, we used these results along with additional simulations, SAXS, size exclusion chromatography (SEC) coupled with static light scattering (SLS), isothermal titration calorimetry (ITC), and surface plasmon resonance (SPR) experiments to rationalize Ub_x binding with NF-κB essential modulator (NEMO).

Atomistic MD (aMD) simulations are routinely used to characterize the conformational ensemble of proteins under equilibrium conditions, but system size is often a limiting factor for their application. To overcome this limitation, one can use coarse-grained force fields such as MARTINI (see section 2.4). However, this usually comes at the expense of accuracy. Specifically, our cgMD simulation using the MARTINI force field 2.2 produced a too compact structural

ensemble (*162-164*) that cannot reproduce the experimental SAXS data (see **Figure** 3.5). In particular, the simulations underestimate the gyration radius (R_g) by 5 Å as compared to the experimental R_g^{exp} of 22.3 Å. We found that a more realistic ensemble could be modeled by increasing the water-protein Lennard-Jones interaction by 5%, an approach used to improve atomistic and coarse-grained force fields, resulting in a much larger ensemble with an average R_g of 20.5 Å.
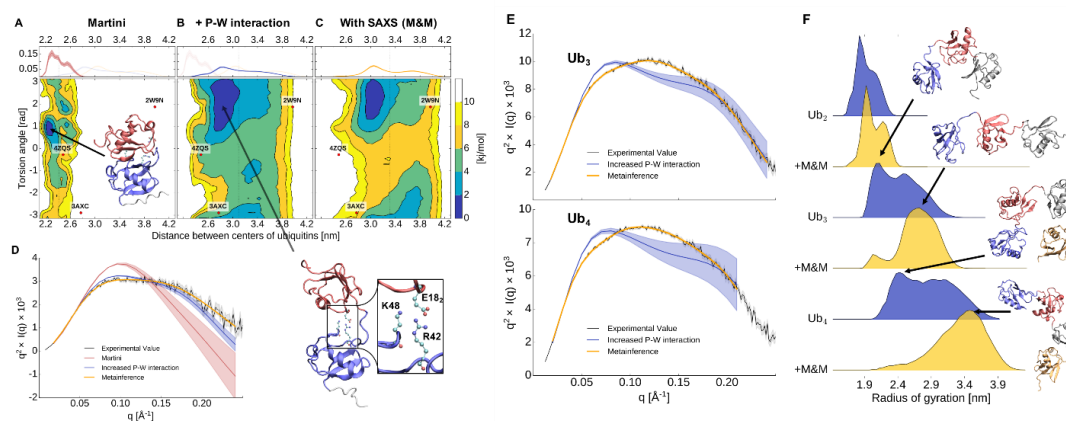

Figure 3.5: Characterization of the dynamics of linear Ub_x. (**A** to **C**) Free energy landscapes as a function of the distance between the center of mass of the two ubiquitin domains and their relative orientation. (**D**) Experimental and from-simulation calculated Kratky plot. (**E**) Experimental and from-simulations calculated Kratky plot for Ub_3 and Ub_4. (**F**) Distribution of the radius of gyration from the cgMD ensembles.

Integration of experimental data can further increase the accuracy of cgMD simulations (see section 2.8). SAXS data appear to be well suited for use with MD simulations (see publication S-II and S-III). Although SAXS does not provide high-resolution structural detail, it gives valuable information on the general shape and dynamics of (multidomain) proteins. In paper S-I, we introduce the use of SAXS data together with Martini cgMD simulations in the metadynamic Metainference (99) (M&M, see section 2.8.3) framework. In linear Ub_2, the use of M&M further increases the agreement with experimental SAXS data, but the global energy landscape does not change considerably (**Figure** 3.5 A-C). We performed additional simulations with the Lys63-linked Ub_2 to compare our results with the atomistic ensemble from publication S-III. Overall, our global energy landscape and the distribution of R_g and contact maps obtained by cgMD simulations with M&M are comparable to the SAXS-restricted aMD ensembles, except for a very compact state, which was also absent in the unrestricted simulations. Simulations on linear Ub_3 and Ub_4 using the enhanced protein-water interaction alone, i.e. without inclusion of exp data, did not reproduce the experimental data, but rather overestimated the amount of contacts between nonadjacent Ub pairs (**Figure** 3.5 E,F). Therefore, integration of SAXS data is

required to characterize the correct behavior of linear polyubiquitin chains at coarse-grained level.

Ub_2, Ub_3 and Ub_4 prefer extended conformations, resulting in an almost linear increase in R_g (20 Å, 27 Å, and 33 Å). Thus, the interaction between nonadjacent ubiquitins contributes very little to the overall ensemble. Interdomain contacts between neighboring Ub pairs appear to be independent, but the free energy landscape describing these interactions between Ub pairs becomes flatter for longer chains. Furthermore, the contacts between domains are not dominated by specific interactions, as the most common contacts are present in only 10% to 30% of all compact states. Nevertheless, these residues belong to three distinct surfaces that define a preferred orientation between the neighboring Ub pair: The surface around the hydrophobic patch Ile44 interacts preferentially with the surfaces (of the second Ub) around Glu18-2 (located on the opposite side of the Ile44) and Ile13-2 (located axially rotated by 90°) (**Figure** 3.6). However, the overall interaction between both ubiquitins is relatively weak and isotropic.
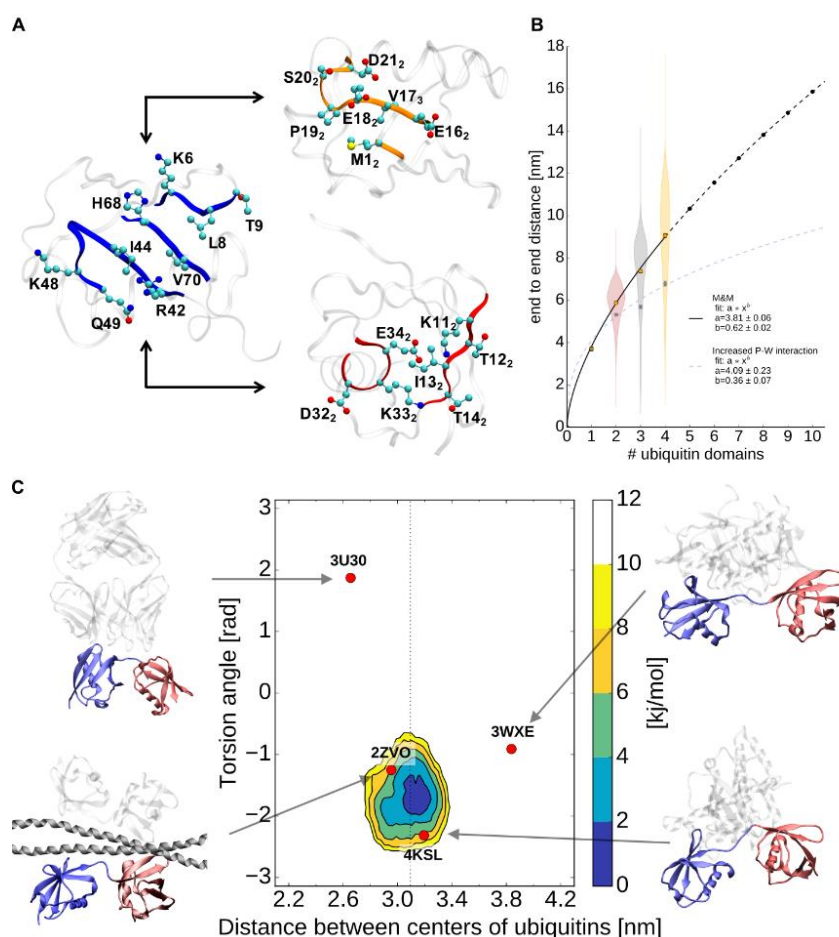
Figure 3.6 (**A**) Interaction surface of two neighboring ubiquitins. (**B**) Average end-to-end distance of a linear polyubiquitin chain. (**C**) Free energy landscapes (in kJ/mol) as a function of the distance between the centers of the two ubiquitin domains and their relative orientation for Ub2 bound to NEMO. Figure modified from publication C-III.

Remarkably, the behavior of linear ubiquitin chains is in strong agreement with the Flory theory for self-avoiding polymers(*194*), a behavior that is rather untypical for folded proteins. We also observe this trend for K63 Ub chains. This points to the fact that polyubiquitin chains avoid ordering and remain flexible until they are bound to a client. Indeed, a comparison of the crystal structures of bound Ub_2 shows that adjacent Ub_2 dimers adopt a completely different conformation depending on the binding partner (**Figure** 3.6 B).

To gain further insight into protein interactions with linear polyubiquitin chains, we studied Ub_x interactions with the dimeric UBAN motif in regulatory NF-κB essential modulator NEMO (95, 97). Here, we specifically investigated the binding of the NEMO dimer with additional simulations, SAXS, size exclusion chromatography (SEC) coupled with static light scattering (SLS), isothermal titration calorimetry (ITC), and surface plasmon resonance (SPR) experiments. Crystal structures suggest that a 2:1 and 2:2 stoichiometry is possible for Ub2. Also, longer Ub chains have more potential binding sites and could, in principle, bind multiple NEMO dimers (or the homolog ABIN such as in PDB ID: 5H07). However, our experiments show only the formation of 2:1 NEMO:Ub_3 and NEMO:Ub_4 complexes characterized by similar binding enthalpy (ΔH of -17.9 kJ/mol and -18.8 kJ/mol vs. -16.9 kJ/mol for NEMO:Ub_2) and affinities (1.6 and 4.1 µM vs. 1.8 µM for NEMO:Ub_2). The conformational entropy of linear polyubiquitins may account for this contradictory stoichiometry. The binding of NEMO strongly reduces the conformational freedom of linear Ub2 (**Figure** 3.6 C) with a disproportionate reduction for Ub_3 and Ub_4 in the case of conformations suitable for binding two NEMO dimers. The difference between 2:1 and 2:2 binding can be rationalized based on our NEMO:Ub_2 ensemble, suggesting that the solvent-accessible surface area (SASA) of the occupied site is 13 nm2. In comparison, the SASA of the unoccupied site is 11.9 nm^2. Both binding sites are not the same after binding of the first Ub_2 to the NEMO dimer.

3.2.2 hySAXS

Small-angle X-ray scattering (SAXS) data can provide valuable insights into the dynamics of biomolecular systems. Therefore, this type of data is well suited to be utilized in conjunction with atomistic MD simulations (*25*). However, the computation of SAXS profiles from atomistic structures is associated with a high computational cost. In publication SII, we present a hybrid multi-resolution strategy to perform atomistic SAXS-constrained simulations by back-calculating scattering intensities over the virtual positions of Martini beads (hySAXS) and used this approach for the structural refinement of an RNA/protein complex. Our approach is publicly available in the PLUMED library.

Since the Debye equation for SAXS back-mapping scales quadratically with the number of required form factors, reducing the number of sites, hereby using Martini beads instead of atoms, leads to a significant reduction in computational cost. To generate SAXS form factors for the Martini beads of nucleic acids, we used the same approach as in Niebling *et al.* (*192*) for protein beads (see section 2.9.2). Our new nucleic acid coarse-grain form factors appear to be reliable for a scattering vector q of up to 0.45 Å. In turn, we used these new martini form factors to refine the ComE-comcde DNA-protein complex model (ComE is part of the ComD-ComE two-component signaling system (*195*)) in the Metainference framework. Metainference enables us to account for various sources of error, such as uncertainties from the forward model (e.g., the possible excess electron density in the hydration shell). The refined structure shows a significantly improved agreement with experimental scattering intensities (x^2 of 1.58 compared to $x^2 = 4.01$ for the initial model (PDB ID: 4CBV (*195*)) and $x^2 = 7.21$ for a structure obtained from an unrestrained simulation). In addition, the use of SAXS constraints results in an excellent *molprobity* score (*196*) with only a slight deviation from the crystal structure (2.5 Å vs. 6.3 Å) and retention of the critical interaction sites between ComE and DNA (**Figure** 3.7).
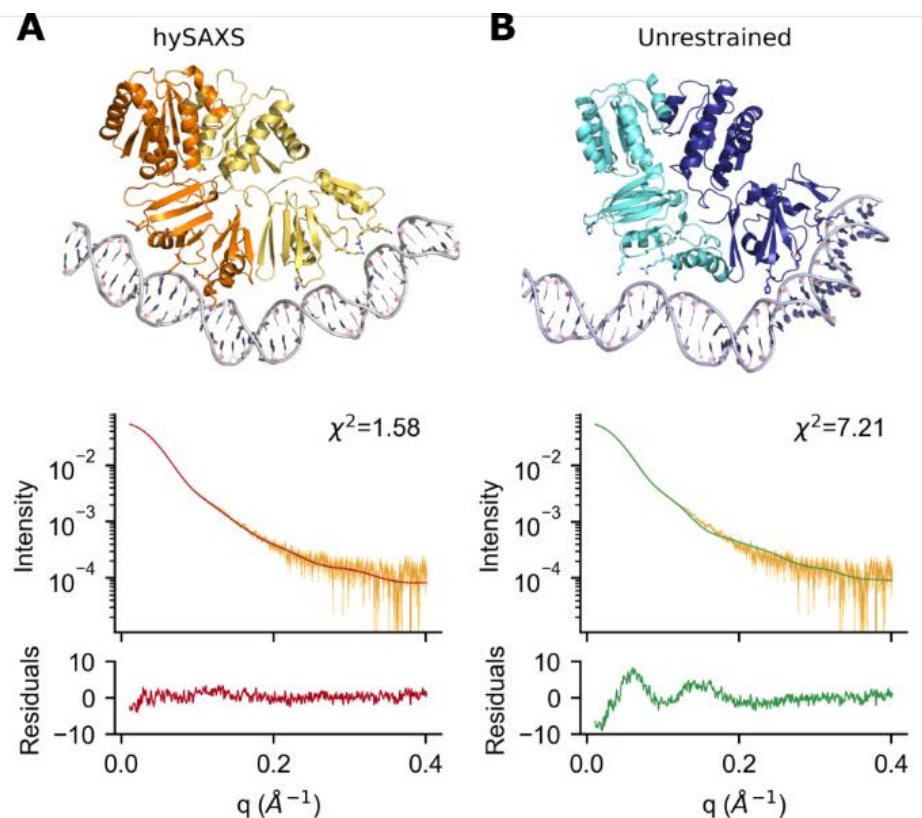
Figure 3.7: The representative ComE–DNA structures extracted from (**A**), the metainference and (**B**) the unrestrained simulations, along with their fits to the experimental data and residuals, according to CRYSOL. Image modified from S-II.

In publication S-III, we extended our hySAXS approach to the multi-replica Metadynamics Metainference (M&M) framework to accurately produce ensembles of dynamical systems with Lys63-linked diubiquitin as a test system. Using a multiple time-step protocol (by applying the Metainference bias only every few time steps), the computational efficiency becomes competitive with unconstrained atomistic simulations (**Figure** 3.8 A, B). Adding SAXS restraints significantly modifies the relative orientation of both ubiquitin cores and the average protein elongation (**Figure** 3.8 C). Specifically, the average gyration radius (R_g) changes from 21.7 ± 0.5 Å in the unrestrained ensemble to 20.6 ± 0.3 Å in the SAXS-restrained once (experimental value: 21.0 ± 0.1 Å). Still, like linear diubiquitin in publication S-I, Lys63-linked diubiquitin can adopt a wide range of open and compact conformations. To test whether the addition of scattering intensities can also enhance the accuracy of local inter-domain interactions, we compared our results with independent paramagnetic relaxation enhancement (PRE) measurements. Indeed, additional SAXS restraints increase the agreement with the N25-PRE data and explicitly allow the reproduction of the N25-PRE intensities for the proximal unit residues 8-14.
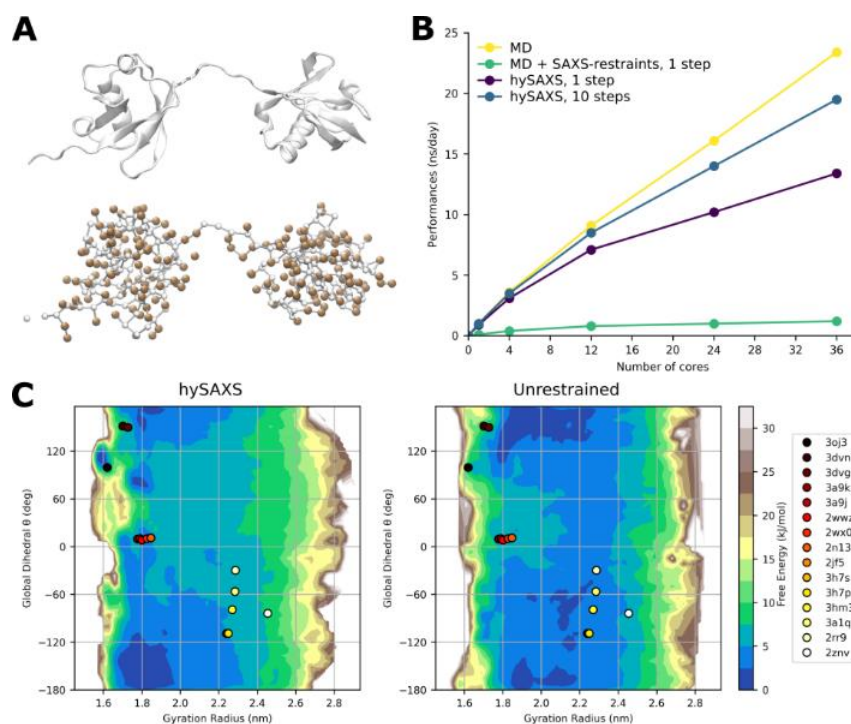
Figure 3.8: (**A**) K63-Ub2 shown as a cartoon representation (top) or highlighting the centers of the 328 Martini beads (bottom) (**B**): Performances, as a function of the number of cores (**C**) Two-dimensional free-energy surface for K63-Ub2, derived by the hySAXS (left panel) and the unrestrained (right panel) ensembles, as a function of the Cα-gyration radius and the global dihedral angle. Image modified from publication S-III.

This finding is consistent with our results from publication S-IV, where we systematically compared ensembles of an intrinsically disordered peptide (EGAAWAASS (*197*)) generated by two diverse force fields and integration of different types of NMR data. Specifically, we compared CHARMM22* (*198*) and AMBER99SB (*155*) integrating chemical shifts (CS), 3J couplings, and residual dipolar couplings (RDCs) using the gyration radius distribution as a metric for comparison. Adding experimental restraints can produce a force-field-independent ensemble (**Figure** 3.9). This point can be exploited by using a computationally less demanding methodology. Here, we were able to use the low-cost implicit solvent model EEF1-SB with CHARMM36 (*199*) and M&M to produce an ensemble comparable to the atomistic simulations with explicit solvent.
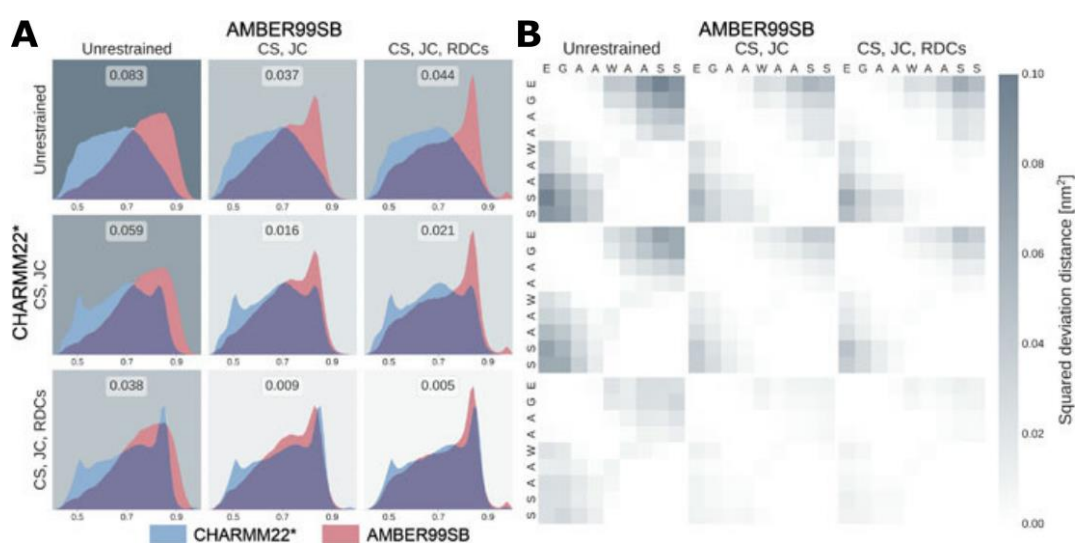
Figure 3.9: (**A**) Ensemble convergence shown as Jensen-Shannon divergences between the probability distributions of the radius of gyration for all pairs of simulations (**B**) Squared-deviation inter-residue distance matrices between each pair of simulations. Image modified from publication S-IV.

3.3 Hsp90 dynamics (H-I)

Own contribution: In publication H-I, I performed and analyzed all aMD and cgMD simulations, was involved in the project design, analysis and interpretation of the experimental and computational results, and manuscript creation.

The molecular chaperone heat shock protein 90 (Hsp90) plays an essential role in many protein folding processes and in preventing protein aggregation in eukaryotic cells (*100-103*). To accomplish this, Hsp90 undergoes large-scale conformational changes driven by ATP splitting (*115-118*). However, this process is poorly understood. Using a combination of atomistic (aMD) and coarse-grained simulations (cgMD), small-angle X-ray scattering (SAXS) experiments, and paramagnetic relaxation enhancement (PRE) measurements, we demonstrate the presence of (partially) extended conformations which could be crucial for client detection and processing.

Hsp90 is a dimer in which the monomer is composed of three domains: the ATP-binding N-terminal domain (NTD), a middle domain (MD), and a C-terminal domain (CTD) that is critical for dimerization. The NTD and MD are connected by a ~60 amino acid long highly charged linker. A full-length (nucleotide-linked) X-ray crystal structure of Hsp90 from *E. coli* (PDB ID: 2CG9 (*122*)) shows Hsp90 in a compact conformation with interdomain NTD-NTD and MD-MD contacts. In this state, Hsp90 has a radius of gyration (R_g) of ~40 Å and a maximum pair-distance D_{max} of ~ 140 Å. However, our SAXS measurement, in agreement with the literature, suggests that Hsp90 has an R_g >60 Å and a D_{max} of ~250 Å. Moreover, even an open "V-shaped" conformation is not sufficient to explain the SAXS data, as it only accounts for an R_g of 50 Å or less (**Figure** 3.10). Therefore, more extended conformations are needed to explain the dynamics of Hsp90. In publication H-I, we propose, based on experimental and computational results, that these extended conformations are caused in part by a separation of the NTD from the middle domain.
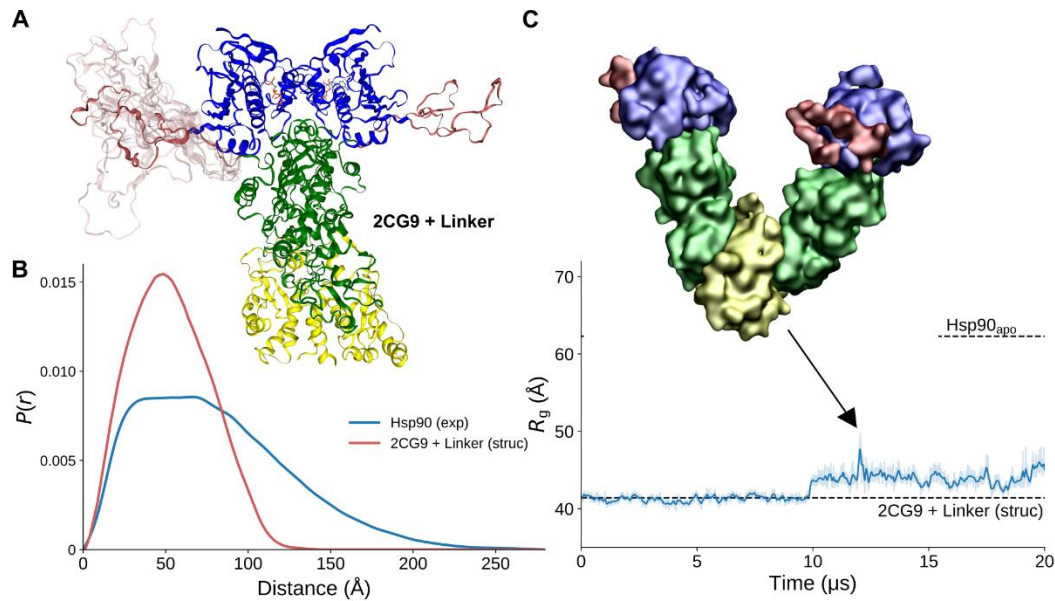
Figure 3.10 **A**) X-ray structure of Hsp90 (PDB ID: 2CG9) + modeled linker (NTD: blue, linker: red, MD: green, CTD: yellow). **B**) Pair-distance distribution functions of Hsp90 from SAXS and crystal structure (2CG9) with a modeled charged linker. **C**) The radius of gyration (R_g) during CGMD simulations of the Hsp90 dimer with bound ATP during the transition from the closed to open states at around 10 µs. Figure modified from publication H-I.

To test our hypothesis, we performed SAXS and PRE experiments on an NTD-MD construct and used aMD simulations to rationalize the results. As in the case of the full-length Hsp90, the NTD-MD structure obtained from the full-length compact crystal structure is insufficient to explain the SAXS data, in particular underestimating the R_g by ~3 Å (R_g^{exp} ~ 35.7 Å). Additionally, our PRE data suggest a mixture between a compact-like state and a semi-open state similar to the crystal structure of GRP94 (PDB ID: 2OU1 (*200*)), a homolog to Hsp90 (**Figure** 3.11). The difference between the two conformations is a ~90° rotation of the NTD relative to the MD. This is consistent with our aMD simulations, where we observe a highly flexible system with multiple compact and transiently open states with a 4:1 ratio between attached and detached conformations. Detached conformations can reach an R_g of up to 40 Å, which balances the R_g of the compact conformations to reproduce the experimental data. We also see short distances for either residues Val114, Ile117, and Val112 (on the NTD-MD interface), which are sensitive to the closed state, or residues Leu50, Leu56, and Leu62 (on the solvent-exposed end of the long helix α2 in NTD), that are sensitive to the semi-open state (**Figure** 3.11). Thus, the combined data show that NTD-MD dissociation is a significant contributor to the NTD-MD ensemble. To investigate the effect of the linker between NTD and MD, we performed additional SAXS measurements and aMD simulations with a deleted charged linker (CL) construct in which residues 211-273 were removed (NTD-MD$_{\Delta CL}$ construct). Our results indicate that the

charged linker region stabilizes the interaction between the NTD and MD by ~0.7kBT, similar to previous pulling experiments.
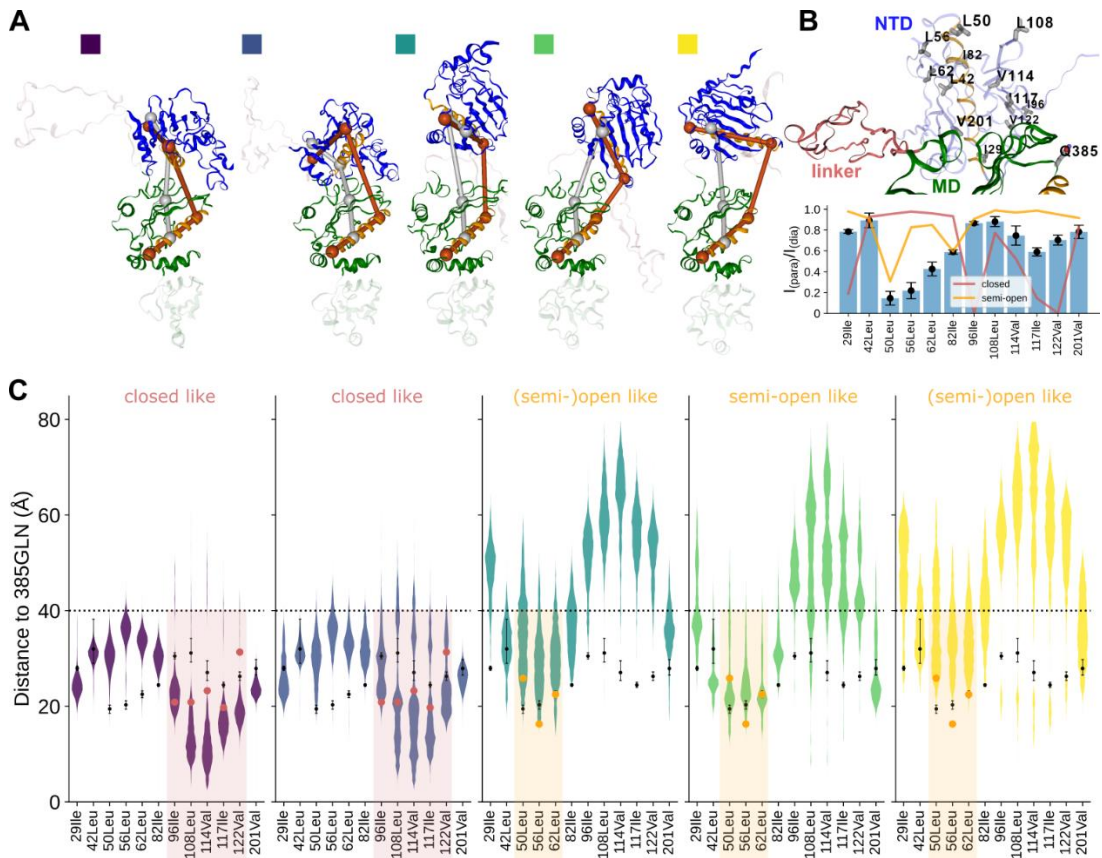

Figure 3.11 **A**) Conformational dynamics of the NTD-MD construct (NM) from the aMD simulations. **B**) Top: Residues with PRE intensity ratio < 0.9 (sticks) shown on the NM structure. Bottom: Comparing experimental PRE data (bars) and values calculated for the closed and semi-open states (lines). **C**) Comparison between the distances derived by PRE (black dots) and the distance distributions from the MD-based clusters (colored violin plots). The residues highlighted are sensitive to conformations observed in the semi-open (orange) and closed (red) state of Hsp90. Figure modified from publication H-I

To test whether these extended conformations are also feasible in full-length Hsp90, we performed additional cgMD simulations using the Martini course-grained force field. We used the same approach to address the problem with the overestimation of the protein-protein interaction as in publication S-I. We observe qualitatively similar NTD-MD dynamics as in the aMD simulations. Interestingly, in the cgMD simulations with ATP bound, we did not observe detached NTD-MD states, suggesting that nucleotide binding may stabilize the NTD-MD interaction. This is consistent with previous SAXS measurements indicating a lower Rg for nucleotide-bound Hsp90 constructs.

4 Conclusions

In this thesis, we employed a vast spectrum of state of the art computational methods to study challenging systems from multiple perspectives - QM methods were used to characterize electron transfer reactions in Complex I, atomistic and coarse-grained molecular dynamics simulations were utilized to study the dynamics of Complex I, and Hsp90; and coarse-grained simulation integrated with experimental SAXS data and enhanced sampling methods were applied to characterize the global conformational ensemble of linear polyubiquitin chains. The results are summarized in publications C-I - C-IV focused on Complex I, publications S-I - S-IV dedicated to linear polyubiquitin, and publication H-I on Hsp90.

In publications CI-CIV, we identified different substrate binding modes of ubiquinone in Complex I validated with cryo-EM and functional studies. We also provided a mechanistic explanation of how cardiolipin could regulate the activity of Complex I by coupling the large-scale twisting and bending motions with the quinone dynamics and the formation putative channels leading to the quinone cavity. These channels could be responsible for the reprotonation of Tyr87 and His39, two residues that protonate quinone under catalysis. By combining QM cluster models, QM/MM, aMD and cgMD simulations, we could successfully link the local behavior of quinone within the binding cavity of Complex I to global conformational changes affecting the hydrophilic and hydrophobic domains of the protein.

In publications SI-SIV, we integrated SAXS data into molecular dynamics simulations using the metainference framework. In doing so, we introduced a hybrid SAXS approach (hySAXS), allowing us to reduce the computational cost of SAXS restraints for aMD simulations. We also presented a novel approach to combining SAXS data with cgMD simulations. To this end, we used the MARTINI force field with a simple modification to tune the protein-solvent interactions and improving the quality of the force field. We employed the same strategy to investigate the large-scale dynamics of Hsp90. These strategies allowed us to accurately characterize the dynamics of linear polyubiquitin chains, which would not be possible with the limited sampling in atomistic simulations. Our results suggest that the individual ubiquitin pairs within the polyubiquitin chains transiently interact and behave independently of each other. Thus, a simple polymer model can be applied to describe the behavior of all polyubiquitin chains.

Our results on Hsp90 strongly imply that the separation of the NTD and MD domains is pivotal for the function of this molecular chaperone. The observed dynamics explain crucial experimental data and suggest the existence of much more open conformations as compared to those observed in crystal structures.

Intriguingly, there appears to be a parallel between Hsp90 and polyubiquitin dynamics: since both proteins can interact with various binding partners, a high degree of structural flexibility in their free state is beneficial. As such, both proteins exhibit behavior resembling intrinsically disordered proteins (IDPs). While IDPs frequently form a stable secondary structure after finding their binding partner, the quaternary structure of Hsp90 or polyubiquitin adapts to their respective client proteins.

In summary, multi-scale computational methods have been used in this thesis as a powerful approach to study the structure, function, and dynamics of several complex biological systems and provide atomistic insights complementary to those obtained experimentally.

5 References

1.   F. Himo, Recent Trends in Quantum Chemical Modeling of Enzymatic Reactions. *J Am Chem Soc* **139**, 6780-6786 (2017).

2.   S. A. Hollingsworth, R. O. Dror, Molecular Dynamics Simulation for All. *Neuron* **99**, 1129-1143 (2018).

3.   P. Saura, M. Ropke, A. P. Gamiz-Hernandez, V. R. I. Kaila, Quantum Chemical and QM/MM Models in Biochemistry. *Methods Mol Biol* **2022**, 75-104 (2019).

4.   S. Kmiecik, D. Gront, M. Kolinski, L. Wieteska, A. E. Dawid, A. Kolinski, Coarse-Grained Protein Models and Their Applications. *Chem Rev* **116**, 7898-7936 (2016).

5.   A. Hospital, J. R. Goni, M. Orozco, J. L. Gelpi, Molecular dynamics simulations: advances and applications. *Adv Appl Bioinform Chem* **8**, 37-47 (2015).

6.   P. C. T. Souza, R. Alessandri, J. Barnoud, S. Thallmair, I. Faustino, F. Grunewald, I. Patmanidis, H. Abdizadeh, B. M. H. Bruininks, T. A. Wassenaar, P. C. Kroon, J. Melcr, V. Nieto, V. Corradi, H. M. Khan, J. Domanski, M. Javanainen, H. Martinez-Seara, N. Reuter, R. B. Best, I. Vattulainen, L. Monticelli, X. Periole, D. P. Tieleman, A. H. de Vries, S. J. Marrink, Martini 3: a general purpose force field for coarse-grained molecular dynamics. *Nat Methods* **18**, 382-388 (2021).

7.   L. Monticelli, S. K. Kandasamy, X. Periole, R. G. Larson, D. P. Tieleman, S. J. Marrink, The MARTINI Coarse-Grained Force Field: Extension to Proteins. *J Chem Theory Comput* **4**, 819-834 (2008).

8.   S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, A. H. de Vries, The MARTINI force field: coarse grained model for biomolecular simulations. *J Phys Chem B* **111**, 7812-7824 (2007).

9.   J. Habchi, P. Tompa, S. Longhi, V. N. Uversky, Introducing protein intrinsic disorder. *Chem Rev* **114**, 6561-6588 (2014).

10.   A. Warshel, M. Levitt, Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *Journal of Molecular Biology* **103**, 227-249 (1976).

11.   M. Bonomi, G. T. Heller, C. Camilloni, M. Vendruscolo, Principles of protein structural ensemble determination. *Curr Opin Struct Biol* **42**, 106-116 (2017).

12.   Y. I. Yang, Q. Shao, J. Zhang, L. Yang, Y. Q. Gao, Enhanced sampling in molecular dynamics. *J Chem Phys* **151**, 070902 (2019).

13.   A. Laio, M. Parrinello, Escaping free-energy minima. *Proc Natl Acad Sci U S A* **99**, 12562-12566 (2002).

14.   A. Laio, F. L. Gervasio, Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Reports on Progress in Physics* **71**, (2008).

15.   G. M. Torrie, J. P. Valleau, Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics* **23**, 187-199 (1977).

16. S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, P. A. Kollman, THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *Journal of Computational Chemistry* **13**, 1011-1021 (1992).

17. L. Bonati, Y. Y. Zhang, M. Parrinello, Neural networks-based variationally enhanced sampling. *Proc Natl Acad Sci U S A* **116**, 17641-17647 (2019).

18. M. S. M, V. S. Pande, tICA-Metadynamics: Accelerating Metadynamics by Using Kinetically Selected Collective Variables. *J Chem Theory Comput* **13**, 2440-2447 (2017).

19. S. Bottaro, T. Bengtsen, K. Lindorff-Larsen, Integrating Molecular Simulation and Experimental Data: A Bayesian/Maximum Entropy Reweighting Approach. *Methods Mol Biol* **2112**, 219-240 (2020).

20. A. Cesari, S. Reißer, G. Bussi, Using the Maximum Entropy Principle to Combine Simulations and Solution Experiments. *Computation* **6**, (2018).

21. M. Bonomi, C. Camilloni, A. Cavalli, M. Vendruscolo, Metainference: A Bayesian inference method for heterogeneous systems. *Sci Adv* **2**, e1501177 (2016).

22. C. Camilloni, A. Cavalli, M. Vendruscolo, Assessment of the use of NMR chemical shifts as replica-averaged structural restraints in molecular dynamics simulations to characterize the dynamics of proteins. *J Phys Chem B* **117**, 1838-1843 (2013).

23. Z. Gong, C. D. Schwieters, C. Tang, Theory and practice of using solvent paramagnetic relaxation enhancement to characterize protein conformational dynamics. *Methods* **148**, 48-56 (2018).

24. M. Bonomi, C. Camilloni, Integrative structural and dynamical biology with PLUMED-ISDB. *Bioinformatics* **33**, 3999-4000 (2017).

25. L. Boldon, F. Laliberte, L. Liu, Review of the fundamental theories behind small angle X-ray scattering, molecular dynamics simulations, and relevant integrated application. *Nano Rev* **6**, 25661 (2015).

26. M. Bonomi, M. Vendruscolo, Determination of protein structural ensembles using cryo-electron microscopy. *Curr Opin Struct Biol* **56**, 37-45 (2019).

27. U. Brandt, Energy converting NADH:quinone oxidoreductase (complex I). *Annu Rev Biochem* **75**, 69-92 (2006).

28. J. Hirst, Mitochondrial complex I. *Annu Rev Biochem* **82**, 551-575 (2013).

29. L. A. Sazanov, A giant molecular proton pump: structure and mechanism of respiratory complex I. *Nat Rev Mol Cell Biol* **16**, 375-388 (2015).

30. V. Zickermann, C. Wirth, H. Nasiri, K. Siegmund, H. Schwalbe, C. Hunte, U. Brandt, Structural biology. Mechanistic insight from the crystal structure of mitochondrial complex I. *Science* **347**, 44-49 (2015).

31. J. Zhu, K. R. Vinothkumar, J. Hirst, Structure of mammalian respiratory complex I. *Nature* **536**, 354-358 (2016).

32. K. Fiedorczuk, J. A. Letts, G. Degliesposti, K. Kaszuba, M. Skehel, L. A. Sazanov, Atomic structure of the entire mammalian mitochondrial complex I. *Nature* **538**, 406-410 (2016).

33. A. A. Agip, J. N. Blaza, H. R. Bridges, C. Viscomi, S. Rawson, S. P. Muench, J. Hirst, Cryo-EM structures of complex I from mouse heart mitochondria in two biochemically defined states. *Nat Struct Mol Biol* **25**, 548-556 (2018).

34. A. J. Jones, J. N. Blaza, F. Varghese, J. Hirst, Respiratory Complex I in Bos taurus and Paracoccus denitrificans Pumps Four Protons across the Membrane for Every NADH Oxidized. *J Biol Chem* **292**, 4987-4995 (2017).

35. V. R. I. Kaila, Long-range proton-coupled electron transfer in biological energy conversion: towards mechanistic understanding of respiratory complex I. *J R Soc Interface* **15**, (2018).

36. P. Mitchell, Chemiosmotic coupling in oxidative and photosynthetic phosphorylation. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* **1807**, 1507-1538 (2011).

37. J. P. Hosler, S. Ferguson-Miller, D. A. Mills, Energy transduction: proton transfer through the respiratory complexes. *Annu Rev Biochem* **75**, 165-187 (2006).

38. Y. Zu, R. J. Shannon, J. Hirst, Reversible, electrochemical interconversion of NADH and NAD+ by the catalytic (Ilambda) subcomplex of mitochondrial NADH:ubiquinone oxidoreductase (complex I). *J Am Chem Soc* **125**, 6020-6021 (2003).

39. M. Alcazar-Fabra, P. Navas, G. Brea-Calvo, Coenzyme Q biosynthesis and its role in the respiratory chain structure. *Biochim Biophys Acta* **1857**, 1073-1078 (2016).

40. L. A. Schurig-Briccio, T. Yano, H. Rubin, R. B. Gennis, Characterization of the type 2 NADH:menaquinone oxidoreductases from Staphylococcus aureus and the bactericidal action of phenothiazines. *Biochim Biophys Acta* **1837**, 954-963 (2014).

41. J. M. Schuller, J. A. Birrell, H. Tanaka, T. Konuma, H. Wulfhorst, N. Cox, S. K. Schuller, J. Thiemann, W. Lubitz, P. Setif, T. Ikegami, B. D. Engel, G. Kurisu, M. M. Nowaczyk, Structural adaptations of photosynthetic complex I enable ferredoxin-dependent electron transfer. *Science* **363**, 257-260 (2019).

42. V. Zickermann, S. Kerscher, K. Zwicker, M. A. Tocilescu, M. Radermacher, U. Brandt, Architecture of complex I and its implications for electron transfer and proton pumping. *Biochim Biophys Acta* **1787**, 574-583 (2009).

43. X. Zhou, W. Fenical, The unique chemistry and biology of the piericidins. *J Antibiot (Tokyo)* **69**, 582-593 (2016).

44. M. Degli Esposti, Inhibitors of NADH–ubiquinone reductase: an overview. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* **1364**, 222-235 (1998).

45. D. J. Horgan, T. P. Singer, J. E. Casida, Studies on the Respiratory Chain-linked Reduced Nicotinamide Adenine Dinucleotide Dehydrogenase. *Journal of Biological Chemistry* **243**, 834-843 (1968).

46. J. Hirst, J. Carroll, I. M. Fearnley, R. J. Shannon, J. E. Walker, The nuclear encoded subunits of complex I from bovine heart mitochondria. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* **1604**, 135-150 (2003).

47. D. M. Elurbe, M. A. Huynen, The origin of the supernumerary subunits and assembly factors of complex I: A treasure trove of pathway evolution. *Biochim Biophys Acta* **1857**, 971-979 (2016).

48. K. R. Vinothkumar, J. Zhu, J. Hirst, Architecture of mammalian respiratory complex I. *Nature* **515**, 80-84 (2014).

49. M. L. Verkhovskaya, N. Belevich, L. Euro, M. Wikstrom, M. I. Verkhovsky, Real-time electron transfer in respiratory complex I. *Proc Natl Acad Sci U S A* **105**, 3763-3767 (2008).

50. M. Sato, J. Torres-Bacete, P. K. Sinha, A. Matsuno-Yagi, T. Yagi, Essential regions in the membrane domain of bacterial complex I (NDH-1): the machinery for proton translocation. *J Bioenerg Biomembr* **46**, 279-287 (2014).

51. C. Mathiesen, C. Hägerhäll, Transmembrane topology of the NuoL, M and N subunits of NADH:quinone oxidoreductase and their homologues among membrane-bound hydrogenases and bona fide antiporters. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* **1556**, 121-132 (2002).

52. R. Baradaran, J. M. Berrisford, G. S. Minhas, L. A. Sazanov, Crystal structure of the entire respiratory complex I. *Nature* **494**, 443-448 (2013).

53. J. Michel, J. DeLeon-Rangel, S. Zhu, K. Van Ree, S. B. Vik, Mutagenesis of the L, M, and N subunits of Complex I from Escherichia coli indicates a common role in function. *PLoS One* **6**, e17420 (2011).

54. M. Ropke, P. Saura, D. Riepl, M. C. Poverlein, V. R. I. Kaila, Functional Water Wires Catalyze Long-Range Proton Pumping in the Mammalian Respiratory Complex I. *J Am Chem Soc* **142**, 21758-21766 (2020).

55. A. Di Luca, M. E. Muhlbauer, P. Saura, V. R. I. Kaila, How inter-subunit contacts in the membrane domain of complex I affect proton transfer energetics. *Biochim Biophys Acta Bioenerg* **1859**, 734-741 (2018).

56. V. Sharma, G. Belevich, A. P. Gamiz-Hernandez, T. Rog, I. Vattulainen, M. L. Verkhovskaya, M. Wikstrom, G. Hummer, V. R. Kaila, Redox-induced activation of the proton pump in the respiratory complex I. *Proc Natl Acad Sci U S A* **112**, 11571-11576 (2015).

57. J. G. Fedor, A. J. Y. Jones, A. Di Luca, V. R. I. Kaila, J. Hirst, Correlating kinetic and structural data on ubiquinone binding and reduction by respiratory complex I. *Proc Natl Acad Sci U S A* **114**, 12737-12742 (2017).

58. A. P. Gamiz-Hernandez, A. Jussupow, M. P. Johansson, V. R. I. Kaila, Terminal Electron-Proton Transfer Dynamics in the Quinone Reduction of Respiratory Complex I. *J Am Chem Soc* **139**, 16282-16288 (2017).

59. P. K. Sinha, N. Castro-Guerrero, G. Patki, M. Sato, J. Torres-Bacete, S. Sinha, H. Miyoshi, A. Matsuno-Yagi, T. Yagi, Conserved amino acid residues of the NuoD segment important for structure and function of Escherichia coli NDH-1 (complex I). *Biochemistry* **54**, 753-764 (2015).

60. M. Verkhovsky, D. A. Bloch, M. Verkhovskaya, Tightly-bound ubiquinone in the Escherichia coli respiratory Complex I. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* **1817**, 1550-1556 (2012).

5 References

61.　T. Ohnishi, S. T. Ohnishi, J. C. Salerno, Five decades of research on mitochondrial NADH-quinone oxidoreductase (complex I). *Biol Chem* **399**, 1249-1264 (2018).

62.　J. Warnau, V. Sharma, A. P. Gamiz-Hernandez, A. Di Luca, O. Haapanen, I. Vattulainen, M. Wikstrom, G. Hummer, V. R. I. Kaila, Redox-coupled quinone dynamics in the respiratory complex I. *Proc Natl Acad Sci U S A* **115**, E8413-E8420 (2018).

63.　H. R. Bridges, J. G. Fedor, J. N. Blaza, A. Di Luca, A. Jussupow, O. D. Jarman, J. J. Wright, A. A. Agip, A. P. Gamiz-Hernandez, M. M. Roessler, V. R. I. Kaila, J. Hirst, Structure of inhibitor-bound mammalian complex I. *Nat Commun* **11**, 5261 (2020).

64.　D. Kampjut, L. A. Sazanov, The coupling mechanism of mammalian respiratory complex I. *Science* **370**, (2020).

65.　M. Babot, A. Birch, P. Labarbuta, A. Galkin, Characterisation of the active/de-active transition of mitochondrial complex I. *Biochim Biophys Acta* **1837**, 1083-1092 (2014).

66.　A. Di Luca, V. R. I. Kaila, Global collective motions in the mammalian and bacterial respiratory complex I. *Biochim Biophys Acta Bioenerg* **1859**, 326-332 (2018).

67.　M. Fry, D. E. Green, Cardiolipin requirement for electron transfer in complex I and III of the mitochondrial respiratory chain. *Journal of Biological Chemistry* **256**, 1874-1880 (1981).

68.　S. Dröse, K. Zwicker, U. Brandt, Full recovery of the NADH:ubiquinone activity of complex I (NADH:ubiquinone oxidoreductase) from Yarrowia lipolytica by the addition of phospholipids. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* **1556**, 65-72 (2002).

69.　M. Vos, A. Geens, C. Bohm, L. Deaulmerie, J. Swerts, M. Rossi, K. Craessaerts, E. P. Leites, P. Seibler, A. Rakovic, T. Lohnau, B. De Strooper, S. M. Fendt, V. A. Morais, C. Klein, P. Verstreken, Cardiolipin promotes electron transport between ubiquinone and complex I to rescue PINK1 deficiency. *J Cell Biol* **216**, 695-708 (2017).

70.　C. Arnarez, J. P. Mazat, J. Elezgaray, S. J. Marrink, X. Periole, Evidence for cardiolipin binding sites on the membrane-exposed surface of the cytochrome bc1. *J Am Chem Soc* **135**, 3112-3120 (2013).

71.　A. R. Mehdipour, G. Hummer, Cardiolipin puts the seal on ATP synthase. *Proc Natl Acad Sci U S A* **113**, 8568-8570 (2016).

72.　M. S. Sharpley, R. J. Shannon, F. Draghi, J. Hirst, Interactions between phospholipids and NADH:ubiquinone oxidoreductase (complex I) from bovine mitochondria. *Biochemistry* **45**, 241-248 (2006).

73.　S. Vijay-Kumar, C. E. Bugg, K. D. Wilkinson, R. D. Vierstra, P. M. Hatfield, W. J. Cook, Comparison of the three-dimensional structures of human, yeast, and oat ubiquitin. *Journal of Biological Chemistry* **262**, 6396-6399 (1987).

74.　P. J. Kraulis, Similarity of protein G and ubiquitin. *Science* **254**, 581-582 (1991).

75. L. Hicke, H. L. Schubert, C. P. Hill, Ubiquitin-binding domains. *Nat Rev Mol Cell Biol* **6**, 610-621 (2005).

76. S. Fieulaine, M. D. Witte, C. S. Theile, M. Ayach, H. L. Ploegh, I. Jupin, S. Bressanelli, Turnip yellow mosaic virus protease binds ubiquitin suboptimally to fine-tune its deubiquitinase activity. *J Biol Chem* **295**, 13769-13783 (2020).

77. Y. Sato, H. Fujita, A. Yoshikawa, M. Yamashita, A. Yamagata, S. E. Kaiser, K. Iwai, S. Fukai, Specific recognition of linear ubiquitin chains by the Npl4 zinc finger (NZF) domain of the HOIL-1L subunit of the linear ubiquitin chain assembly complex. *Proc Natl Acad Sci U S A* **108**, 20520-20525 (2011).

78. L. Penengo, M. Mapelli, A. G. Murachelli, S. Confalonieri, L. Magri, A. Musacchio, P. P. Di Fiore, S. Polo, T. R. Schneider, Crystal structure of the ubiquitin binding domains of rabex-5 reveals two modes of interaction with ubiquitin. *Cell* **124**, 1183-1195 (2006).

79. K. Haglund, I. Dikic, Ubiquitylation and cell signaling. *EMBO J* **24**, 3353-3359 (2005).

80. R. J. Deshaies, C. A. Joazeiro, RING domain E3 ubiquitin ligases. *Annu Rev Biochem* **78**, 399-434 (2009).

81. W. Li, Y. Ye, Polyubiquitin chains: functions, structures, and mechanisms. *Cell Mol Life Sci* **65**, 2397-2406 (2008).

82. G. S. McDowell, A. Philpott, Non-canonical ubiquitylation: mechanisms and consequences. *Int J Biochem Cell Biol* **45**, 1833-1842 (2013).

83. S. Zhao, H. D. Ulrich, Distinct consequences of posttranslational modification by linear versus K63-linked polyubiquitin chains. *Proc Natl Acad Sci U S A* **107**, 7704-7709 (2010).

84. D. Komander, M. Rape, The ubiquitin code. *Annu Rev Biochem* **81**, 203-229 (2012).

85. D. Fushman, K. D. Wilkinson, Structure and recognition of polyubiquitin chains of different lengths and linkage. *F1000 Biol Rep* **3**, 26 (2011).

86. M. Akutsu, I. Dikic, A. Bremm, Ubiquitin chain diversity at a glance. *J Cell Sci* **129**, 875-880 (2016).

87. B. Gerlach, S. M. Cordier, A. C. Schmukle, C. H. Emmerich, E. Rieser, T. L. Haas, A. I. Webb, J. A. Rickard, H. Anderton, W. W. Wong, U. Nachbur, L. Gangoda, U. Warnken, A. W. Purcell, J. Silke, H. Walczak, Linear ubiquitination prevents inflammation and regulates immune signalling. *Nature* **471**, 591-596 (2011).

88. I. E. Wertz, K. Newton, D. Seshasayee, S. Kusam, C. Lam, J. Zhang, N. Popovych, E. Helgason, A. Schoeffler, S. Jeet, N. Ramamoorthi, L. Kategaya, R. J. Newman, K. Horikawa, D. Dugger, W. Sandoval, S. Mukund, A. Zindal, F. Martin, C. Quan, J. Tom, W. J. Fairbrother, M. Townsend, S. Warming, J. DeVoss, J. Liu, E. Dueber, P. Caplazi, W. P. Lee, C. C. Goodnow, M. Balazs, K. Yu, G. Kolumam, V. M. Dixit, Phosphorylation and linear ubiquitin direct A20 inhibition of inflammation. *Nature* **528**, 370-375 (2015).

89.   D. Senft, J. Qi, Z. A. Ronai, Ubiquitin ligases in oncogenic transformation and cancer therapy. *Nat Rev Cancer* **18**, 69-88 (2018).

90.   D. A. Catici, J. E. Horne, G. E. Cooper, C. R. Pudney, Polyubiquitin Drives the Molecular Interactions of the NF-kappaB Essential Modulator (NEMO) by Allosteric Regulation. *J Biol Chem* **290**, 14130-14139 (2015).

91.   A. V. Hauenstein, G. Xu, V. Kabaleeswaran, H. Wu, Evidence for M1-Linked Polyubiquitin-Mediated Conformational Change in NEMO. *Journal of Molecular Biology* **429**, 3793-3800 (2017).

92.   M. Hinz, C. Scheidereit, The IκB kinase complex in NF-κB regulation and beyond. *EMBO reports* **15**, 46-61 (2014).

93.   K. Iwai, H. Fujita, Y. Sasaki, Linear ubiquitin chains: NF-kappaB signalling, cell death and beyond. *Nat Rev Mol Cell Biol* **15**, 503-508 (2014).

94.   K. Iwai, F. Tokunaga, Linear polyubiquitination: a new regulator of NF-kappaB activation. *EMBO Rep* **10**, 706-713 (2009).

95.   M. Vincendeau, K. Hadian, A. C. Messias, J. K. Brenke, J. Halander, R. Griesbach, U. Greczmiel, A. Bertossi, R. Stehle, D. Nagel, K. Demski, H. Velvarska, D. Niessing, A. Geerlof, M. Sattler, D. Krappmann, Inhibition of Canonical NF-kappaB Signaling by a Small Molecule Targeting NEMO-Ubiquitin Interaction. *Sci Rep* **6**, 18934 (2016).

96.   T. Kirisako, K. Kamei, S. Murata, M. Kato, H. Fukumoto, M. Kanie, S. Sano, F. Tokunaga, K. Tanaka, K. Iwai, A ubiquitin ligase complex assembles linear polyubiquitin chains. *EMBO J* **25**, 4877-4887 (2006).

97.   S. Rahighi, F. Ikeda, M. Kawasaki, M. Akutsu, N. Suzuki, R. Kato, T. Kensche, T. Uejima, S. Bloor, D. Komander, F. Randow, S. Wakatsuki, I. Dikic, Specific recognition of linear ubiquitin chains by NEMO is important for NF-kappaB activation. *Cell* **136**, 1098-1109 (2009).

98.   T. T. Thach, D. Shin, S. Han, S. Lee, New conformations of linear polyubiquitin chains from crystallographic and solution-scattering studies expand the conformational space of polyubiquitin. *Acta Crystallogr D Struct Biol* **72**, 524-535 (2016).

99.   M. Bonomi, C. Camilloni, M. Vendruscolo, Metadynamic metainference: Enhanced sampling of the metainference ensemble using metadynamics. *Sci Rep* **6**, 31232 (2016).

100.   D. F. Nathan, M. H. Vos, S. Lindquist, In vivo functions of the Saccharomyces cerevisiae Hsp90 chaperone. *Proc Natl Acad Sci U S A* **94**, 12949-12956 (1997).

101.   D. Picard, Heat-shock protein 90, a chaperone for folding and regulation. *Cell Mol Life Sci* **59**, 1640-1648 (2002).

102.   J. C. Young, V. R. Agashe, K. Siegers, F. U. Hartl, Pathways of chaperone-mediated protein folding in the cytosol. *Nat Rev Mol Cell Biol* **5**, 781-791 (2004).

103.   M. Taipale, D. F. Jarosz, S. Lindquist, HSP90 at the hub of protein homeostasis: emerging mechanistic insights. *Nat Rev Mol Cell Biol* **11**, 515-528 (2010).

104. P. Csermely, T. Schnaider, C. So″ti, Z. Prohászka, G. Nardai, The 90-kDa Molecular Chaperone Family. *Pharmacology & Therapeutics* **79**, 129-168 (1998).

105. S. Wickner, J. L. Camberg, S. M. Doyle, D. M. Johnston, "Molecular Chaperones☆" in *Reference Module in Life Sciences* (Elsevier, 2017).

106. J. S. Isaacs, W. Xu, L. Neckers, Heat shock protein 90 as a molecular target for cancer therapeutics. *Cancer Cell* **3**, 213-217 (2003).

107. M. Taipale, I. Krykbaeva, M. Koeva, C. Kayatekin, K. D. Westover, G. I. Karras, S. Lindquist, Quantitative analysis of HSP90-client interactions reveals principles of substrate recognition. *Cell* **150**, 987-1001 (2012).

108. N. Wayne, D. N. Bolon, Dimerization of Hsp90 is required for in vivo function. Design and analysis of monomers and dimers. *J Biol Chem* **282**, 35386-35395 (2007).

109. S. Tsutsumi, M. Mollapour, C. Graf, C. T. Lee, B. T. Scroggins, W. Xu, L. Haslerova, M. Hessling, A. A. Konstantinova, J. B. Trepel, B. Panaretou, J. Buchner, M. P. Mayer, C. Prodromou, L. Neckers, Hsp90 charged-linker truncation reverses the functional consequences of weakened hydrophobic contacts in the N domain. *Nat Struct Mol Biol* **16**, 1141-1147 (2009).

110. S. Tsutsumi, M. Mollapour, C. Prodromou, C. T. Lee, B. Panaretou, S. Yoshida, M. P. Mayer, L. M. Neckers, Charged linker sequence modulates eukaryotic heat shock protein 90 (Hsp90) chaperone activity. *Proc Natl Acad Sci U S A* **109**, 2937-2942 (2012).

111. O. R. Lorenz, L. Freiburger, D. A. Rutz, M. Krause, B. K. Zierer, S. Alvira, J. Cuellar, J. M. Valpuesta, T. Madl, M. Sattler, J. Buchner, Modulation of the Hsp90 chaperone cycle by a stringent client protein. *Mol Cell* **53**, 941-953 (2014).

112. A. Lopez, A. R. Elimelech, K. Klimm, M. Sattler, The Charged Linker Modulates the Conformations and Molecular Interactions of Hsp90. *Chembiochem*, (2020).

113. S. L. Mader, A. Lopez, J. Lawatscheck, Q. Luo, D. A. Rutz, A. P. Gamiz-Hernandez, M. Sattler, J. Buchner, V. R. I. Kaila, Conformational dynamics modulate the catalytic activity of the molecular chaperone Hsp90. *Nat Commun* **11**, 1410 (2020).

114. F. H. Schopf, M. M. Biebl, J. Buchner, The HSP90 chaperone machinery. *Nat Rev Mol Cell Biol* **18**, 345-360 (2017).

115. H. Wegele, P. Muschler, M. Bunck, J. Reinstein, J. Buchner, Dissection of the contribution of individual domains to the ATPase mechanism of Hsp90. *J Biol Chem* **278**, 39303-39310 (2003).

116. M. Hessling, K. Richter, J. Buchner, Dissection of the ATP-induced conformational cycle of the molecular chaperone Hsp90. *Nat Struct Mol Biol* **16**, 287-293 (2009).

117. M. Mickler, M. Hessling, C. Ratzke, J. Buchner, T. Hugel, The large conformational changes of Hsp90 are only weakly coupled to ATP hydrolysis. *Nat Struct Mol Biol* **16**, 281-286 (2009).

118. B. Hellenkamp, P. Wortmann, F. Kandzia, M. Zacharias, T. Hugel, Multidomain structure and correlated dynamics determined by self-consistent FRET networks. *Nat Methods* **14**, 174-180 (2017).

119. K. A. Krukenberg, T. O. Street, L. A. Lavery, D. A. Agard, Conformational dynamics of the molecular chaperone Hsp90. *Q Rev Biophys* **44**, 229-255 (2011).

120. Y. E. Kim, M. S. Hipp, A. Bracher, M. Hayer-Hartl, F. U. Hartl, Molecular chaperone functions in protein folding and proteostasis. *Annu Rev Biochem* **82**, 323-355 (2013).

121. J. Oroz, J. H. Kim, B. J. Chang, M. Zweckstetter, Mechanistic basis for the recognition of a misfolded protein by the molecular chaperone Hsp90. *Nat Struct Mol Biol* **24**, 407-413 (2017).

122. M. M. Ali, S. M. Roe, C. K. Vaughan, P. Meyer, B. Panaretou, P. W. Piper, C. Prodromou, L. H. Pearl, Crystal structure of an Hsp90-nucleotide-p23/Sba1 closed chaperone complex. *Nature* **440**, 1013-1017 (2006).

123. L. A. Lavery, J. R. Partridge, T. A. Ramelot, D. Elnatan, M. A. Kennedy, D. A. Agard, Structural asymmetry in the closed state of mitochondrial Hsp90 (TRAP1) supports a two-step ATP hydrolysis mechanism. *Mol Cell* **53**, 330-343 (2014).

124. J. D. Huck, N. L. Que, F. Hong, Z. Li, D. T. Gewirth, Structural and Functional Analysis of GRP94 in the Closed State Reveals an Essential Role for the Pre-N Domain and a Potential Client-Binding Site. *Cell Rep* **20**, 2800-2809 (2017).

125. J. Townsend, J. K. Kirkland, K. D. Vogiatzis, "Chapter 3 - Post-Hartree-Fock methods: configuration interaction, many-body perturbation theory, coupled-cluster theory" in *Mathematical Physics in Theoretical Chemistry,* S. M. Blinder, J. E. House, Eds. (Elsevier, 2019), pp. 63-117.

126. P. Hohenberg, W. Kohn, Inhomogeneous Electron Gas. *Physical Review* **136**, B864-B871 (1964).

127. W. Kohn, L. J. Sham, Self-Consistent Equations Including Exchange and Correlation Effects. *Physical Review* **140**, A1133-A1138 (1965).

128. J. P. Perdew, K. Burke, M. Ernzerhof, Generalized Gradient Approximation Made Simple. *Phys Rev Lett* **77**, 3865-3868 (1996).

129. J. Tao, J. P. Perdew, V. N. Staroverov, G. E. Scuseria, Climbing the density functional ladder: nonempirical meta-generalized gradient approximation designed for molecules and solids. *Phys Rev Lett* **91**, 146401 (2003).

130. C. Lee, W. Yang, R. G. Parr, Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys Rev B Condens Matter* **37**, 785-789 (1988).

131. A. D. Becke, Density-functional thermochemistry. III. The role of exact exchange. *The Journal of Chemical Physics* **98**, 5648-5652 (1993).

132. S. H. Vosko, L. Wilk, M. Nusair, Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis. *Canadian Journal of Physics* **58**, 1200-1211 (1980).

133. A. D. Becke, Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys Rev A Gen Phys* **38**, 3098-3100 (1988).

134.   Electronic wave functions - I. A general method of calculation for the stationary states of any molecular system. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* **200**, 542-554 (1950).

135.   F. Weigend, R. Ahlrichs, Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys Chem Chem Phys* **7**, 3297-3305 (2005).

136.   S. Grimme, J. Antony, S. Ehrlich, H. Krieg, A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J Chem Phys* **132**, 154104 (2010).

137.   A. Klamt, G. Schüürmann, COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *J. Chem. Soc., Perkin Trans. 2*, 799-805 (1993).

138.   L. Noodleman, C. Y. Peng, D. A. Case, J. M. Mouesca, Orbital interactions, electron delocalization and spin coupling in iron-sulfur clusters. *Coordination Chemistry Reviews* **144**, 199-244 (1995).

139.   J. C. Phillips, D. J. Hardy, J. D. C. Maia, J. E. Stone, J. V. Ribeiro, R. C. Bernardi, R. Buch, G. Fiorin, J. Henin, W. Jiang, R. McGreevy, M. C. R. Melo, B. K. Radak, R. D. Skeel, A. Singharoy, Y. Wang, B. Roux, A. Aksimentiev, Z. Luthey-Schulten, L. V. Kale, K. Schulten, C. Chipot, E. Tajkhorshid, Scalable molecular dynamics on CPU and GPU architectures with NAMD. *J Chem Phys* **153**, 044130 (2020).

140.   M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, E. Lindahl, GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1-2**, 19-25 (2015).

141.   M. A. Cuendet, W. F. van Gunsteren, On the calculation of velocity-dependent properties in molecular dynamics simulations using the leapfrog integration algorithm. *J Chem Phys* **127**, 184102 (2007).

142.   K. Olesen, N. Awasthi, D. S. Bruhn, W. Pezeshkian, H. Khandelia, Faster Simulations with a 5 fs Time Step for Lipids in the CHARMM Force Field. *J Chem Theory Comput* **14**, 3342-3350 (2018).

143.   N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, E. Teller, Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics* **21**, 1087-1092 (1953).

144.   G. Bussi, D. Donadio, M. Parrinello, Canonical sampling through velocity rescaling. *J Chem Phys* **126**, 014101 (2007).

145.   M. Parrinello, A. Rahman, Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics* **52**, 7182-7190 (1981).

146.   S. Nosé, M. L. Klein, Constant pressure molecular dynamics for molecular systems. *Molecular Physics* **50**, 1055-1076 (2006).

147.   M. G. Evans, M. Polanyi, Some applications of the transition state method to the calculation of reaction velocities, especially in solution. *Transactions of the Faraday Society* **31**, 875-894 (1935).

148. P. V. Coveney, S. Wan, On the calculation of equilibrium thermodynamic properties from molecular dynamics. *Phys Chem Chem Phys* **18**, 30236-30240 (2016).

149. A. Grossfield, P. N. Patrone, D. R. Roe, A. J. Schultz, D. W. Siderius, D. M. Zuckerman, Best Practices for Quantification of Uncertainty and Sampling Quality in Molecular Simulations [Article v1.0]. *Living J Comput Mol Sci* **1**, (2018).

150. A. Grossfield, D. M. Zuckerman, "Chapter 2 Quantifying Uncertainty and Sampling Quality in Biomolecular Simulations" in *Annual Reports in Computational Chemistry,* R. A. Wheeler, Ed. (Elsevier, 2009), vol. 5, pp. 23-48.

151. J. Huang, S. Rauscher, G. Nawrocki, T. Ran, M. Feig, B. L. de Groot, H. Grubmuller, A. D. MacKerell, Jr., CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat Methods* **14**, 71-73 (2017).

152. J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, C. Simmerling, ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J Chem Theory Comput* **11**, 3696-3713 (2015).

153. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein, Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* **79**, 926-935 (1983).

154. S. Piana, P. Robustelli, D. Tan, S. Chen, D. E. Shaw, Development of a Force Field for the Simulation of Single-Chain Proteins and Protein-Protein Complexes. *J Chem Theory Comput* **16**, 2494-2507 (2020).

155. S. Piana, A. G. Donchev, P. Robustelli, D. E. Shaw, Water dispersion interactions strongly influence simulated structural properties of disordered protein states. *J Phys Chem B* **119**, 5113-5123 (2015).

156. T. Darden, D. York, L. Pedersen, Particle mesh Ewald: AnN·log(N) method for Ewald sums in large systems. *The Journal of Chemical Physics* **98**, 10089-10092 (1993).

157. D. H. de Jong, G. Singh, W. F. Bennett, C. Arnarez, T. A. Wassenaar, L. V. Schafer, X. Periole, D. P. Tieleman, S. J. Marrink, Improved Parameters for the Martini Coarse-Grained Protein Force Field. *J Chem Theory Comput* **9**, 687-697 (2013).

158. J. J. Uusitalo, H. I. Ingolfsson, S. J. Marrink, I. Faustino, Martini Coarse-Grained Force Field: Extension to RNA. *Biophys J* **113**, 246-256 (2017).

159. D. H. de Jong, S. Baoukina, H. I. Ingólfsson, S. J. Marrink, Martini straight: Boosting performance using a shorter cutoff and GPUs. *Computer Physics Communications* **199**, 1-7 (2016).

160. M. Bulacu, N. Goga, W. Zhao, G. Rossi, L. Monticelli, X. Periole, D. P. Tieleman, S. J. Marrink, Improved Angle Potentials for Coarse-Grained Molecular Dynamics Simulations. *J Chem Theory Comput* **9**, 3282-3292 (2013).

161. M. M. Tirion, Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. *Phys Rev Lett* **77**, 1905-1908 (1996).

162. A. Jussupow, A. C. Messias, R. Stehle, A. Geerlof, S. M. O. Solbak, C. Paissoni, A. Bach, M. Sattler, C. Camilloni, The dynamics of linear polyubiquitin. *Sci Adv* **6**, (2020).

163. A. C. Stark, C. T. Andrews, A. H. Elcock, Toward optimized potential functions for protein-protein interactions in aqueous solutions: osmotic second virial coefficient calculations using the MARTINI coarse-grained force field. *J Chem Theory Comput* **9**, (2013).

164. M. Javanainen, H. Martinez-Seara, I. Vattulainen, Excessive aggregation of membrane proteins in the Martini model. *PLoS One* **12**, e0187936 (2017).

165. H. Lin, D. G. Truhlar, QM/MM: what have we learned, where are we, and where do we go from here? *Theoretical Chemistry Accounts* **117**, 185-199 (2006).

166. S. Riahi, C. N. Rowley, The CHARMM-TURBOMOLE interface for efficient and accurate QM/MM molecular dynamics, free energies, and excited state properties. *J Comput Chem* **35**, 2076-2086 (2014).

167. M. J. Field, P. A. Bash, M. Karplus, A combined quantum mechanical and molecular mechanical potential for molecular dynamics simulations. *Journal of Computational Chemistry* **11**, 700-733 (1990).

168. U. C. Singh, P. A. Kollman, A combinedab initio quantum mechanical and molecular mechanical method for carrying out simulations on complex molecular systems: Applications to the CH3Cl + Cl? exchange reaction and gas phase protonation of polyethers. *Journal of Computational Chemistry* **7**, 718-730 (1986).

169. P. H. Konig, M. Hoffmann, T. Frauenheim, Q. Cui, A critical evaluation of different QM/MM frontier treatments with SCC-DFTB as the QM method. *J Phys Chem B* **109**, 9082-9095 (2005).

170. C. C. David, D. J. Jacobs, Principal component analysis: a method for determining the essential dynamics of proteins. *Methods Mol Biol* **1084**, 193-226 (2014).

171. K. Pearson, LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **2**, 559-572 (1901).

172. G. Bussi, A. Laio, Using metadynamics to explore complex free-energy landscapes. *Nature Reviews Physics* **2**, 200-212 (2020).

173. D. Branduardi, G. Bussi, M. Parrinello, Metadynamics with Adaptive Gaussians. *J Chem Theory Comput* **8**, 2247-2254 (2012).

174. A. Barducci, G. Bussi, M. Parrinello, Well-tempered metadynamics: a smoothly converging and tunable free-energy method. *Phys Rev Lett* **100**, 020603 (2008).

175. S. Piana, A. Laio, A bias-exchange approach to protein folding. *J Phys Chem B* **111**, 4553-4559 (2007).

176. J. Pfaendtner, M. Bonomi, Efficient Sampling of High-Dimensional Free-Energy Landscapes with Parallel Bias Metadynamics. *J Chem Theory Comput* **11**, 5062-5067 (2015).

177. P. Raiteri, A. Laio, F. L. Gervasio, C. Micheletti, M. Parrinello, Efficient reconstruction of complex free energy landscapes by multiple walkers metadynamics. *J Phys Chem B* **110**, 3533-3539 (2006).

178. T. Lohr, A. Jussupow, C. Camilloni, Metadynamic metainference: Convergence towards force field independent structural ensembles of a disordered peptide. *J Chem Phys* **146**, 165102 (2017).

179. B. E. Warren, *X-ray Diffraction* (Courier Corporation, 1990).

180. P. Debye, Zerstreuung von röntgenstrahlen. *Annalen der Physik* **351**, 809-823 (1915).

181. E. Prince, A. J. C. Wilson, *International tables for crystallography* (Kluwer., 2004), vol. 100.

182. D. T. Cromer, J. T. Waber, Scattering factors computed from relativistic Dirac–Slater wave functions. *Acta Crystallographica* **18**, 104-109 (1965).

183. R. Fraser, T. MacRae, E. Suzuki, An improved method for calculating the contribution of solvent to the X-ray diffraction pattern of biological molecules. *Journal of Applied Crystallography* **11**, 693-694 (1978).

184. D. Svergun, C. Barberato, M. H. Koch, CRYSOL–a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. *Journal of applied crystallography* **28**, 768-773 (1995).

185. P.-c. Chen, J. S. Hub, Interpretation of solution x-ray scattering by explicit-solvent molecular dynamics. *Biophysical journal* **108**, 2573-2584 (2015).

186. O. Glatter, O. Kratky, H. Kratky, *Small angle X-ray scattering* (Academic press, 1982).

187. A. Guinier, G. Fournet, K. L. Yudowitch, Small-angle scattering of X-rays. (1955).

188. M. H. Koch, P. Vachette, D. I. Svergun, Small-angle scattering: a view on the properties, structures and structural changes of biological macromolecules in solution. *Q Rev Biophys* **36**, 147-227 (2003).

189. K. Berlin, N. A. Gumerov, D. Fushman, R. Duraiswami, Hierarchical O(N) computation of small-angle scattering profiles and their associated derivatives. *J Appl Crystallogr* **47**, 755-761 (2014).

190. M. Marchi, A first principle particle mesh method for solution SAXS of large bio-molecular systems. *J Chem Phys* **145**, 045101 (2016).

191. S. Yang, S. Park, L. Makowski, 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. *Biophys J* **96**, 4449-4463 (2009).

192. S. Niebling, A. Bjorling, S. Westenhoff, MARTINI bead form factors for the analysis of time-resolved X-ray scattering of proteins. *J Appl Crystallogr* **47**, 1190-1198 (2014).

193. K. M. Ravikumar, W. Huang, S. Yang, Fast-SAXS-pro: a unified approach to computing SAXS profiles of DNA, RNA, protein, and their complexes. *J Chem Phys* **138**, 024112 (2013).

194. S. M. Bhattacharjee, A. Giacometti, A. Maritan, Flory theory for polymers. *Journal of Physics: Condensed Matter* **25**, 503101 (2013).

195. M. Boudes, D. Sanchez, M. Graille, H. van Tilbeurgh, D. Durand, S. Quevillon-Cheruel, Structural insights into the dimerization of the response regulator ComE from Streptococcus pneumoniae. *Nucleic Acids Res* **42**, 5302-5313 (2014).

196. V. B. Chen, W. B. Arendall, 3rd, J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson, D. C. Richardson, MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr* **66**, 12-21 (2010).

197. S. A. Dames, R. Aregger, N. Vajpai, P. Bernado, M. Blackledge, S. Grzesiek, Residual dipolar couplings in short peptides reveal systematic conformational preferences of individual amino acids. *J Am Chem Soc* **128**, 13508-13514 (2006).

198. S. Piana, K. Lindorff-Larsen, D. E. Shaw, How robust are protein folding simulations with respect to force field parameterization? *Biophys J* **100**, L47-49 (2011).

199. S. Bottaro, K. Lindorff-Larsen, R. B. Best, Variational Optimization of an All-Atom Implicit Solvent Force Field to Match Explicit Solvent Simulation Data. *J Chem Theory Comput* **9**, 5641-5652 (2013).

200. M. S. Kumar, M. Carson, M. M. Hussain, H. M. Murthy, Structures of apolipoprotein A-II and a lipid-surrogate complex provide insights into apolipoprotein-lipid interactions. *Biochemistry* **41**, 11681-11691 (2002).

6 Publications

Order of the papers: C-I, C-II, C-III, S-I, S-II, S-III, S-IV

How cardiolipin modulates the dynamics of respiratory complex I

Alexander Jussupow*, Andrea Di Luca*, Ville R. I. Kaila[†]

Cardiolipin modulates the activity of membrane-bound respiratory enzymes that catalyze biological energy transduction. The respiratory complex I functions as the primary redox-driven proton pump in mitochondrial and bacterial respiratory chains, and its activity is strongly enhanced by cardiolipin. However, despite recent advances in the structural biology of complex I, cardiolipin-specific interaction mechanisms currently remain unknown. On the basis of millisecond molecular simulations, we suggest that cardiolipin binds to proton-pumping subunits of complex I and induces global conformational changes that modulate the accessibility of the quinone substrate to the enzyme. Our findings provide key information on the coupling between complex I dynamics and activity and suggest how biological membranes modulate the structure and activity of proteins.

INTRODUCTION

Complex I (NADH:ubiquinone oxidoreductase) functions as a redox-driven proton pump in aerobic respiratory chains (*1–3*). With a molecular mass of ca. 1 MDa distributed among up to 45 subunits in eukaryotes (*4–7*), complex I is the largest, most intricate, and least understood enzyme of the respiratory chain. The over 100-Å-long hydrophilic domain of complex I catalyzes electron transfer between nicotinamide adenine dinucleotide (NADH) and quinone, and the released free energy is used for proton pumping across the 200-Å-wide membrane domain (Fig. 1) (*8–10*). The established proton motive force powers energy-requiring processes in the cell in the form of active transport and synthesis of adenosine triphosphate (ATP) (*11, 12*). The electron and proton transfer processes are fully coupled despite their large spatial separation. Although the overall mechanism of this process still remains unclear, molecular principles recently started to emerge, suggesting that the signal propagates by combined conformational, electrostatic, and hydration changes across the membrane domain (*13–16*).

In addition to the functional elements of the 14 core subunits (*17*), the activity of complex I strongly depends on cardiolipin (*18–20*), an anionic lipid that accounts for 20% of mitochondrial membranes. The activity of all mitochondrial energy-transducing enzymes depends on this lipid (*21, 22*), but unlike, e.g., F_oF_1–ATPase, where the lipid-protein interactions are transient (*23, 24*), complex I interacts with at least 10 tightly bound cardiolipin molecules that strongly modulate the enzymatic activity (*25*). Recent cryo–electron microscopy (cryo-EM) structures (*6, 7*) suggest that some cardiolipins are likely to bind between the proton-pumping membrane subunits, but the functional role of these lipids remains puzzling, considering that they are located up to 100 Å away from the quinone reduction site. The quinone site is located ca. 20 Å above the membrane plane, at the top of a narrow tunnel with an opening to the lipid membrane (Fig. 1).

To elucidate a molecular mechanism by which cardiolipin modulates the activity of complex I, we study here the dynamics of the bacterial enzyme in cardiolipin-containing membranes on millisecond time scales using a combination of coarse-grained and atomistic molecular dynamics (MD) simulation approaches (Fig. 1, inset).

Department Chemie, Technische Universität München (TUM), Lichtenbergstraße 4, D-85747 Garching, Germany.
*These authors contributed equally to this work.
†Corresponding author. Email: ville.kaila@ch.tum.de

RESULTS

Millisecond simulations identify cardiolipin binding sites

To probe the effect of cardiolipin on the structure and dynamics of complex I, we performed coarse-grained molecular dynamics (CG-MD) simulations on millisecond time scales with the bacterial complex I embedded in a lipid membrane containing 20% cardiolipin and compared these to simulations of complex I in POPC (palmitoyl-oleoyl phosphatidylcholine):POPE (palmitoyl-oleoyl phosphatidylethanolamine) membrane without cardiolipin (see Methods; table S1). The coarse-grained simulations suggest that complex I forms contacts with ca. 150 lipid molecules during the simulations (Fig. 2, A and B). However, in stark contrast to POPE and POPC (see Methods) that transiently interact with the protein, cardiolipin molecules form much stronger interactions with the membrane domain of complex I (Fig. 2C and movie S1). The Nqo8 subunit, which provides a hinge between the membrane domain and the hydrophilic domain, strongly interacts with 10 to 15 cardiolipin molecules (Fig. 2, A to C) that are also supported by atomistic MD simulation of the complex I hinge region (see Methods; fig. S1). We also observe prominent binding sites in the CG-MD models at the interface between the antiporter-like subunits, Nqo12/Nqo13 and Nqo13/Nqo14 (Fig. 2A and fig. S2). These sites are close to the broken transmembrane (TM) helix elements, TM12, which are involved in establishing proton channels across the membrane (*15*). The identified binding sites superimpose with cardiolipin molecules recently resolved from cryo-EM structure of mammalian complex I (Fig. 2A) (*6, 7*), suggesting that the cardiolipin interaction sites are conserved across different species, as in the other respiratory enzymes (*22*).

The cardiolipin binding is induced by positively charged surface regions distributed around the membrane domain of complex I (Fig. 2D) and conserved across different complex I isoforms (fig. S3). In addition, phenylalanine residues interact with cardiolipin, particularly in the Nqo8 subunit (Fig. 2, D and E). Notably, phenylalanine residues provide important cardiolipin binding sites also in cytochrome c oxidase (*21*).

Cardiolipin binding affects complex I structure and dynamics

To analyze how cardiolipin affects the global dynamics of complex I, we projected out thermal fluctuations from the CG-MD simulation trajectories using principal components (PC) analysis (*26*). The analysis indicates that complex I undergoes large-scale bending (PC1) and twisting (PC2) motion around the hydrophilic and membrane
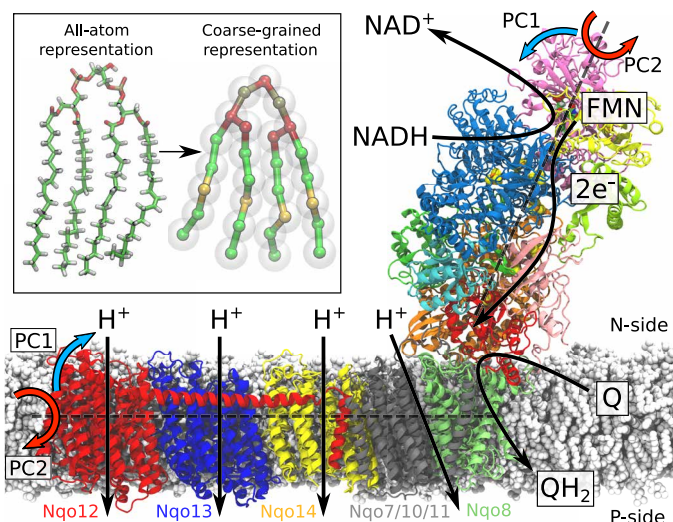
Fig. 1. Structure and function of the respiratory complex I (PDB ID: 4HEA). Electrons are transferred from the NADH/FMN (flavin mononucleotide) site to quinone (Q) via a tunneling wire comprising eight iron-sulfur centers. The Q reduction takes place at the interface between the hydrophilic and membrane domains and couples to the pumping of four protons across the membrane. The global bending (PC1, blue arrows) and twisting motions (PC2, red arrows) are shown along two principal axes (see text). Inset: Comparison of atomistic and coarse-grained molecular simulation models of a cardiolipin molecule.

domains with Nqo8 acting as a hinge (Fig. 1), consistent with global motions identified from previous network models and atomistic simulations of the bacterial complex I in a POPC membrane (*27*). The PC analysis suggests that the coupled twisting-bending motions (PC3) around the hydrophilic and membrane domains is strongly modulated by cardiolipin (Fig. 3A and fig. S4). This global motion displaces half of the membrane and hydrophilic domains (Fig. 3, A and B), and its origin can be traced to cardiolipin bound at the Nqo13/Nqo14 interface and at Nqo8 that induce a 0.4-nm structural shift between TM1 of Nqo8 and helix 1 of Nqo9 (Fig. 3C). These structural rearrangements may have important functional implications as they modulate the accessibility of the quinone to its binding site (see below).

Cardiolipin modulates quinone access and channel formation dynamics

To probe how the quinone dynamics is affected by cardiolipin, we studied the global motions of complex I with and without a quinone molecule inside the quinone tunnel (see Methods; table S1). The quinone finds stable binding sites in the CG-MD simulations at the top ($d \sim 0.5$ to 0.7 nm), slightly below ($d \sim 0.7$ to 0.9 nm), and at the bottom part ($d \sim 3.2$ to 3.7 nm from Tyr[87]) of the cavity (Fig. 4, A and B). These binding sites are consistent with minima observed in free energy profiles of the quinone diffusion along the tunnel derived from a recent atomistic study of complex I in a POPC membrane on nanosecond time scales (*28*). We find that the enzyme undergoes a structural rearrangement in the CG-MD simulations along the bending mode when the quinone molecule moves from the lower edge of the tunnel opening ($d \sim 3.2$ to 4 nm in Fig. 4, A to C) to the upper binding site ($d \sim 0.6$ to 0.8 nm in Fig. 4, A to C). When the quinone passes a strongly bent kink region ($d \sim 2$ nm; Fig. 4B) (*29*) in the middle of the tunnel, we observe a structural rearrangement along the twisting mode that facilitates the substrate motions

along its tunnel, indicated by a transition around $d \sim 2$ nm in Fig. 4A. In contrast, when the quinone is located at the lower edge of the tunnel around $d \sim 3.2$ to 4 nm, complex I samples conformations similar to those observed in simulations of the unbound (*apo*) state (Fig. 4A), suggesting that the initial quinone binding to complex I does not trigger large-scale conformational changes in the protein structure.

The activity of the mammalian complex I is regulated by the so-called active (A)–to–deactive (D) transition (*30*) that is also characterized by structural changes along these twisting and bending modes (*5, 7, 27*). Since we observe a strong correlation between the quinone position in the tunnel and these global motions (Fig. 4 and fig. S5), our findings suggest that quinone dynamics could also be modulated by the structural changes required for the A-D transition in the eukaryotic complex I.

In the absence of cardiolipin in the membrane, the quinone rarely visits the upper binding site, but it finds more stable binding poses at the edge of the tunnel ($d \sim 2.7$ to 3.6 nm; Fig. 4A). We find that complex I also undergoes similar global twisting and bending motions around the hydrophilic and membrane domains as in the cardiolipin-bound state (fig. S4A). However, in stark contrast to the latter, complex I now samples a configurational space that is very similar to that in the different quinone-bound states as well as that in the unbound (*apo*) state of the enzyme, with large fluctuations along the twisting and bending modes throughout the complete distance range ($d \sim 0.6$ to 4.2 nm; Fig. 4A).

This suggests that the twisting and bending motions are now decoupled from the quinone diffusion, indicating that cardiolipin alters the free energy landscape of the enzyme (Fig. 4A). In other words, the bound cardiolipin molecules seem to direct complex I to sample conformations that favor the quinone motion along its cavity and stabilize quinone binding. These findings thus provide a possible molecular explanation to how cardiolipin enhances complex I activity.

The bound cardiolipin enhances the stability of the quinone tunnel (Fig. 4D), in addition to two prominent cavities that are large enough for water molecules to enter from the negatively charged side of the membrane (N-side) to the quinone tunnel. Coarse-grained simulations usually have a tendency to overestimate protein-protein interactions (*31*), and they may therefore underestimate the probability of finding such channels.

One of these putative channels comprises several conserved residues (table S2) and leads from the N-side to the conserved His[38] and Tyr[87] residues of Nqo4 (Fig. 4C and movie S2), which function as proton donors upon reduction of quinone to quinol (QH_2) (*32*). Reprotonation of these residues is thus a prerequisite for complex I to initiate the next catalytic cycle. This cavity is close to a loop region in Nqo7 that also undergoes conformational changes during the A-D transition in the mammalian complex I (*5*). It is important to note that a functional redox-driven pump requires tight gating of the protonation reactions to avoid leaks. A continuous water connectivity between the quinone reduction site and the N-side could thus compromise the proton-pumping machinery. To this end, we find that the stability of this putative channel depends on the position of the quinone in the tunnel (Fig. 4D), suggesting that opening of the channel could be controlled by the redox chemistry of the quinone site.

The second channel forms at the Nqo8/Nqo9 interface near TM1 (Fig. 4D). This channel also starts at the N-side of the membrane, but it leads to a unique kink region in the middle of the quinone tunnel. This region coincides with the putative second quinone binding site around $d \sim 3.2$ to 4.0 nm (Fig. 4C), to which the quinone
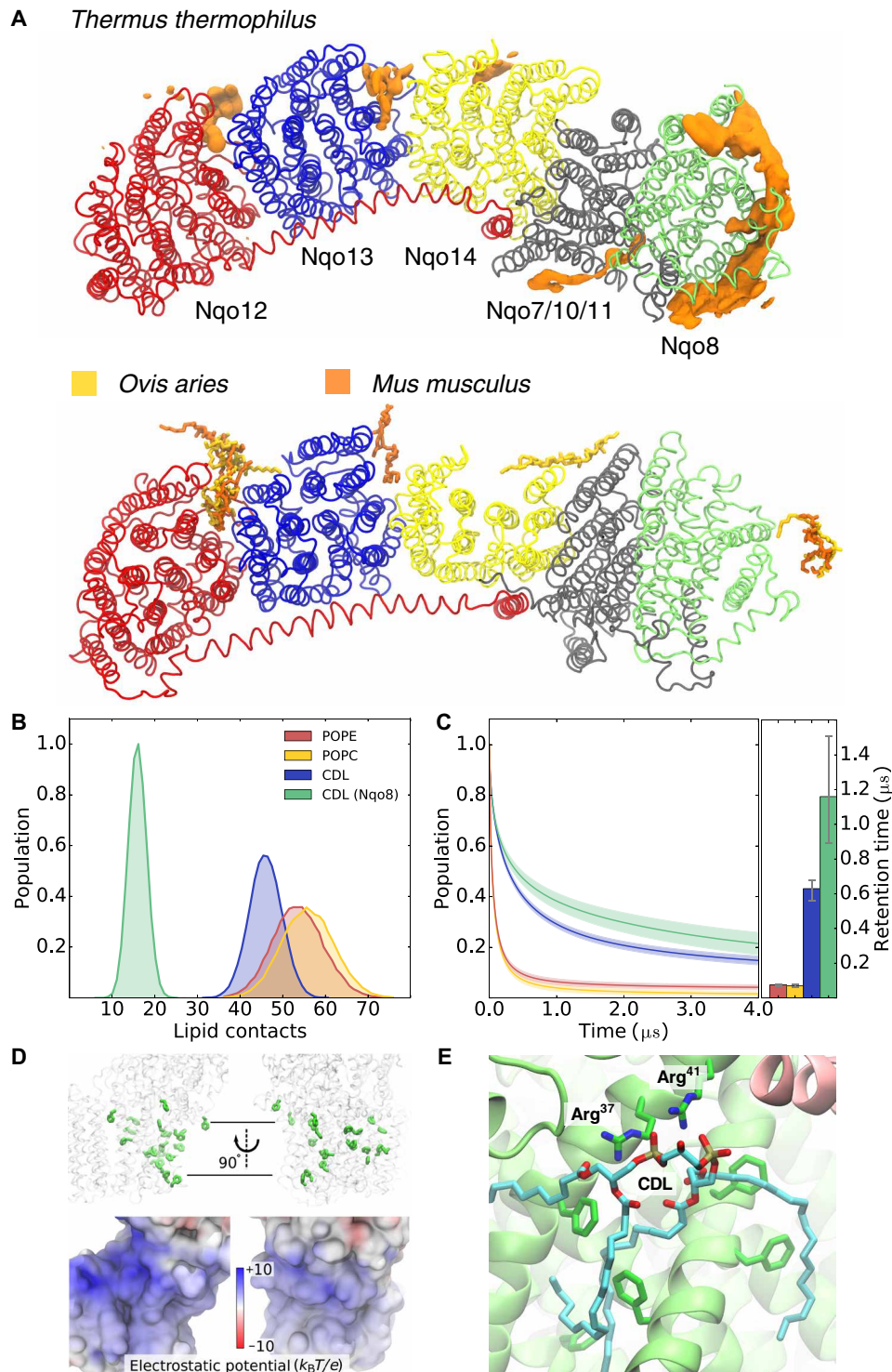
Fig. 2. Cardiolipin binding sites in complex I. (**A**) Top: Cardiolipin (CDL) binding sites identified by CG-MD simulations of complex I from *Thermus thermophilus*. Bottom: Experimentally refined CDL molecules around cryo-EM structures of complex I from *Ovis aries* (*6*) (PDB ID: 5LNK) and *Mus musculus* (*7*) (PDB ID: 6G2J). Note that TM helices 1 to 3 of subunit Nqo14 (subunit in yellow, *T. thermophilus*) are not present in the homologous mammalian ND2 subunits. Both enzymes are viewed from the N-side of the membrane. The supernumerary subunits are omitted for visual clarity. (**B**) Statistics of lipid contacts with the membrane domain of complex I obtained from CG-MD simulations. (**C**) Autocorrelation function of the lipid binding dynamics. The average retention time for CDL is about six times higher than that for POPE or POPC, and ca. 12 times higher than that for Nqo8-bound CDL. (**D**) Top: Membrane-exposed phenylalanine residues in Nqo8. Bottom: Electrostatic potential map of complex I (front view). The color scale ranges from +10 k_BT/e (blue, ca. +260 mV) to −10 k_BT/e (red, ca. −260 mV). (**E**) Putative CDL binding sites in Nqo8, obtained from atomistic simulations, highlighting interactions with Arg[37], Arg[41], and phenylalanine residues.
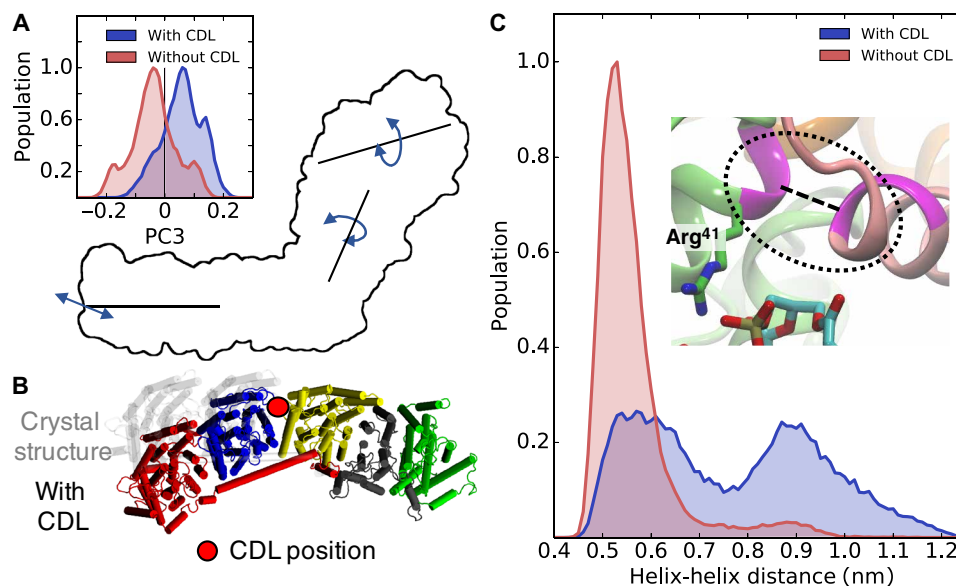
Fig. 3. Effect of CDL on the dynamics of complex I. (**A**) The coupled twisting-bending motion (PC3) is strongly influenced by CDL binding to complex I. The crystal structure is taken as reference conformation for the PC. (**B**) Structural changes in the membrane domain (view from the N-side) induced by cardiolipin binding at the Nqo13/Nqo14 interface. (**C**) Cardiolipin induces conformational changes at the interface between Nqo8 and the hydrophilic domain. The figure shows the distance between helix TM1 of Nqo8 and helix 1 of Nqo9.

molecule docks before exiting to the membrane [see above, and cf. also (*28*)]. The kink site could be relevant for the activation of the proton-pumping machinery in Nqo8, as this site involves many conserved residues that have been linked with human mitochondrial disorders. In simulations without cardiolipin, this channel remains strictly closed, whereas both cardiolipin and quinone bound to the active site strongly enhance opening of this channel (ca. 20 to 30% open channels with quinone and cardiolipin relative to 0.4 to 3% without cardiolipin; Fig. 4D).

DISCUSSION
We have identified here, on the basis of millisecond molecular simulations, how cardiolipin modulates the structure and dynamics of respiratory complex I. Our simulations indicate that cardiolipin binds at specific sites in the membrane domain of complex I, in particular around the Nqo8 subunit and at the interface between antiporter-like subunits Nqo12/Nqo13 and Nqo13/Nqo14. Some of the predicted cardiolipin binding sites are supported by recent cryo-EM structures of the mammalian enzyme (*6*, *7*), whereas others remain to be experimentally validated. Our findings suggest that in addition to the conserved structural core features of complex I, the lipid binding sites are also conserved across different species.

Cardiolipin binding arises from a combination of electrostatic effects between the cardiolipin headgroup and the positively charged protein residues, as well as by dispersive interactions that are dominated by phenylalanine residues. The lipid binding modulates inter-subunit contacts, which are important for the protonation signal propagation across the membrane domain (*16*). This suggests that cardiolipin binding could indirectly also affect the energetics of the proton pumping process. Cardiolipin molecules bound at Nqo8 might also stabilize the highly tilted TM helices of the subunit and affect the energetics of conformational changes linked with the protonation dynamics in this region.

In addition to these local conformational changes, our simulations suggested that complex I also undergoes global conformational changes along the bending and twisting modes that are strongly influenced by cardiolipin. These global motions were recently linked with the active-to-deactive transition (*27*). Our simulations indicate that bacterial complex I undergoes displacement along a similar structural motion when the quinone molecule moves along its tunnel. Recent studies suggest that the bacterial enzyme could also undergo a fast resting-to-active transition (*33–35*). In the absence of cardiolipin, the twisting and bending motions are decoupled from the quinone position along its tunnel, resulting in a lower accessibility of the quinone to its binding site. These findings might provide a molecular explanation to the regulatory role of cardiolipin in complex I: The lipid interactions modify the free energy landscape of complex I along the bending and twisting motions, which in turn modulate the quinone dynamics. Although detailed free energy calculations and atomistic simulations will be important to elucidate structural details of this effect, differences between the simulations with and without cardiolipin are systematic and robust and support that cardiolipin plays a central role in the regulation of the complex I dynamics.

We found that cardiolipin also strongly modulates the formation of two channels from the N-side of the membrane, one around the Nqo7 loop region that could provide protonation pathways for the quinone reduction site, and another channel at the Nqo8/Nqo9 interface that leads to a kink site of the quinone cavity, which could be functionally relevant for activating the pumping machinery (*28*). The functional role of these putative channels can be probed by site-directed mutagenesis experiments, to which our simulations provide important possible candidates (table S2).

In conclusion, our simulations identified how the lipid membrane may modulate global enzyme dynamics that in turn regulate the accessibility of substrates to the active site. These results provide a molecular understanding of complex I function and the role of cardiolipin in the regulation of this respiratory enzyme.
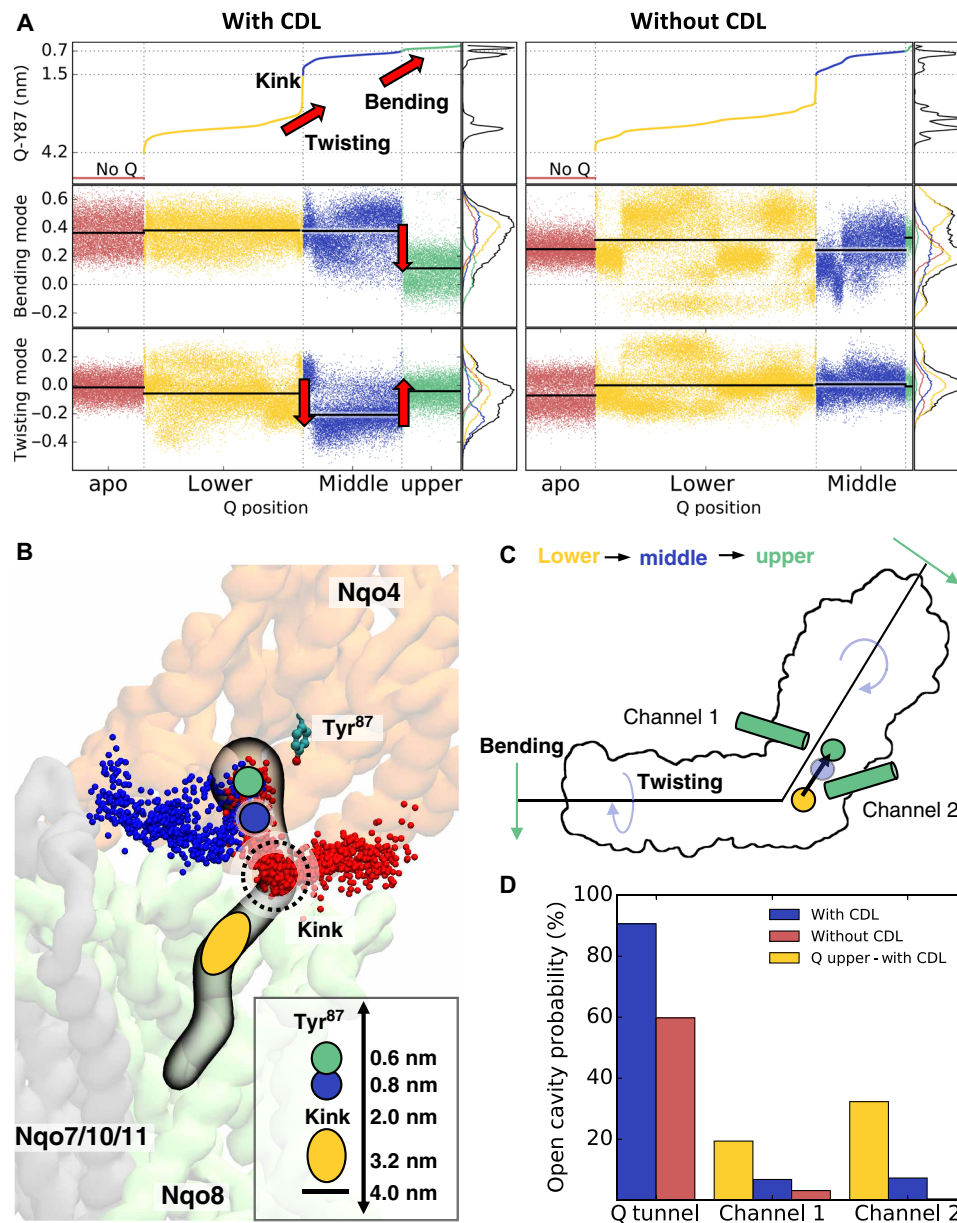
Fig. 4. CDL modulates quinone and complex I dynamics. (**A**) Projection of complex I dynamics with and without CDL along the bending (PC1) and twisting (PC2) modes sorted by the quinone (Q) position along the tunnel. The quinone position is determined by the distance between the quinone headgroup and Tyr[87]. The quinone motion along the cavity is coupled to the complex I motion along the bending and twisting modes. (**B**) Quinone binding sites and channels connecting the quinone cavity with the N-side of the membrane. The upper binding sites ($d \sim 0.6$ to 0.8 nm from Tyr[87]) are shown as green and blue circles, the kink region is marked at $d \sim 2.0$ nm, and the lower binding sites ($d \sim 3.2$ to 4.0 nm) are shown as a yellow oval. Channel 1 (blue) is located near the Nqo7 loop region, and channel 2 (in red) is located at the interface between Nqo8 and Nqo9. Both channels are more stable in simulations with CDL and when quinone is bound at the upper site ($d \sim 0.6$ to 0.8 nm) (see also movie S2). Inset: Distance between Tyr[87] and the kink region and different quinone binding sites. (**C**) Schematic representation of structural changes in complex I that couple to the quinone motion. (**D**) Probability of finding open cavities connected to the quinone binding site.

METHODS

General methodological overview

CG-MD simulations were used to study the interaction between cardiolipin and complex I up to millisecond time scales. The CG-MD trajectories were analyzed by PC analysis (*26*) and protein cavity search algorithms (*36*) to connect the large-scale motion with the quinone dynamics and the formation of channels leading to the quinone cavity. Cardiolipin binding around the Nqo8 region was validated by atomistic MD

simulations to refine the molecular details of the cardiolipin-protein interactions. All CG-MD results were compared with global motions obtained from previous atomistic simulations and network models (*27*), as well as to atomistic simulations of the quinone dynamics (*28*).

CG-MD simulations

Coarse-grained simulation models of complex I were built from the x-ray crystal structure from *Thermus thermophilus* (PDB ID: 4HEA) (*13*).

The protein was embedded in cardiolipin:POPE:POPC and POPE:POPC membranes with lipid ratios of 1:2:2 and 1:1, respectively, solvated using MARTINI water (*37*) and neutralized. The cardiolipin was modeled in its singly pronated state (Fig. 1B). CG-MD simulations were performed at $T = 300$ K using an NpT ensemble with a 20-fs time step, a thermostat with a coupling constant of $\tau_t = 1.0$ ps (*38*), a semi-isotropic Parrinello-Rahman barostat (*39*) with a coupling constant of $\tau_p = 12.0$ ps, and a compressibility of $\chi = 3.0 \times 10^{-4}$ bar^{-1}. Nonbonded interactions were treated with a dielectric constant of 15 and using a cutoff distance of 1.1 nm. The CG-MD simulations were performed in total for ca. 1.2 ms, with each individual trajectory at least 50 μs long. Simulation details are summarized in table S1. All CG-MD simulations were performed using Gromacs 2016.2 (*40*) and the MARTINI 2.2 force field (*37, 41*), and quinone parameters were obtained from the literature (*42*). Visual molecular dynamics (VMD) (*43*), Gromacs select tool, and PLUMED (*44*) were used for analysis and/or visualization, and the Backward script (*45*) was used to convert coarse-grained structures back to an atomistic representation. PyEMMA (*46*) was used to perform the PC analysis (*26*) on the coarse-grained simulation trajectories on the backbone atom level. Caver (*36*) was used with a probe radius of 2.25 Å to identify the channels that are shown in fig. S4.

Atomistic molecular simulation

Molecular models of the membrane domain of complex I were built from the x-ray crystal structure from *T. thermophilus* (PDB ID: 4HEA) (*13*). The system comprised subunits Nqo4/Nqo6/Nqo7, Nqo8 to Nqo11, Nqo14, and the C-terminal end of the HL helix (residues 569 to 605). The model was embedded in a lipid membrane with a cardiolipin:POPC:POPE ratio of 1:2:2 built using the CHARMM-GUI module (*47*), and solvated with TIP3P (*48*) water molecules. Sodium and chloride ions were added to neutralize the system with an ionic strength of ca. 100 mM. The total system comprised ca. 332,100 atoms. After minimization, relaxation, and equilibration using harmonic forces of 2 kcal mol^{-1} Å^{-2} on the protein backbone atoms, MD simulations without restraints were carried out for 0.5 μs at $T = 310$ K using a 2-fs time step, and the long-range electrostatics were treated using the Particle Mesh Ewald approach (*49*). The system was treated using the CHARMM36 force field (*50*) and the density functional theory–derived parameters for the redox-active cofactors. Classical MD simulations were performed with NAMD2 (*51*), and VMD (*43*) was used for visualization and analysis. The electrostatic potential calculations were performed using the Adaptive Poisson-Boltzmann Solver (*52*).

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/5/3/eaav1850/DC1

Fig. S1. Comparison of cardiolipin molecules bound to Nqo8 from coarse-grained and atomistic models.
Fig. S2. The structure of the membrane domain of complex I from *T. thermophilus*.
Fig. S3. Electrostatic potential surface of the bacterial and mammalian complex I.
Fig. S4. Global motion of complex I with and without cardiolipin.
Fig. S5. Projection of complex I dynamics with and without cardiolipin along the bending (PC1) and twisting (PC2) modes.
Table S1. Overview of all coarse-grained MD simulations.
Table S2. Residues within 3 Å of putative channels 1 and 2.
Movie S1. Cardiolipin binding to complex I from a coarse-grained simulation trajectory.
Movie S2. Cardiolipin-induced channel formation dynamics.

REFERENCES AND NOTES

1. U. Brandt, Energy converting NADH: Quinone oxidoreductase (complex I). *Annu. Rev. Biochem.* **75**, 69–92 (2006).
2. J. Hirst, Mitochondrial complex I. *Annu. Rev. Biochem.* **82**, 551–575 (2013).
3. L. A. Sazanov, A giant molecular proton pump: Structure and mechanism of respiratory complex I. *Nat. Rev. Mol. Cell Biol.* **16**, 375–388 (2015).
4. V. Zickermann, C. Wirth, H. Nasiri, K. Siegmund, H. Schwalbe, C. Hunte, U. Brandt, Mechanistic insight from the crystal structure of mitochondrial complex I. *Science* **347**, 44–49 (2015).
5. J. Zhu, K. R. Vinothkumar, J. Hirst, Structure of mammalian respiratory complex I. *Nature* **536**, 354–358 (2016).
6. K. Fiedorczuk, J. A. Letts, G. Degliesposti, K. Kaszuba, M. Skehel, L. A. Sazanov, Atomic structure of the entire mammalian mitochondrial complex I. *Nature* **538**, 406–410 (2016).
7. A. A. Agip, J. N. Blaza, H. R. Bridges, C. Viscomi, S. Rawson, S. P. Muench, J. Hirst, Cryo-EM structures of complex I from mouse heart mitochondria in two biochemically defined states. *Nat. Struct. Mol. Biol.* **25**, 548–556 (2018).
8. A. J. Y. Jones, J. N. Blaza, F. Varghese, J. Hirst, Respiratory complex I in *Bos taurus* and *Paracoccus denitrificans* pumps four protons across the membrane for every NADH oxidized. *J. Biol. Chem.* **292**, 4987–4995 (2017).
9. M. Wikström, G. Hummer, Stoichiometry of proton translocation by respiratory complex I and its mechanistic implications. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 4431–4436 (2012).
10. A. S. Galkin, V. G. Grivennikova, A. D. Vinogradov, →H$^+$/2e stoichiometry in NADH-quinone reductase reactions catalyzed by bovine heart submitochondrial particles. *FEBS Lett.* **451**, 157–161 (1999).
11. P. Mitchell, Coupling of phosphorylation to electron and hydrogen transfer by a chemi-osmotic type of mechanism. *Nature* **191**, 144–148 (1961).
12. M. Yoshida, E. Muneyuki, T. Hisabori, ATP synthase — A marvellous rotary engine of the cell. *Nat. Rev. Mol. Cell Biol.* **2**, 669–677 (2001).
13. R. Baradaran, J. M. Berrisford, G. S. Minhas, L. A. Sazanov, Crystal structure of the entire respiratory complex I. *Nature* **494**, 443–448 (2013).
14. V. R. I. Kaila, Long-range proton-coupled electron transfer in biological energy conversion: Towards mechanistic understanding of respiratory complex I. *J. R. Soc. Interface* **15**, 20170916 (2018).
15. A. Di Luca, A. P. Gamiz-Hernandez, V. R. I. Kaila, Symmetry-related proton transfer pathways in respiratory complex I. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E6314–E6321 (2017).
16. A. Di Luca, M. E. Mühlbauer, P. Saura, V. R. I. Kaila, How inter-subunit contacts in the membrane domain of complex I affect proton transfer energetics. *Biochim. Biophys. Acta Bioenerg.* **1859**, 734–741 (2018).
17. J. Hirst, J. Carroll, I. M. Fearnley, R. J. Shannon, J. E. Walker, The nuclear encoded subunits of complex I from bovine heart mitochondria. *Biochim. Biophys. Acta Bioenerg.* **1604**, 135–150 (2003).
18. M. Fry, D. E. Green, Cardiolipin requirement for electron transfer in complex I and III of the mitochondrial respiratory chain. *J. Biol. Chem.* **256**, 1874–1880 (1981).
19. S. Dröse, K. Zwicker, U. Brandt, Full recovery of the NADH: Ubiquinone activity of complex I (NADH:ubiquinone oxidoreductase) from *Yarrowia lipolytica* by the addition of phospholipids. *Biochim. Biophys. Acta Bioenerg.* **1556**, 65–72 (2002).
20. M. Vos, A. Geens, C. Böhm, L. Deaulmerie, J. Swerts, M. Rossi, K. Craessaerts, E. P. Leites, P. Seibler, A. Rakovic, T. Lohnau, B. de Strooper, S.-M. Fendt, V. A. Morais, C. Klein, P. Verstreken, Cardiolipin promotes electron transport between ubiquinone and complex I to rescue *PINK1* deficiency. *J. Cell Biol.* **216**, 695–708 (2017).
21. C. Arnarez, S. J. Marrink, X. Periole, Identification of cardiolipin binding sites on cytochrome *c* oxidase at the entrance of proton channels. *Sci. Rep.* **3**, 1263 (2013).
22. C. Arnarez, J.-P. Mazat, J. Elezgaray, S.-J. Marrink, X. Periole, Evidence for cardiolipin binding sites on the membrane-exposed surface of the cytochrome *bc*1. *J. Am. Chem. Soc.* **135**, 3112–3120 (2013).
23. A. R. Mehdipour, G. Hummer, Cardiolipin puts the seal on ATP synthase. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 8568–8570 (2016).
24. A. L. Duncan, A. J. Robinson, J. E. Walker, Cardiolipin binds selectively but transiently to conserved lysine residues in the rotor of metazoan ATP synthases. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 8687–8692 (2016).
25. M. S. Sharpley, R. J. Shannon, F. Draghi, J. Hirst, Interactions between phospholipids and NADH: Ubiquinone oxidoreductase (complex I) from bovine mitochondria. *Biochemistry* **45**, 241–248 (2006).
26. C. C. David, D. J. Jacobs, Principal component analysis: A method for determining the essential dynamics of proteins. *Methods Mol. Biol.* **1084**, 193–226 (2014).
27. A. Di Luca, V. R. I. Kaila, Global collective motions in the mammalian and bacterial respiratory complex I. *Biochim. Biophys. Acta Bioenerg.* **1859**, 326–332 (2018).
28. J. Warnau, V. Sharma, A. P. Gamiz-Hernandez, A. Di Luca, O. Haapanen, I. Vattulainen, M. Wikström, G. Hummer, V. R. I. Kaila, Redox-Coupled Quinone Dynamics in the Respiratory Complex I. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E8413–E8420 (2018).

29. J. G. Fedor, A. J. Y. Jones, A. Di Luca, V. R. I. Kaila, J. Hirst, Correlating kinetic and structural data on ubiquinone binding and reduction by respiratory complex I. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 12737–12742 (2017).

30. M. Babot, A. Birch, P. Labarbuta, A. Galkin, Characterisation of the active/de-active transition of mitochondrial complex I. *Biochim. Biophys. Acta Bioenerg.* **1837**, 1083–1092 (2014).

31. M. Javanainen, H. Martinez-Seara, I. Vattulainen, Excessive aggregation of membrane proteins in the Martini model. *PLOS ONE* **12**, e0187936 (2017).

32. V. Sharma, G. Belevich, A. P. Gamiz-Hernandez, T. Róg, I. Vattulainen, M. L. Verkhovskaya, M. Wikström, G. Hummer, V. R. I. Kaila, Redox-induced activation of the proton pump in the respiratory complex I. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 11571–11576 (2015).

33. N. Belevich, C. von Ballmoos, M. Verkhovskaya, Activation of proton translocation by respiratory complex I. *Biochemistry* **56**, 5691–5697 (2017).

34. N. Belevich, M. Verkhovskaya, Resting state of respiratory complex I from *Escherichia coli*. *FEBS Lett.* **590**, 1570–1575 (2016).

35. N. Belevich, G. Belevich, Z. Chen, S. C. Sinha, M. Verkhovskaya, Activation of respiratory Complex I from Escherichia coli studied by fluorescent probes. *Heliyon* **3**, e00224 (2017).

36. E. Chovancova, A. Pavelka, P. Benes, O. Strnad, J. Brezovsky, B. Kozlikova, A. Gora, V. Sustr, M. Klvana, P. Medek, L. Biedermannova, J. Sochor, J. Damborsky, CAVER 3.0: A tool for the analysis of transport pathways in dynamic protein structures. *PLOS Comput. Biol.* **8**, e1002708 (2012).

37. S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, A. H. de Vries, The MARTINI force field: Coarse grained model for biomolecular simulations. *J. Phys. Chem. B* **111**, 7812–7824 (2007).

38. G. Bussi, D. Donadio, M. Parrinello, Canonical sampling through velocity rescaling. *J. Chem. Phys.* **126**, 014101 (2007).

39. M. Parrinello, A. Rahman, Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **52**, 7182–7190 (1981).

40. M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, E. Lindahl, GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25 (2015).

41. D. H. de Jong, G. Singh, W. F. D. Bennett, C. Arnarez, T. A. Wassenaar, L. V. Schäfer, X. Periole, D. P. Tieleman, S. J. Marrink, Improved parameters for the martini coarse-grained protein force field. *J. Chem. Theory Comput.* **9**, 687–697 (2013).

42. D. H. de Jong, N. Liguori, T. van den Berg, C. Arnarez, X. Periole, S. J. Marrink, Atomistic and coarse grain topologies for the cofactors associated with the photosystem II core complex. *J. Phys. Chem. B* **119**, 7791–7803 (2015).

43. W. Humphrey, A. Dalke, K. Schulten, VMD: Visual molecular dynamics. *J. Mol. Graph.* **14**, 33–38 (1996).

44. G. A. Tribello, M. Bonomi, D. Branduardi, C. Camilloni, G. Bussi, PLUMED 2: New feathers for an old bird. *Comput. Phys. Commun.* **185**, 604–613 (2014).

45. T. A. Wassenaar, K. Pluhackova, R. A. Böckmann, S. J. Marrink, D. P. Tieleman, Going backward: A flexible geometric approach to reverse transformation from coarse grained to atomistic models. *J. Chem. Theory Comput.* **10**, 676–690 (2014).

46. M. K. Scherer, B. Trendelkamp-Schroer, F. Paul, G. Pérez-Hernández, M. Hoffmann, N. Plattner, C. Wehmeyer, J.-H. Prinz, F. Noé, PyEMMA 2: A software package for estimation, validation, and analysis of Markov models. *J. Chem. Theory Comput.* **11**, 5525–5542 (2015).

47. S. Jo, T. Kim, V. G. Iyer, W. Im, CHARMM-GUI: A web-based graphical user interface for CHARMM. *J. Comput. Chem.* **29**, 1859–1865 (2008).

48. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein, Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935 (1983).

49. T. Darden, D. York, L. Pedersen, Particle mesh Ewald: An $N \cdot \log(N)$ method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089–10092 (1993).

50. R. B. Best, X. Zhu, J. Shim, P. E. M. Lopes, J. Mittal, M. Feig, A. D. MacKerell Jr., Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone φ, ψ and side-chain χ_1 and χ_2 dihedral angles. *J. Chem. Theory Comput.* **8**, 3257–3273 (2012).

51. J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé, K. Schulten, Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **26**, 1781–1802 (2005).

52. N. A. Baker, D. Sept, S. Joseph, M. J. Holst, J. A. McCammon, Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 10037–10041 (2001).

Citation: A. Jussupow, A. Di Luca, V. R. I. Kaila, How cardiolipin modulates the dynamics of respiratory complex I. *Sci. Adv.* **5**, eaav1850 (2019).

ScienceAdvances

How cardiolipin modulates the dynamics of respiratory complex I

Alexander Jussupow, Andrea Di Luca and Ville R. I. Kaila

ARTICLE TOOLS	http://advances.sciencemag.org/content/5/3/eaav1850
SUPPLEMENTARY MATERIALS	http://advances.sciencemag.org/content/suppl/2019/03/18/5.3.eaav1850.DC1
REFERENCES	This article cites 52 articles, 12 of which you can access for free http://advances.sciencemag.org/content/5/3/eaav1850#BIBL
PERMISSIONS	http://www.sciencemag.org/help/reprints-and-permissions

Use of this article is subject to the Terms of Service

Terminal Electron−Proton Transfer Dynamics in the Quinone Reduction of Respiratory Complex I

Ana P. Gamiz-Hernandez,*,† Alexander Jussupow,† Mikael P. Johansson,†,‡ and Ville R. I. Kaila*,†

†Department Chemie, Technische Universität München (TUM), Lichtenbergstraße 4, Garching D-85747, Germany
‡Department of Chemistry, University of Helsinki, P.O. Box 55, Helsinki FI-00014, Finland

S *Supporting Information*

ABSTRACT: Complex I functions as a redox-driven proton pump in aerobic respiratory chains. By reducing quinone (Q), complex I employs the free energy released in the process to thermodynamically drive proton pumping across its membrane domain. The initial Q reduction step plays a central role in activating the proton pumping machinery. In order to probe the energetics, dynamics, and molecular mechanism for the proton-coupled electron transfer process linked to the Q reduction, we employ here multiscale quantum and classical molecular simulations. We identify that both ubiquinone (UQ) and menaquinone (MQ) can form stacking and hydrogen-bonded interactions with the conserved Q-binding-site residue His-38 and that conformational changes between these binding modes modulate the Q redox potentials and the rate of electron transfer (eT) from the terminal N2 iron−sulfur center. We further observe that, while the transient formation of semiquinone is not proton-coupled, the second eT process couples to a semiconcerted proton uptake from conserved tyrosine (Tyr-87) and histidine (His-38) residues within the active site. Our calculations indicate that both UQ and MQ have low redox potentials around −260 and −230 mV, respectively, in the Q-binding site, respectively, suggesting that release of the Q toward the membrane is coupled to an energy transduction step that could thermodynamically drive proton pumping in complex I.
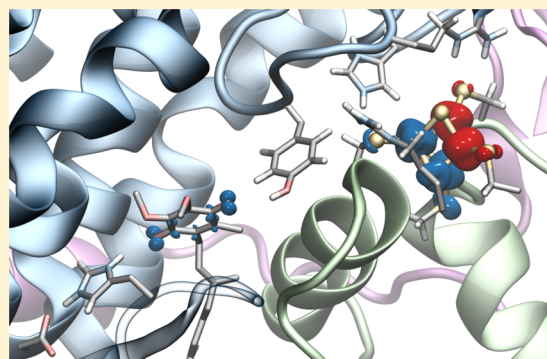
1. INTRODUCTION

Respiratory complex I (NADH:ubiquinone oxidoreductase) is a redox-driven proton pump that reduces quinone (Q) by electron transfer (eT) from nicotinamide adenine dinucleotide (NADH) and couples the energy released in the process to transfer of protons (pT) across the mitochondrial or bacterial membrane,[1−7] (Figure 1). The eight to nine iron−sulfur centers (ISCs) of the hydrophilic domain of complex I are responsible for catalyzing the eT reaction,[8] while the proton pumping entirely takes place in the membrane domain of the enzyme.[1,3−6] Remarkably, the most distant subunits responsible for the pT are separated by up to ca. 200 Å from the eT domain.[9−14] This long-range coupling is believed to take place by combined conformational and electrostatic transitions, in which the terminal eT step between the N2 ISC and Q is likely to trigger the proton-pumping activity,[1,3−6,8−18] but the exact molecular mechanism remains elusive.

Although the exact binding mode of Q has not been experimentally resolved in complex I, structural and biochemical studies[2,9−14] show that the Q-binding site has a unique location in complex I.[19] In contrast to many membrane proteins, in which the Q site is embedded within the membrane domain, the Q site in complex I is located ca. 20 Å above the membrane plane and ca. 8−11 Å from the terminal N2 ISC[20] (see Figure S1 for nomenclature of ISCs in complex I), with Tyr-87[21] and His-38 interacting with the Q headgroup (Figure

1).[22] Although proton-coupled electron transfer (PCET) reactions of Qs in different solvents and oxidoreductases have been intensively studied over the years,[23−27] a detailed understanding of the Q-reduction-linked PCET processes in complex I is still unclear. Recent electrochemical studies[28] indicate that the redox potential of ubiquinone (UQ) for the Q/QH$_2$ redox couple is unusually low, in the −300 mV range, which is considerably downshifted from the E_m of Q in membranes of ca. +90 mV; electron paramagnetic resonance studies[8,29] show that the E_m of the terminal N2 ISC is ca. −150 mV in *Escherichia coli* (all E_m values are relative to NHE if not otherwise stated). Moreover, experiments suggest[28] that semiquinone (SQ) does not accumulate during turnover (but cf. ref 30), further supporting that the SQ species is thermodynamically unstable and transient. Interestingly, recent experiments[19] also indicate that the eT rate in complex I may be modulated by the redox state of the terminal N2 cluster, which in turn may be important for understanding the eT dynamics along the ISC chain. Importantly, elucidating the thermodynamics of the terminal eT step that is linked to the Q reduction in complex I is important for understanding how the redox energy is employed for proton pumping.
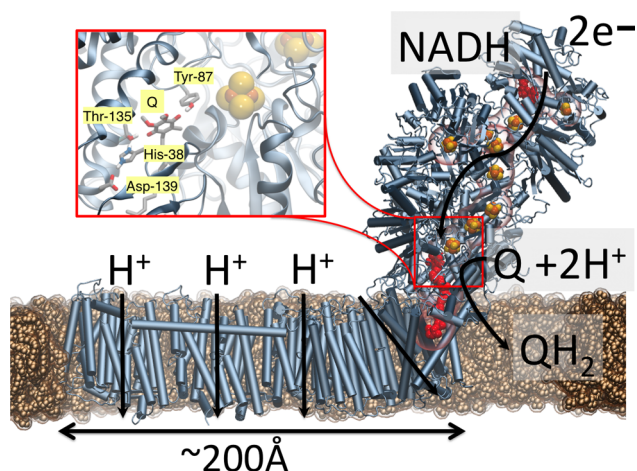
Figure 1. Complex I, the initial electron acceptor in aerobic respiratory chains. Electron transfer from NADH to Q via eight to nine ISCs leads to pumping of protons in the membrane domain of complex I. (inset) Q and N2 sites are located about 20 Å above the membrane plane; Q has been modeled in computationally (see the Computational Methods section).

In order to address the mechanism of the Q reduction by the terminal N2 cluster and the E_m of Q in complex I, we performed here first-principles quantum chemical calculations (QM), hybrid quantum mechanics/classical mechanics (QM/MM) simulations, classical molecular dynamics (MD) simulations, as well as continuum Poisson–Boltzmann (PB) electrostatics calculations on molecular models that are based on the experimentally resolved structure of complex I from *Thermus thermophilus*,[11] for which no Q-bound structure has yet been released. Q-binding models were constructed for both UQ and menaquinone (MQ), which complex I can employ as a substrate, for example, under anaerobic conditions.[31] In order to control the localization of the electron between N2 and Q, we develop here a computational approach, where we initiate the simulations from reduced or oxidized QM subsystems that are created by merging converged electronic structures from the individual donor, N2, and acceptor, Q, subsystems together. This allows us to study the energetics and dynamics of the eT process from *localized* diabatic electronic states of the reduced/oxidized N2 and Q. Our combined results identify important UQ- and MQ-binding configurations within the Q-binding site of complex I. We also characterize the E_m of UQ/MQ in these states and show how the terminal eT from N2 is linked to coupled protonation changes of nearby residues.

2. RESULTS AND DISCUSSION

2.1. Binding Modes of Q in the Active Site of Complex I. Our hybrid QM/MM simulations, quantum chemical density functional theory (DFT) models, as well as classical MD simulations of UQ and MQ in different oxidation states show that both species can bind in two possible conformations within the binding pocket, shown in Figure 2 for UQ and Figure S3 for MQ. In the *hydrogen-bonded* binding mode, UQ/MQ forms a hydrogen-bonded interaction with both His-38 and Tyr-87, while in the *stacking* binding mode, UQ/MQ interacts dispersively with His-38, while retaining the hydrogen bond with Tyr-87. Comparison of computed IR spectra of these conformations predict a ca. 200 cm^{-1} shift on the His-38 Nε-H bond, which could be employed to spectroscopically identify the two binding poses (Figure S7) by using, for example,
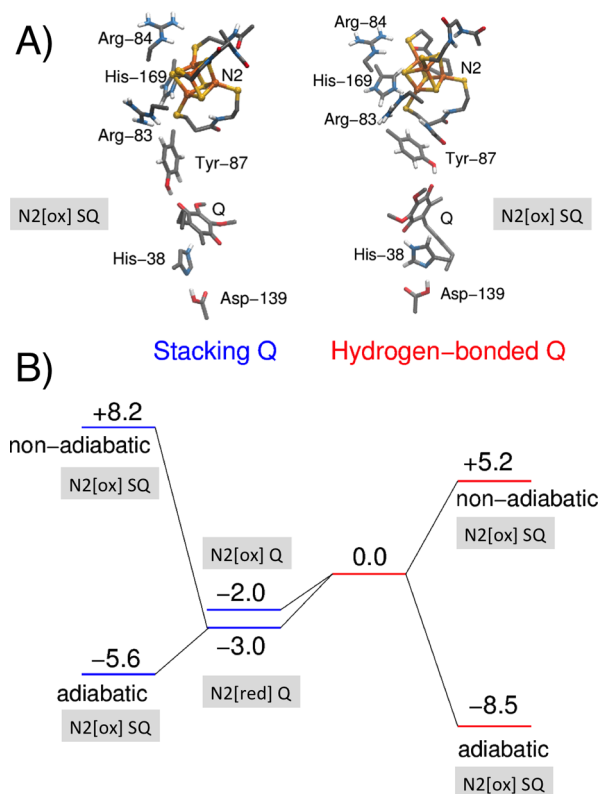

Figure 2. (A) Structure and (B) B3LYP-D3/def2-TZVPP/ε = 4 energetics (in kcal mol^{-1}) of hydrogen-bonded (right) and stacked (left) binding modes of UQ in complex I. Q and SQ refer to oxidized quinone and semiquinone species, while N2[ox] and N2[red] refer to oxidized (2Fe[II]2Fe[III]) and reduced (3Fe[II]1Fe[III]) states of N2, respectively. Nonadiabatic and adiabatic eT energetics are calculated by relaxing the initial electron donor state (N2/Q) or by relaxing both the initial (N2[red]/Q) and final (N2[ox]/SQ) states, respectively.

Fourier-transform IR methodology as applied for Photosystem II.[32] While UQ swaps between these two binding modes on nanosecond time scales in the oxidized state (Figure S2), MQ forms a stronger stacking interaction with His-38 in the oxidized state and favors more strongly the hydrogen-bonded interaction in the SQ state. This one-electron reduced UQ/MQ species forms an anionic semiquinone (Q$^{\bullet/-}$), without proton abstraction of the surrounding residues, consistent with previous results.[17] Our QM calculations further indicate that the stacked Q conformation is favored by ca. 2 kcal mol^{-1} over the hydrogen-bonded binding mode and is weakly affected by the reduction of N2 (Figures 2 and S18). On the basis of the computed energetics, we expect that the hydrogen-bonded mode would have a low occupation, while structural studies might identify only the stacked conformation. In contrast to the oxidized UQ, we find that the SQ favors the hydrogen-bonded binding mode by ca. 2 kcal mol^{-1} (Table S1), whereas the menasemiquinone (i.e., MQ$^{\bullet/-}$) favors the hydrogen-bonded binding mode by ca. 5 kcal mol^{-1} due to the stronger dispersive π–π interaction between His-38 and the MQ rings in the oxidized state (Table S2). Our QM/MM calculations support these results (Figure 3), and a similar behavior is also found in our classical MD simulations (Figure S2). Structural comparison of the two binding modes for the DFT cluster models (Figure S4) shows that there is little variation between the Q and the N2 distances, but our MD simulations indicate that, in
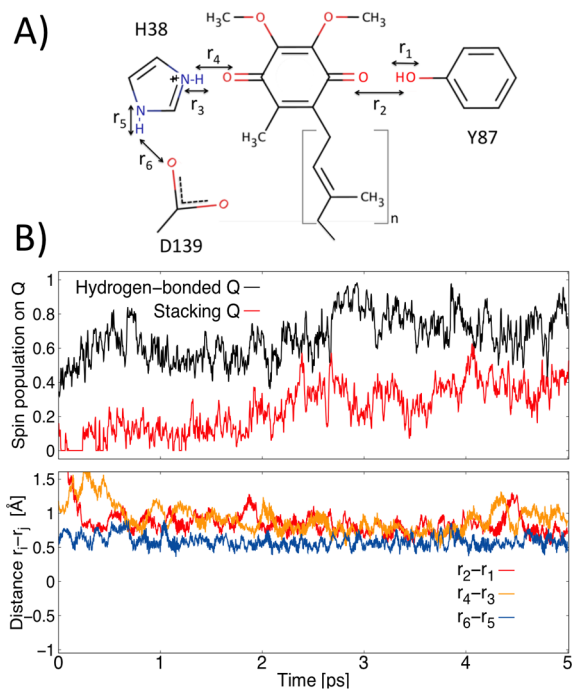
Figure 3. (A) pT distances between UQ and the proton donors, Y87 (r_2-r_1) and H38 (r_4-r_3), as well as between H38 and D139 (r_6-r_5). (B) QM/MM dynamics of eT between N2 and UQ in the stacked (in red) and hydrogen-bonded conformations (in black). The hydrogen-bonded conformation leads to formation of SQ, indicated by the higher spin population relative to the stacked conformation. The fractional spin occupation on the UQ in the stacked conformation might result from DFT charge transfer problems[38] (see Figure S12). The early steps of eT indicate that, within the first 0.5 ps, the hydrogen bonds with Y87 and H38 become stronger but no pT is observed.

the stacked conformation, the N2–Q distance is ca. 0.5 Å shorter than in the hydrogen-bonded conformation (Figures S5 and S6), which could affect the rate of eT from N2 (see below).

2.2. Energetics of eT between N2 and Q. By combining DFT (B3LYP-D3) molecular orbitals obtained from the UQ (or MQ) and N2 broken-symmetry spin-flip subsystems, we converged the electronic structure into localized diabatic electronic states of reduced/oxidized N2 and Q/SQ (MQ/MSQ). We find that the UQ reductions become thermodynamically favored by ca. −9 kcal mol^{-1} in the hydrogen-bonded conformation as compared to ca. −4 kcal mol^{-1} in the stacked conformation. This indicates that the first eT step may couple to a conformational change of the Q from the stacked to the hydrogen-bonded binding mode. The conformational change is also reflected in the rather large nonadiabatic eT energy gap of ca. 11 kcal mol^{-1}, disfavoring formation of SQ, which is obtained without relaxing the geometry of the electron-accepting Q subsystem. In contrast, the nonadiabatic eT energy is reduced to ca. 5 kcal mol^{-1} in the hydrogen-bonded binding mode. This suggests that the thermodynamics of the eT between N2 and Q is nearly degenerate in the stacked conformation (Table S1). We find that the various spin-state configurations, that is, different broken-symmetry[33] combinations of the unpaired electrons around the individual high-spin iron centers leading to the same total spin at the ISCs, are energetically quite different. We use the lowest energy configuration, but we note that other choices may tune the relative stability of the reduced N2 center by up to ca. 6 kcal mol^{-1}. Interestingly, in our DFT models of the other ISCs, we

do not observe such a large energy difference between different spin-state configurations, which may relate to the uniquely connected cysteine residue, Cys-45/Cys-46.[2] Moreover, we observe that the relative stability of SQ is qualitatively similar using other density functionals (Table S1). For MQ, the eT from N2 to MQ is exergonic by ca. 2 kcal mol^{-1} and ca. 9 kcal mol^{-1} in the stacked and hydrogen-bonded binding modes, respectively, suggesting that Q and MQ might be redox tuned in a similar way in the Q-binding pocket (Table S2).

In order to probe the effect of the protein environment and estimate E_m values, we performed classical PB continuum electrostatics calculations with Monte Carlo (MC) sampling of surrounding residues in different protonation/redox states (Figures S8 and S9). For UQ, we obtain an average E_m for the Q/SQ redox couple of −260 mV with N2[ox] and −320 mV with N2[red] in the hydrogen-bonded conformation that contrast the average E_m of −380 mV for the stacked Q conformation with N2[ox]. For the N2 center, we obtain an E_m of −300 mV, which is considerably upshifted by ca. 200 mV from the other ISC of complex I (Figures S8 and S9). Although, somewhat downshifted from the experimental E_m values of N2 (Table S1), these values are in qualitative agreement with the difference reported for E_m values between N2 and other ISCs in complex I.[20] The computed values are not strongly affected by applying other dielectric constants in the PB framework (Figure S10) or by the number of titrated ISCs included in the calculations (Table S4). Moreover, the values are also in qualitative agreement with electron affinities calculated based on the QM cluster models, where we find that the relative electron affinity of N2 is ca. 200 mV higher than that for a QM model of the N4 center model, giving an electron affinity of N2 in the −50 mV range, based on the employed experimental E_m of N4 at ca. −250 mV (Table S1). Our calculations suggest that the negative charge of the N2 center downshifts the electron affinity of Q by ca. 400 mV, which would place the E_m of Q near −300 mV, considering a reference value of +90 mV in membranes.

Similar to UQ, we obtain for MQ in its hydrogen-bonded binding mode an E_m value of −230 mV (Table S5), supporting that the protein environment tunes the redox potentials of both UQ and MQ in a similar way, although MQ has a ca. 170 mV lower E_m value in membranes relative to UQ.[34] For the stacked MQ, we obtain a strongly downshifted E_m, most likely due to the conformational switching between the two conformations upon reduction of MQ.

The E_m values calculated at the PB level are thus consistent with experiments[28] and the QM calculations, suggesting that the Q/SQ redox couple is less than −300 mV.[28,35] Interestingly, the PB calculations suggest that reduction of N2 increases the proton affinity of a nearby His-169, consistent with site-directed mutagenesis studies by Zwicker and co-workers[36] (Table S3). Our calculations thus show that the binding mode of Q, hydrogen-bonded or stacked, may modulate its redox potential, while the binding energetics itself is linked to the redox state of the N2 center. A similar stacking-induced redox-tuning effect has also been suggested for Photosystem I, where a low-potential Q stacks with a tryptophan residue.[37]

We next probed the energetics for the second eT between N2 and SQ, initiated by re-reduction of N2. Consistent with previous simulations,[17] our DFT models suggest that the process is coupled to proton abstraction from surrounding Tyr-87 residues, whereas no complete proton abstraction from His-

38 was observed, leading to the effective formation of QH⁻. We obtain an overall endergonicity of ca. 20 kcal mol⁻¹ in the DFT models, which might result from the protonated His-169 that disfavors the reduction of N2. However, DFT calculations suggest that deprotonation of His-169 could stabilize this eT by ca. 5.6 kcal mol⁻¹, which is also supported by our PB calculations. Taken together, the results suggest that reduction of N2 is linked to deprotonation of His-169 (see below and Table S3), further stabilizing the eT process.

2.3. Simulating the eT Dynamics. In order to address the eT dynamics between N2 and UQ, we performed QM/MM MD simulations, with both UQ and N2 forming the QM region and the remaining system treated classically (Figure S11). Also, for these models, the reduced donor and oxidized acceptor subsystems were generated by merging together their individual molecular orbitals to generate initial eT states with the electron on N2. The results, shown in Figure 3, indicate that the conformation of Q indeed modulates the rate of eT, consistent with the results obtained from the QM calculations and the PB calculations of E_m values. In the hydrogen-bonded conformation, we observe a rapid eT between N2 and UQ (Figure 3), whereas in the stacked conformation the SQ formation is slowed down. On the basis of a Moser–Dutton treatment,[39,40] by using the calculated ΔG values (−40 and +80 mV), N2–UQ *edge-to-edge* distance (10.5 Å), generic reorganization energy (λ = 0.7 eV), and protein-packing density (ρ = 0.76), we obtain predicted eT rates of ca 1×10^{-7} and 6×10^{-7} s⁻¹ for the hydrogen-bonded and stacked conformations, respectively, suggesting that the eT could nevertheless be possible from both binding conformations. However, an accurate estimation of eT requires an explicit calculation of the electronic overlap, which is outside the scope of the present work.

As DFT spin energetics are known to be sensitive to the amount of exact exchange (EXX), we employed functionals with different amount of EXX in order to establish the reliability of the results. We find that the SQ is somewhat more favored when decreasing the EXX to 10% (using the TPSSh functional[41]), while increasing the EXX to 50% (using BHLYP) or using the range-corrected CAM-B3LYP functional[42] decreases the stability of SQ (Figure S12). This indicates that the B3LYP-D3/MM calculations might somewhat overestimate the stability of SQ. However, the overall conclusions remain the same with all functionals employed, validating the robustness of the results. We do not observe any linked deprotonation of the surrounding Tyr-87 or His-38 residues, but the eT process is coupled to a spontaneous reorganization to the hydrogen-bonded conformation, further supporting the conformational reorganization of Q in the eT process. Interestingly, the eT also seems to couple to reorganization of the Arg-84 side chain near the N2 center that transiently forms a contact with the N6B center upon oxidation of the former (Figure S13).

In order to understand the role of Arg-84 and His-169 on N2 pH redox potentials,[43−45] we prepared in silico H169M and R84A mutations and recomputed the E_m values at different pHs (Figure S14). We estimated in our models that R84A downshifts the N2 E_m value by ca. 250 mV and shows almost no pH dependence, while the H169 M mutation downshifts the E_m value by 150 mV and has a pH dependence of ca. 30 mV/pH unit. His-169 has previously been suggested to function as a redox Bohr group for N2,[43,44] whereas our results suggest that a combined conformational change between Arg-84 and His-169 could also be linked to N2 reduction and in part contribute to the pH dependence of the N2 center.[45] The influence of other

mutations[43−45] could further shed light into the dynamics of the eT process and will be addressed in future work.

In order to address the dynamics of the second eT step from N2 to SQ, we re-reduced the N2 center and initiated QM/MM MD simulations. We find that, upon deprotonation of His-169 (Nε), as suggested by our PB calculations (see above), the electron moves within 0.3 ps to the SQ, which in turn, further triggers a rapid deprotonation of His-38 and Tyr-87 on the ca. 0.3−1.3 ps time scale, suggesting that the pT immediately follows the eT process (Figure 4). Interestingly, in the QM/
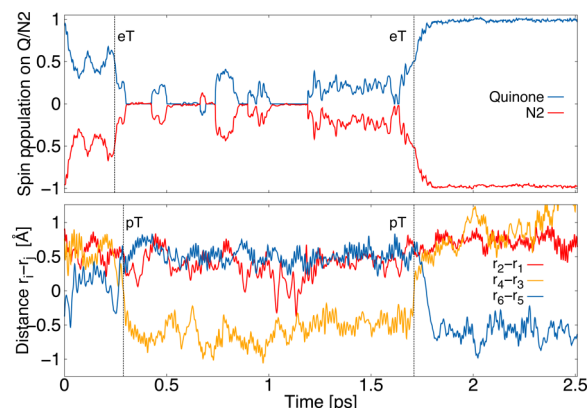

Figure 4. Dynamics of the second eT from N2 to SQ (top) upon deprotonation of His-169 and Tyr-87 (partial deprotonation at 1 ps, below). The eT process is coupled with pT from His-38 to SQ forming QH⁻.

MM trajectory, we also observe a back eT at around ca. 1.7 ps, which follows a similar but reverse PCET process, suggesting that the second eT from N2 to SQ is nearly isoenergetic; this further supports that the E_m of the SQ/QH₂ process is close to that of N2. We were not able to observe the second eT from N2 without deprotonation of His-169 (Figure S15), possibly due to the limited time scale of the simulations, whereas deprotonation of the Nδ proton on His-169 also resulted in a fast (>0.1 ps) eT process, coupled to a similar PCET from His-38 and Tyr-87 (Figure S16). We could not identify likely proton donors that would participate in this putative deprotonation step, but we observe that the Q-binding site is accessible to water from the bulk (Figure S17). Our QM/MM MD simulations, which treat both the eT and pT processes between the N2 and Q quantum mechanically, thus support our previous results of the Q-site dynamics, suggesting that Tyr-87 and His-38 may act as local proton donors in the QH₂ formation for the second eT step (Figure 5B).[17] For the first eT step, our simulations suggest that SQ formation is not linked to proton uptake (Figure 5A). Our classical MD simulations of the resulting QH₂/deprotonated Tyr-87/His-38 state were further suggested to trigger conformational changes in the membrane domain Nqo8 that may in turn activate the proton pumping machinery.[15]

3. CONCLUSIONS

Despite significant structural and biochemical work on complex I, exact structural information on the Q-binding poses in its usual binding pocket, ca. 20 Å above the membrane plane, has not been available. We have shown here by computational methods that both UQ and MQ can bind in two energetically possible binding modes to complex I: one with a hydrogen-bonded interaction with Tyr-87/His-38 of subunit Nqo4, and
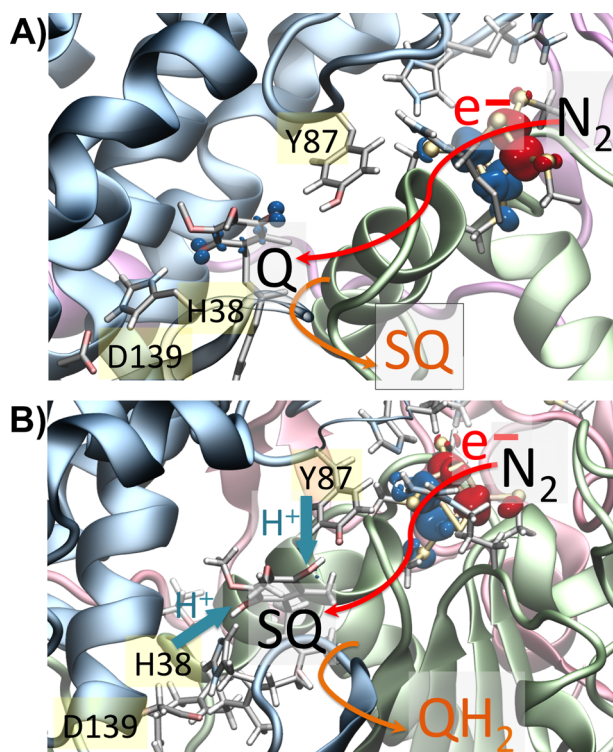
Figure 5. Sequence of eT (red arrow) and pT (blue arrow) between N2 and UQ based on QM/MM MD calculations. (A) First eT from N2 to UQ in the hydrogen-bonded conformation forming SQ. (B) Second electron moves from N2 to SQ and couples with pT from H38 and Y87, leading to formation of QH_2/QH^-.

one where the hydrogen bond with His-38 is replaced by a stacked conformation. The latter conformation was found to be energetically favored upon N2 reduction, and the Q-binding mode was further found to tune the E_m values as dispersive $\pi-\pi$ interactions cause smaller shifts in redox potentials relative to hydrogen bonds. We obtained a redox potential of Q that is unusually low due to its interaction with N2. Our QM/MM MD simulations of the first eT steps between the reduced N2 and Q suggest that eT is linked to a conformational change of Q between stacked and hydrogen-bonded conformations that increases its electron affinity. The second eT step was further found to lead to a local PCET from Tyr-87/His-38 of subunit Nqo4 and is strongly favored by deprotonation of His-169 near the N2 center (Figure 4). Similar binding poses and E_m values were also observed for MQ. Our calculations suggest that redox tuning of Q could be important for controlling the rate of the terminal eT steps in complex I, while local PCET in the Q chamber may in turn be responsible for triggering the proton pump in the membrane domain of complex I. Both eT steps between N2 and UQ were found to be nearly isoenergetic, suggesting that there might not be a large energy drop between NADH (E_m = −320 mV) and UQ/MQ (E_m <−300 mV). Moreover, because both Q (+90 mV) and MQ (−80 mV) have higher E_m values in membranes, this further indicates that the main energy transduction event that is coupled to proton pumping could follow after Q release from its binding site near N2.

4. COMPUTATIONAL METHODS

DFT cluster and QM/MM models of complex I were built based on the X-ray structure of complex I from *T. thermophilus* (PDB ID:

4HEA).[46] The model comprised subunits Nqo4, Nqo5, Nqo6, Nqo7, Nqo8, and Nqo9, the Q and three ISCs N2, N6B, and N6A (Figure S11). Protonation states based on the crystal structure were assigned by PB continuum electrostatics calculations.[47,48] Classical MD simulations were performed by using the whole protein and considering the UQ/MQ in oxidized (Q), SQ (i.e., $Q^{\bullet/-}$), and reduced (Q^{2-}) states, while the ISCs were fixed in their N2[ox] state ($2Fe^{3+} 2Fe^{2+}$) with atomic partial charges that included the inorganic iron−sulfur and four ligated cysteine residues. For the QM calculations, we also considered the N2 cluster in its N2[red] state ($1Fe^{3+} 3Fe^{2+}$). The system was classically relaxed with NAMD2[49] for 10 ns at T = 310 K using a 1 fs integration time step with harmonic restrains on the Cα atoms and by employing the CHARMM27 force field.[50] The classical MD simulations were followed by QM/MM geometry optimization and first-principles dynamics for 5 ps for each state. The QM region consisted of three models: the N2 model, the Q model, and the combined N2−Q model. The N2 model comprised the N2 center and residues Cys-45, Cys-46, Cys-111, Cys-140, Arg-86, Arg-83, and His-169. The Q model comprised the Q and residues His-38, Asp-139, Thr-135, and Tyr-87, while the combined N2−Q model comprised both subsystems (see Figure S11). The QM/MM simulations were performed at the B3LYP-D3/def2-TZVP(Fe)/def2-SVP(S, C, H, N, O) level,[51−53] and the N2 center was modeled with antiferromagnetic coupling using the broken-symmetry spin-flip approach.[33] QM cluster models were optimized at the dispersion corrected B3LYP-D3/def2-TZVP(Fe, S)/def2-SVP(C, H, N, O).[54,55] Single point energies were computed for all atoms at B3LYP-D3 and TPSSh-D3/def2-TZVPP levels (see Table S1). The QM models comprised the Q, the N2 cluster, and residues His-38, Asp-139, Thr-135, Tyr-87, Cys-45, Cys-46, Cys-111, Cys-140, Arg-86, and Arg-83 (Figure S4); terminal carbon atoms were fixed during the optimization process. Solvation effects were treated with COSMO[56] with a dielectric constant ε = 4. Calculations were performed with the CHARMM/TURBOMOLE[57] interface, TURBOMOLE,[58] and NWChem.[59] VMD was used for visualization.[60] Redox potentials were calculated based on electrostatic potentials obtained by solving numerically the linearized PB equation using APBS[48,61] and MC titration sampling.[47,62] The protein was described using atomic partial charges, embedded in an inhomogeneous dielectric continuum with dielectric constants of ε_p = 4 inside the protein and ε_w = 80 for the bulk water. The boundary interface between the protein and the solvent was calculated by the molecular surface routine implemented in APBS, using a solvent probe radius of 1.4 Å and modeling an implicit ionic strength of 100 mM potassium chloride. The redox potential was computed as a difference of electrostatic free energy shifts between a model compound in water and the model compound in the protein.[25,62−64] To assess the influence of using a small or large model of the protein, we performed redox potential computations using the whole crystal structure and compared them with the small QM/MM models observing only a small deviation of Q/SQ and N2 E_m values of ca. 40 mV (see Table S4). We also addressed the influence of the redox state of other ISCs on the Q/SQ and N2 values, suggesting that, in the current PB model, only N6B and N6A strongly influence the Q/SQ and N2 E_m values (Table S4). A reference value of the UQ (MQ) Q/SQ redox couple was estimated to −188 mV (−241 mV) in water by B3LYP/def2-TZVP calculation, based on the experimental E_m values of UQ and MQ in DMF and acetonitrile,[25] (Table S5) by the approach of Knapp and co-workers.[65−67] All our reported E_m values are given relative to the NHE.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/jacs.7b08486.

> Detailed energetics, coordinates of QM and QM/MM models, and supplementary figures and tables (PDF)
>
> Movie of the second eT (GIF)

■ AUTHOR INFORMATION

Corresponding Authors
*ana.gamiz@tum.de
*ville.kaila@ch.tum.de

ORCID ●
Mikael P. Johansson: 0000-0002-9793-8235
Ville R. I. Kaila: 0000-0003-4464-6324

Notes
The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

■ REFERENCES

(1) Brandt, U. *Annu. Rev. Biochem.* **2006**, *75*, 69−92.
(2) Sazanov, L. A.; Hinchliffe, P. *Science* **2006**, *311*, 1430−1436.
(3) Verkhovskaya, M.; Bloch, D. A. *Int. J. Biochem. Cell Biol.* **2013**, *45*, 491−511.
(4) Hirst, J. *Annu. Rev. Biochem.* **2013**, *82*, 551−575.
(5) Sazanov, L. A. *Nat. Rev. Mol. Cell Biol.* **2015**, *16*, 375−388.
(6) Wikström, M.; Sharma, V.; Kaila, V. R. I.; Hosler, J. P.; Hummer, G. *Chem. Rev.* **2015**, *115*, 2196−2221.
(7) Hirst, J.; Roessler, M. M. *Biochim. Biophys. Acta, Bioenerg.* **2016**, *1857*, 872−883.
(8) Verkhovskaya, M. L.; Belevich, N.; Euro, L.; Wikström, M.; Verkhovsky, M. I. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 3763−3767.
(9) Zickermann, V.; Kerscher, S.; Zwicker, K.; Tocilescu, M. A.; Radermacher, M.; Brandt, U. *Biochim. Biophys. Acta, Bioenerg.* **2009**, *1787*, 574−583.
(10) Hunte, C.; Zickermann, V.; Brandt, U. *Science* **2010**, *329*, 448−451.
(11) Baradaran, R.; Berrisford, J. M.; Minhas, G. S.; Sazanov, L. A. *Nature* **2013**, *494*, 443−448.
(12) Zickermann, V.; Wirth, C.; Nasiri, H.; Siegmund, K.; Schwalbe, H.; Hunte, C.; Brandt, U. *Science* **2015**, *347*, 44−49.
(13) Fiedorczuk, K.; Letts, J. A.; Degliesposti, G.; Kaszuba, K.; Skehel, M.; Sazanov, L. A. *Nature* **2016**, *538*, 406−410.
(14) Zhu, J.; Vinothkumar, K. R.; Hirst, J. *Nature* **2016**, *536*, 354−358.
(15) Wikström, M.; Hummer, G. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 4431−4436.
(16) Kaila, V. R. I.; Wikström, M.; Hummer, G. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111*, 6988−6993.
(17) Sharma, V.; Belevich, G.; Gamiz-Hernandez, A. P.; Rog, T.; Vattulainen, I.; Verkhovskaya, M. L.; Wikström, M.; Hummer, G.; Kaila, V. R. I. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, 11571−11576.
(18) Di Luca, A.; Gamiz-Hernandez, A. P.; Kaila, V. R. I. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, E6314−E6321.
(19) de Vries, S.; Dörner, K.; Strampraad, M. J. F.; Friedrich, T. *Angew. Chem., Int. Ed.* **2015**, *54*, 2844−2848.
(20) Ohnishi, T. *Biochim. Biophys. Acta, Bioenerg.* **1998**, *1364*, 186.
(21) Tocilescu, M. A.; Fendel, U.; Zwicker, K.; Dröse, S.; Kerscher, S.; Brandt, U. *Biochim. Biophys. Acta, Bioenerg.* **2010**, *1797*, 625−632.
(22) Sinha, P. K.; Castro-Guerrero, N.; Patki, G.; Sato, M.; Torres-Bacete, J.; Sinha, S.; Miyoshi, H.; Matsuno-Yagi, A.; Yagi, T. *Biochemistry* **2015**, *54*, 753−764.
(23) Warren, J. J.; Tronic, T. A.; Mayer, J. M. *Chem. Rev.* **2010**, *110*, 6961−7001.
(24) Weinberg, D. R.; Gagliardi, C. J.; Hull, J. F.; Murphy, C. F.; Kent, C. A.; Westlake, B. C.; Paul, A.; Ess, D. H.; McCafferty, D. G.; Meyer, T. J. *Chem. Rev.* **2012**, *112*, 4016−4093.
(25) Prince, R. C.; Dutton, P. L.; Bruce, J. M. *FEBS Lett.* **1983**, *160*, 273−276.
(26) Gunner, M. R.; Madeo, J.; Zhu, Z. *J. Bioenerg. Biomembr.* **2008**, *40*, 509−519.
(27) Trumpower, B. *Function of Quinones in Energy Conserving Systems*; Academic Press: New York, 1982.
(28) Verkhovskaya, M.; Wikström, M. *Biochim. Biophys. Acta, Bioenerg.* **2014**, *1837*, 246−250.
(29) Ohnishi, T. *Biochim. Biophys. Acta, Bioenerg.* **1998**, *1364*, 186−206.
(30) Yano, T.; Dunham, W. R.; Ohnishi, T. *Biochemistry* **2005**, *44*, 1744−1754.
(31) Ingledew, W. J.; Poole, R. K. *Microbiol Rev.* **1984**, *48*, 222−271.
(32) Noguchi, T.; Inoue, Y.; Tang, X. S. *Biochemistry* **1999**, *38*, 399−403.
(33) Mouesca, J.-M.; Noodleman, L.; Case, D. A. *Int. J. Quantum Chem.* **1995**, *56*, 95−102.
(34) Brandt, U. *Annu. Rev. Biochem.* **2006**, *75*, 69−92.
(35) Verkhovsky, M.; Bloch, D. A.; Verkhovskaya, M. *Biochim. Biophys. Acta, Bioenerg.* **2012**, *1817*, 1550−1556.
(36) Zwicker, K.; Galkin, A.; Drose, S.; Grgic, L.; Kerscher, S.; Brandt, U. *J. Biol. Chem.* **2006**, *281*, 23013−7.
(37) Hanley, J.; Deligiannakis, Y.; MacMillan, F.; Bottin, H.; Rutherford, A. W. *Biochemistry* **1997**, *36*, 11543−11549.
(38) Cohen, A. J.; Mori-Sánchez, P.; Yang, W. *Science* **2008**, *321*, 792−794.
(39) Moser, C. C.; Keske, J. M.; Warncke, K.; Farid, R. S.; Dutton, P. L. *Nature* **1992**, *355*, 796−802.
(40) Page, C. C.; Moser, C. C.; Chen, X.; Dutton, P. L. *Nature* **1999**, *402*, 47−52.
(41) Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. *Phys. Rev. Lett.* **2003**, *91*, 146401.
(42) Yanai, T.; Tew, D. P.; Handy, N. C. *Chem. Phys. Lett.* **2004**, *393*, 51−57.
(43) Zwicker, K.; Galkin, A.; Dröse, S.; Grgic, L.; Kerscher, S.; Brandt, U. *J. Biol. Chem.* **2006**, *281*, 23013−7.
(44) Grgic, L.; Zwicker, K.; Kashani-Poor, N.; Kerscher, S.; Brandt, U. *J. Biol. Chem.* **2004**, *279*, 21193.
(45) Kashani-Poor, N.; Zwicker, K.; Kerscher, S.; Brandt, U. *J. Biol. Chem.* **2001**, *276*, 24082−24087.
(46) Baradaran, R.; Berrisford, J. M.; Minhas, G. S.; Sazanov, L. A. *Nature* **2013**, *494*, 443−448.
(47) Kieseritzky, G.; Knapp, E. W. *J. Comput. Chem.* **2008**, *29*, 2575−2581.
(48) Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; McCammon, J. A. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98*, 10037−10041.
(49) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26*, 1781−1802.
(50) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586−3616.
(51) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648−5652.
(52) Lee, C. T.; Yang, W. T.; Parr, R. G. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1988**, *37*, 785−789.
(53) Schäfer, A.; Horn, H.; Ahlrichs, R. *J. Chem. Phys.* **1992**, *97*, 2571−2577.
(54) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. *J. Chem. Phys.* **2010**, *132*, 154104.
(55) Weigend, F.; Ahlrichs, R. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297−3305.

(56) Klamt, A.; Schüürmann, G. *J. Chem. Soc., Perkin Trans. 2* **1993**, *5*, 799−805.

(57) Riahi, S.; Rowley, C. N. *J. Comput. Chem.* **2014**, *35*, 2076−86.

(58) Ahlrichs, R.; Bär, M.; Häser, M.; Horn, H.; Kölmel, C. *Chem. Phys. Lett.* **1989**, *162*, 165−169.

(59) Valiev, M.; Bylaska, E. J.; Govind, N.; Kowalski, K.; Straatsma, T. P.; van Dam, H. J. J.; Wang, D.; Nieplocha, J.; Apra, E.; Windus, T. L.; de Jong, W. A. *Comput. Phys. Commun.* **2010**, *181*, 1477−1489.

(60) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 33−38.

(61) Holst, M.; Saied, F. *J. Comput. Chem.* **1993**, *14*, 105−113.

(62) Rabenstein, B.; Ullmann, G. M.; Knapp, E. W. *Eur. Biophys. J.* **1998**, *27*, 626−637.

(63) Ishikita, H.; Knapp, E. W. *J. Am. Chem. Soc.* **2004**, *126*, 8059−8064.

(64) Zhu, Z.; Gunner, M. R. *Biochemistry* **2005**, *44*, 82−96.

(65) Schmidt am Busch, M.; Knapp, E. W. *J. Am. Chem. Soc.* **2005**, *127*, 15730−15737.

(66) Galstyan, A. S.; Knapp, E. W. *J. Comput. Chem.* **2009**, *30*, 203−211.

(67) Gamiz-Hernandez, A. P.; Galstyan, A. S.; Knapp, E. W. *J. Chem. Theory Comput.* **2009**, *5*, 2898−2908.

ARTICLE

Check for updates

Structure of inhibitor-bound mammalian complex I

Hannah R. Bridges [1,6], Justin G. Fedor [1,6], James N. Blaza [1,5,6], Andrea Di Luca [2,3,6], Alexander Jussupow [2], Owen D. Jarman[1], John J. Wright[1,4], Ahmed-Noor A. Agip[1], Ana P. Gamiz-Hernandez [2,3], Maxie M. Roessler [4], Ville R. I. Kaila [2,3✉] & Judy Hirst [1✉]

Respiratory complex I (NADH:ubiquinone oxidoreductase) captures the free energy from oxidising NADH and reducing ubiquinone to drive protons across the mitochondrial inner membrane and power oxidative phosphorylation. Recent cryo-EM analyses have produced near-complete models of the mammalian complex, but leave the molecular principles of its long-range energy coupling mechanism open to debate. Here, we describe the 3.0-Å resolution cryo-EM structure of complex I from mouse heart mitochondria with a substrate-like inhibitor, piericidin A, bound in the ubiquinone-binding active site. We combine our structural analyses with both functional and computational studies to demonstrate competitive inhibitor binding poses and provide evidence that two inhibitor molecules bind end-to-end in the long substrate binding channel. Our findings reveal information about the mechanisms of inhibition and substrate reduction that are central for understanding the principles of energy transduction in mammalian complex I.

[1] The Medical Research Council Mitochondrial Biology Unit, University of Cambridge, The Keith Peters Building, Cambridge Biomedical Campus, Hills Road, Cambridge CB2 0XY, UK. [2] Center for Integrated Protein Science Munich (CIPSM) at Department of Chemistry, Technische Universität München, 85748 Garching, Germany. [3] Department of Biochemistry and Biophysics, The Arrhenius Laboratories for Natural Sciences, Stockholm University, 106 91 Stockholm, Sweden. [4] Department of Chemistry, Molecular Sciences Research Hub, Imperial College London, White City Campus, London W12 0BZ, UK. [5] Present address: Department of Chemistry, University of York, Heslington, York YO10 5DD, UK. [6] These authors contributed equally: Hannah R. Bridges, Justin G. Fedor, James N. Blaza, Andrea Di Luca. ✉email: ville.kaila@dbb.su.se; jh@mrc-mbu.cam.ac.uk

Mitochondrial complex I (NADH:ubiquinone oxidor-eductase)[1] is central to oxidative phosphorylation in mammalian cells. It captures the free energy released from reduction of ubiquinone by NADH to pump protons across the inner mitochondrial membrane, support the proton-motive force and power ATP synthesis. As an essential respiratory enzyme, and an important contributor to cellular oxidative stress, complex I dysfunctions resulting from mutations in its subunits and assembly factors cause a diverse set of inherited neuromus-cular and metabolic diseases[2].

Mammalian complex I comprises 45 subunits: fourteen "core" subunits, which are conserved from bacteria to humans and sufficient for catalysis, and 31 "supernumerary" subunits, which are required for assembly, stability, regulation, or fulfil indepen-dent metabolic roles[1,3]. Understanding of the structure of mammalian complex I has leapt forward in recent years due to the advent of high-resolution single particle electron cryo-microscopy (cryoEM). The structure of complex I from mouse heart mitochondria was described recently at 3.3 Å resolution[4] and that of the yeast *Yarrowia lipolytica* at 3.2 Å[5]. The structures illustrate how, as shown previously in the enzyme from *Thermus thermophilus*[6], electrons enter from NADH oxidation, at the top of the hydrophilic domain, and are transferred towards the membrane along a chain of iron-sulphur (FeS) clusters. The terminal cluster, a [4Fe-4S] cluster known as N2, then donates the electrons to ubiquinone-10 (Q_{10}). The hydrophobic Q_{10} enters the enzyme from the mitochondrial inner membrane through a long binding channel, elevating its redox active head-group out of the membrane plane to within electron-transfer distance of N2. Although ubiquinone binding, reduction, and dissociation are now beginning to be defined by structural and functional data[5,7,8], the mechanism that couples the redox reac-tion to proton translocation remains poorly understood.

The quinone binding tunnel in complex I is long and hetero-geneous; the top and bottom sections are hydrophobic, while the central section is surrounded by many charged residues[6,8]. The charged region may be important in linking redox catalysis to proton translocation, because it sits at the start of a chain of charged residues that leads into the membrane plane[4,6,9], and involves the loops from membrane-bound subunits ND1 and ND3 that may move during catalysis[10]. The structures of the mammalian enzyme in both its active and deactive states depict how these loops become disordered when complex I converts from its ready-to-catalyse "active" state to its "deactive" state. The deactive state is a pronounced resting state that forms sponta-neously at physiological temperatures in the absence of turnover, and requires reactivation by both NADH and ubiquinone in order to re-enter the catalytic cycle[3,4,11].

The large, heterogeneous and conformationally-labile nature of the ubiquinone binding site may explain, why such a wide variety of compounds with little resemblance to the substrate inhibit complex I[12–14]. However, only limited structural information is available on their bound states. Density for piericidin was shown pictorially in the structure of *Thermus thermophilus* complex I[6], but neither the model nor data were made available, precluding evaluation of the information. The inhibitor 2-decyl-4-quinazolinyl amine has been observed with its headgroup part way up the ubiquinone binding channel of complex I from *Y. lipolytica*[10]. For several families of inhibitor, including pier-icidin[12,15] extensive structure-function studies have been under-taken, but the information is difficult to interpret without detailed knowledge of the inhibitor-binding mode. No structures of inhibitors bound to the mammalian enzyme have been presented, despite their inherent biomedical interest[16,17], and that inhibitor-bound structures present unrivalled opportunities to access dif-ferent mechanistically-relevant enzyme states.

Piericidin is a natural insecticide that was first isolated in the 1960s from the spore forming bacterium *Streptomyces mobar-aensis*[18], and has now been synthesised chemically[19]. It is a tight-binding complex I inhibitor[20] that resembles a short chain ubiquinone (Q_3). The headgroup resembles the ubiquinone headgroup (see Fig. 1), except one of the two carbonyls (reduced to hydroxyls in ubiquinol) is replaced by a 4-pyridone nitrogen. The hydrophobic tail contains an initial isoprenoid followed by two isoprenoid-like groups, with a hydroxyl group close to its end. Piericidin has been described to compete for the same or overlapping binding sites as the inhibitors rotenone and DQA[21], and to display partially-competitive inhibition with the substrate analogue Q_2[22] but mixed behaviour towards decylubiquinone

a The piericidin molecule is located at the interface between the redox (blue) and proton-transfer (green) domains; supernumerary subunits in grey. Charged residues in the proton-transfer domain are marked in purple. M matrix, IMM inner mitochondrial membrane, IMS intermembrane space.
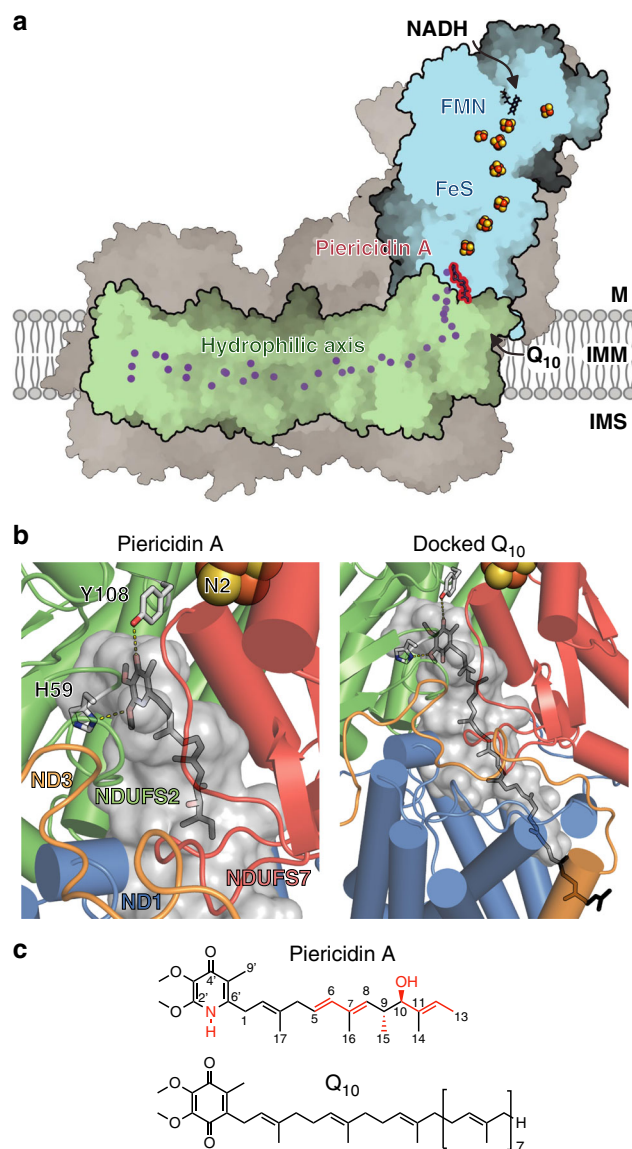
Fig. 1 Piericidin bound in the structure of mammalian complex I. a The piericidin molecule is located at the interface between the redox (blue) and proton-transfer (green) domains; supernumerary subunits in grey. Charged residues in the proton-transfer domain are marked in purple. M matrix, IMM inner mitochondrial membrane, IMS intermembrane space. **b** Piericidin (left) and a docked-in Q_{10} (right) in the proposed ubiquinone-binding cavity, leading from Tyr108 and His59 to outside the protein. The internal surface of the cavity was identified using PyMol-1.8.4.0. NDUFS7 helix 4 (residues 104–117) has been removed for clarity. **c** The structures of piericidin and Q_{10} with differences highlighted in red.

(DQ)[23]. It inhibits both the forward NADH:ubiquinone oxidoreductase reaction, and reverse electron transfer (succinate-driven reduction of NAD[+] in submitochondrial particles) with equivalent efficacy[24]. ^{14}C-labelled piericidin studies describe two noncooperative binding sites for piericidin for complex I in its native lipid environment, but only one in the delipidated enzyme[24,25].

Here, we have used the canonical complex I inhibitor piericidin A1 (referred to hereon as piericidin for simplicity) to determine the structure of complex I with a ubiquinone-analogue inhibitor bound in its active site. We combine structural, kinetic, spectroscopic, and computational analyses to elucidate the inhibitor-binding mode and the relationships between piericidin and ubiquinone binding, and find evidence for a distal ubiquinone/inhibitor binding site in the central region of the quinone-binding channel. Our piericidin-bound structure is the highest resolution structure of mammalian complex I in an active-like state so far reported, and it reveals information about the mechanisms of inhibition and ubiquinone reduction.

Results

Determination of the structure of piericidin-bound complex I. Piericidin-inhibited complex I was prepared by adding piericidin to the enzyme during turnover with NADH and DQ to ensure exposure of the inhibitor-binding state. Residual substrates and inhibitor were then removed to ensure any inhibitor present was specifically bound. The sample of mouse complex I analysed by

cryoEM was 89 ± 3% inhibited by comparison with a control sample prepared identically but without inhibitor. The complex was frozen onto PEGylated gold grids and two datasets (Supplementary Table 1) with similar pixel sizes, and numbers of micrographs were collected on FEI Titan Krios microscopes, using either a Gatan Quantum K2 Summit detector with energy filter (piericidin-K2) or an FEI Falcon III detector in counting mode (piericidin-FIII). Both datasets were processed by RELION-3.0 to 3.0 Å resolution (Supplementary Figs. 1, 2, 3 and Supplementary Table 1). Once sharpened, the two maps are essentially indistinguishable: the piericidin-K2 map was used unless otherwise stated. The dataset for the active state of mouse complex I (referred to as active complex I) described previously[4], which reached 3.3 Å resolution with RELION-2.1, was then reprocessed with RELION-3.0 to 3.1 Å (Supplementary Figs. 2, 3 and Supplementary Table 1). The resolutions achieved enabled confident modelling of 97% of the 8430 residues of both the piericidin-bound and active enzymes (Supplementary Table 2).

Piericidin binding at the top of the ubiquinone-binding channel. Density for a bound piericidin was readily identified at the top of the ubiquinone-binding cavity, where the ubiquinone ring and first three isoprenoids of ubiquinone-10 are expected to bind (Figs. 1 and 2a, b). Overall, the piericidin-bound maps and model match those for active complex I very closely (98% correlation between the piericidin-bound and active maps
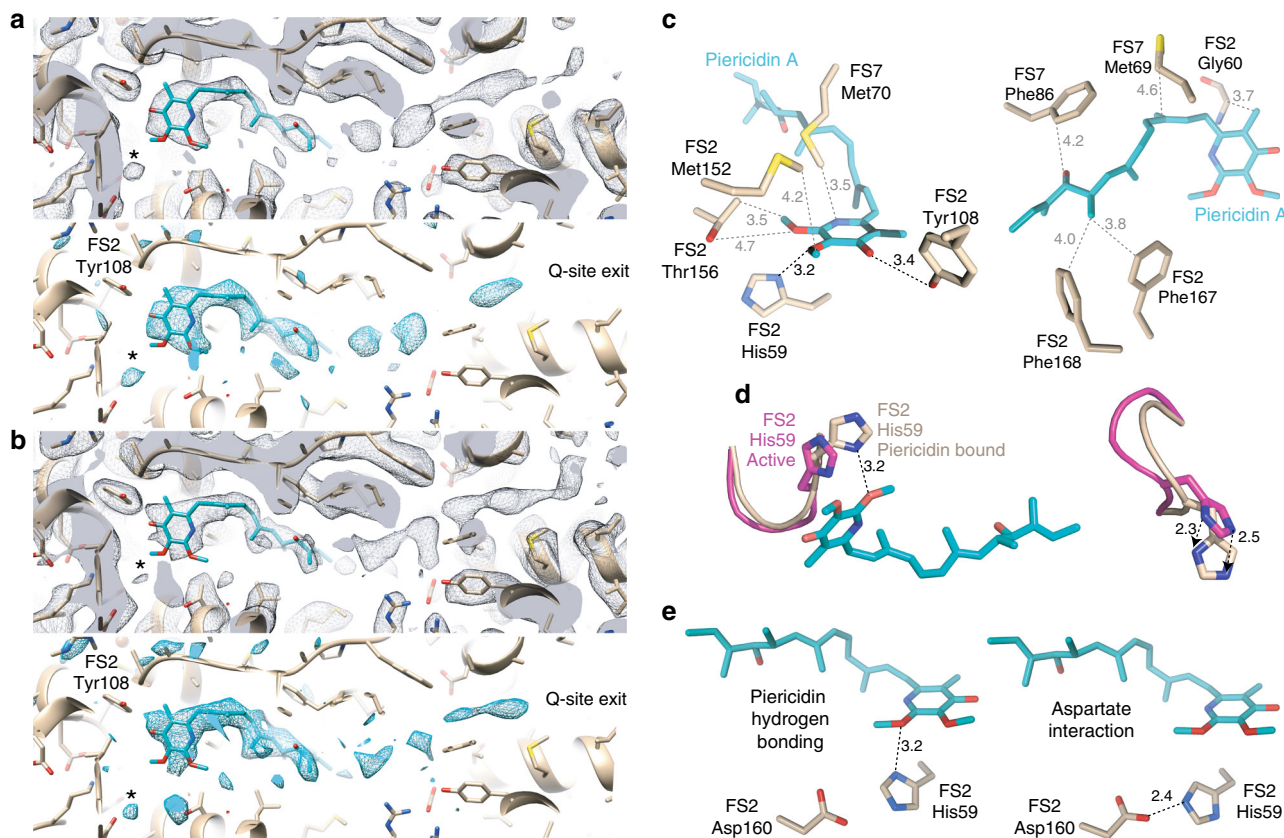

Fig. 2 Cryo-EM densities for the piericidin molecule, and interactions between the protein and the bound piericidin. a, b Cryo-EM densities for the piericidin-containing region (black mesh, top) and the difference between the piericidin-free protein model and the piericidin-bound density (teal mesh, calculating using the phenix.real_space_diff_map command) for **a** the piericidin-K2 and **b** the piericidin-FIII datasets. A putative water molecule can be observed in the lower left of each panel (marked by an asterisk). Densities rendered in UCSF Chimera at thresholds of **a** 0.0163, 6.07 and **b** 0.0294 and 5.82. **c** Residues around the piericidin; proposed hydrogen bonds marked in black, other distances in grey. **d** Movement of the NDUFS2 β1-β2 loop between the active (magenta) and piericidin-bound (wheat) models. **e** Interactions of NDUFS2-His59 with the piericidin headgroup and Asp160, depending on the rotation of the His ring; refined structure with interaction with piericidin 2′ methoxy, and ring flipped to form interaction with Asp160. All distances in Å. Subunits NDUFS2 and NDUFS7 are abbreviated as FS2 and FS7, respectively.

in UCSF Chimera and all atom RMSD = 0.6 Å, relative to 2.7 Å for the deactive state (PDB:6G72)) and exhibit all the hallmarks of it: the extended interface between NDUFA5 and NDUFA10 that defines the arrangement of the hydrophilic and membrane domains[4]; clear densities for the mobile loops in NDUFS2, ND3, and ND1[3,4,11]; and α-helical structure throughout ND6-TMH3[4]. Two short stretches of NDUFS7 (residues 58–60 and 86–92, where R91 is known to be hydroxylated)[26] also match the active state. Therefore, we observe piericidin bound to the active state of complex I, in which the ubiquinone-binding site is fully config-ured and ready for catalysis.

The piericidin headgroup binds adjacent to NDUFS2 His59 and Tyr108, the two proposed ligands of the ubiquinone headgroup (Fig. 2c). Its ring carbonyl accepts a hydrogen bond from the Tyr108 hydroxyl (O–O distance 3.4 Å), consistent with the keto form of the piericidin headgroup (rather than the tautomeric enol form with an aromatic ring, a deprotonated pyridone-N and a hydroxyl). Because piericidin lacks a second carbonyl it cannot form a second ubiquinone-like hydrogen bond to His59. Instead, His59 Nδ1 forms a hydrogen bond (N–O distance 3.2 Å) with the piericidin 2′ methoxy (Fig. 2c). Alternatively, a good fit to the density can be obtained from flipping the imidazole ring to form a hydrogen bond (albeit with poor geometry) between His59 Nε2 and Asp160 (N–O distance 2.4 Å) (Fig. 2e). To accommodate the piericidin, His59 has shifted by ~2.5 Å (Fig. 2d); this is the only difference that could be identified between the piericidin-bound and active structures.

Additional residues around the piericidin may further stabilise it. In particular, NDUFS7-Met70 and NDUFS2-Met152 point towards one face of the ring (Fig. 2c) and the sidechain of NDUFS2-Thr156 is modelled 3.5 Å from the 2′ piericidin methoxy, and could be rotated to bring its hydroxyl group into a hydrogen-bonding configuration (but compromising the fit to the density). Interestingly, weak densities observed between the

Thr156 hydroxy, the piericidin methoxy groups, and NDUFS2 Lys371 may arise from bound water molecules and reflect a wider hydrogen-bonded network for ubiquinone protonation/reduction (see below). However, we are not sufficiently confident to model water networks at the current resolution. The piericidin isoprenoid-like tail tracks along the proposed ubiquinone-binding channel, overlapping the predicted positions of iso-prenoids one to three (Figs. 1, 2) and surrounded by a series of hydrophobic sidechains (Fig. 2c), including NDUFS7 Phe86 in a π–π interaction and NDUFS2-Phe167 and Phe168 framing the final isoprenoid-like unit. The sidechains of NDUFS7-Thr59 and ND1-Glu204 are near to the hydroxy group on the piericidin chain, but too far to form hydrogen bonds. Glu204 is on the TMH5-6 loop of ND1 at the start of the "E-channel" that connects the ubiquinone-binding site to charged residues in the membrane domain[4,6].

Molecular dynamics (MD) of piericidin in the substrate binding pocket.

To probe the piericidin-binding mode further, we performed classical MD simulations, starting from the piericidin-bound cryoEM model. The piericidin was modelled into the density, or docked in using MD flexible fitting (MDFF)[27] with the density as a biasing potential. During the MD simula-tions (Fig. 3 and Supplementary Fig. 4), the NDUFS2 Tyr108-OH forms a stable hydrogen bond with the piericidin 4′ carbonyl, consistent with the cryoEM model, but at a slightly shorter dis-tance of ~2.9 Å (Fig. 3c). NDUFS2 His59, modelled as the doubly-protonated imidazolium (HisH+) samples a hydrogen-bonded conformation with the piericidin 2′ methoxy, but prefers an alternative conformation, resembling that obtained by flipping the His59 ring as in Fig. 2e (right) and stabilised by van der Waals interactions. The histidine further forms an ion-pair interaction with Asp160 (Fig. 3c). Interestingly, simulations with a neutral histidine (Nδ-protonated only) resulted in His59 and the
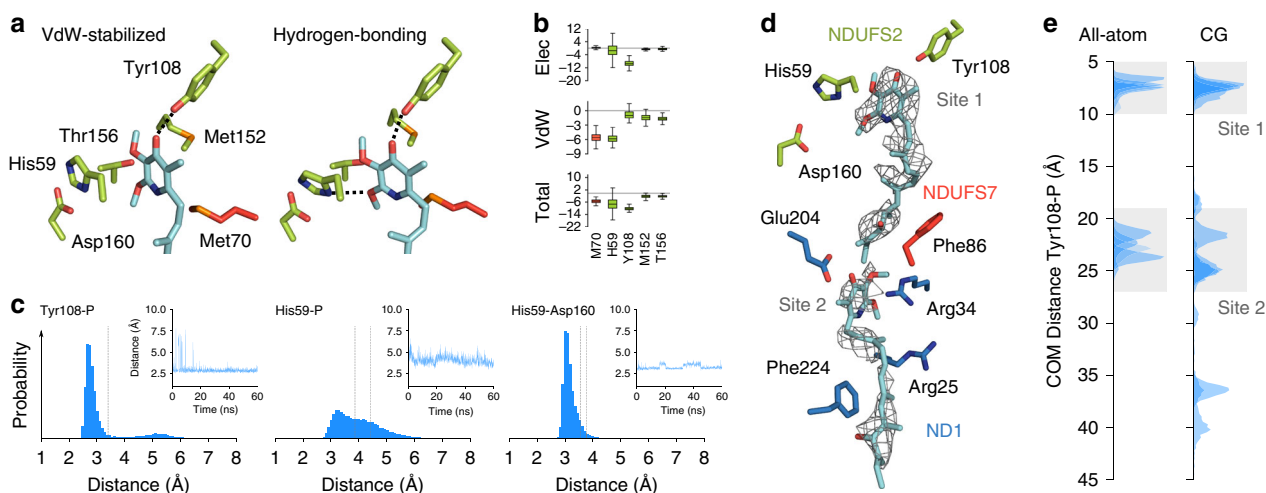

Fig. 3 Molecular dynamics simulations of piericidin in complex I. a Snapshots from the MD simulations of piericidin in van der Waals (VdW)-stabilised and hydrogen-bonded conformations. Hydrogen bonds are indicated by dotted lines. **b** Interaction energies between piericidin and selected surrounding residues (total, van der Waals, and electrostatic contributions in kcal mol⁻¹). Green, residues in NDUFS2; red, residues in NDUFS7. The boxes extend from the 25th to the 75th percentile, the middle line represents the median. The whiskers show the range of the data from the 10th to 90th percentile ($n = 1500$, snapshots calculated every 40 ps). **c** Distance distributions (simulations A1–A4, Supplementary Table 4) and time traces of distances (simulation A1) between piericidin (P) and key residues. Left: piericidin 4′ carbonyl oxygen to Tyr108 hydroxyl oxygen. Middle: piericidin 2′ methoxy oxygen to His59 (centre of mass of Nδ/Cε/Nε). Right: His59 (centre-of-mass of Nδ/Cε/Nε) to Asp160 (centre-of-mass of carboxylic group). Experimental distances (Fig. 2c) for the refined structure (dark grey) and ring-flipped model (light grey) are shown as dotted lines. See Supplementary Fig. 4 for further time traces from individual simulations, and distances to Thr156. **d** The two modelled piericidin molecules with the experimental cryo-EM density (PyMol-2.2.3 density level 4.0, carve radius = 1.8) and surrounding residues. **e** Distance distributions of piericidin in the substrate cavity obtained from classical (left, simulation A1–A5, 220 ns total) and coarse-grained MD simulations (right, simulation C1–C10, 100 μs total). Distances are between the centre of masses of the Tyr108 and piericidin rings.

NDUFS2 β1–β2 loop moving away from the piericidin and Asp160, suggesting histidine protonation is required for stable piericidin binding. Hybrid quantum/classical (QM/MM) simulations, in which the binding site was modelled using density functional theory (DFT) polarised by a classical model for the surrounding protein, also revealed the hydrogen-bonded and van der Waals binding poses. In addition, at the expense of atomistic detail but with enhanced statistical sampling, coarse-grained (CG) MD simulations support the piericidin remaining stable in the binding pocket for several tens of microseconds. Finally, energy decomposition analyses based on the MD simulations (Fig. 3b and Supplementary Fig. 5) suggest that His59, Tyr108, and Asp160, as well as NDUFS2 Val424 and Phe425, which move closer to the piericidin ring in some of the simulations, make strong contributions to the interaction energy. Although NDUFS2 Thr156 samples transient hydrogen-bonded interactions with the 2′ methoxy (Fig. 3a and Supplementary Fig. 4d) it contributes only moderately to the interaction energy, along with NDUFS7 Met70 and NDUFS2 Met152 (Supplementary Fig. 5). Commensurate with the additional densities observed in the cryoEM analyses, water clustering analyses (see Supplementary Fig. 5b and see "Methods" section) suggest that this region could form a stable water binding site, stabilised by the nearby conserved residues NDUFS2 Asp107 and Asp422.

The environment of cluster N2. During the inhibitor-binding incubation, complex I was exposed to NADH to induce turnover and ensure exposure of the inhibitor-binding site. Inhibitor binding prevents reoxidation by DQ, so to determine if the enzyme was prepared in the reduced state, the same protocol was used to prepare a larger amount of the piericidin-bound bovine complex (inhibited by $72 \pm 6\%$) for EPR analyses. With no extra NADH added to reduce the sample, the signal from reduced FeS cluster N2 was clearly observed in the inhibited sample (Fig. 4a). It was also observed in the turnover control (that was prepared identically but not exposed to piericidin), but not in the matching "nonturnover" control to which NADH had not been added (Fig. 4b). No signals from any other reduced clusters were observed. The reduction of only N2 is consistent with the relatively high reduction potential of this cluster in the mammalian complex[28,29], disfavouring electrons transferring from N2 back to the low-potential flavin where they may slowly escape to O_2[30]. The samples were subsequently thawed and NADH added to reduce N2 completely, showing that N2 was 78% reduced in the inhibited sample, 44% reduced in the turnover control, and fully oxidised in the nonturnover control. Therefore, we infer that cluster N2 was predominantly reduced in the piericidin-bound sample used for cryoEM, although the small scale of the cryoEM preparation precludes direct measurement. Strikingly, Fig. 4c shows that the environments of N2 in the piericidin-bound (N2-reduced) and active (N2-oxidised) states of mouse complex I are indistinguishable. Even NDUFS2-His190, which becomes protonated at pH < 7.4 when N2 is reduced[29,31] and has a well-defined density in both maps, does not move. Furthermore, the lack of any substantial structural change "downstream" of N2 in the proton pumping modules is consistent with our previous conclusion[29] that N2 redox cycling is not the coupling reaction that initiates proton translocation.

Kinetic evidence for the binding of two inhibitor molecules. Inhibition of complex I by piericidin was studied in proteoliposomes reconstituted with complex I, Q_{10}, and an excess of the cyanide-insensitive alternative oxidase (AOX) to recycle the ubiquinol back to ubiquinone[32,33]. Proteoliposomes were prepared with varying ubiquinone concentrations to span the K_M
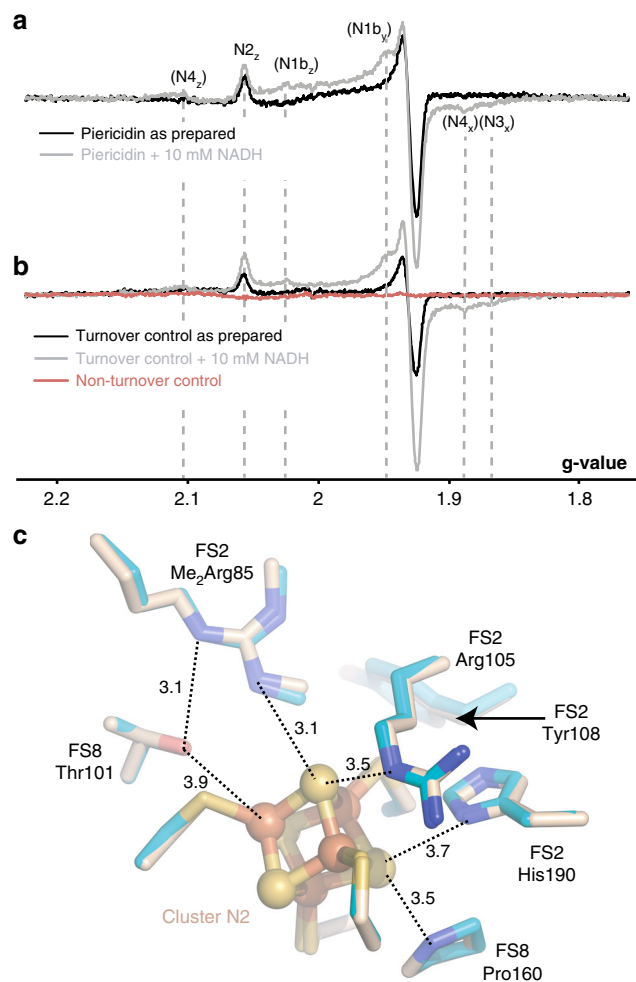
Fig. 4 Cluster N2 is reduced in the piericidin-bound enzyme but its surrounding structure does not change. a EPR spectra showing reduced clusters in the piericidin-treated sample as prepared (black) and following addition of 10 mM NADH (grey). **b** EPR spectra showing reduced clusters in the piericidin-free control samples: the turnover control as prepared (black) and following addition of 10 mM NADH (grey) and the nonturnover control as prepared (red). EPR conditions: 9.5 GHz (X-band), temperature 20 K, microwave power 2.0 mW, modulation frequency 100 kHz, modulation amplitude 0.7 mT. **c** Key residues around N2 in the piericidin-bound (teal) and active (wheat) states. Distances are in Å. NDUFS2-Tyr108, the ligand to the piericidin headgroup, is shown for reference.

curve, and characterised by their phospholipid content (hydrophobic phase volume), ubiquinone concentration, and complex I concentration and orientation. Complex I inhibition was then investigated by titrating the rate of NADH oxidation with piericidin. Figure 5 shows that the data can be modelled by a series of Michaelis–Menten curves for their Q_{10} dependence, and a series of IC_{50} curves for their piericidin dependence. Interestingly, both the apparent K_M and V_{max} values decrease with increasing piericidin concentration, contrary to the behaviour expected for a competitive inhibitor (increasing K_M and constant V_{max})—although the piericidin is clearly observed in the binding site for the ubiquinone headgroup in the cryoEM map. Different models were therefore tested as follows. The complex I, ubiquinone-10, and piericidin concentrations were defined relative to the membrane volume, and the ratio of membrane to aqueous piericidin were set by its partition constant ($\log_{10}P = 4.73$). Reoxidation of ubiquinol-10 was assumed to be fast so inhibition by ubiquinol-
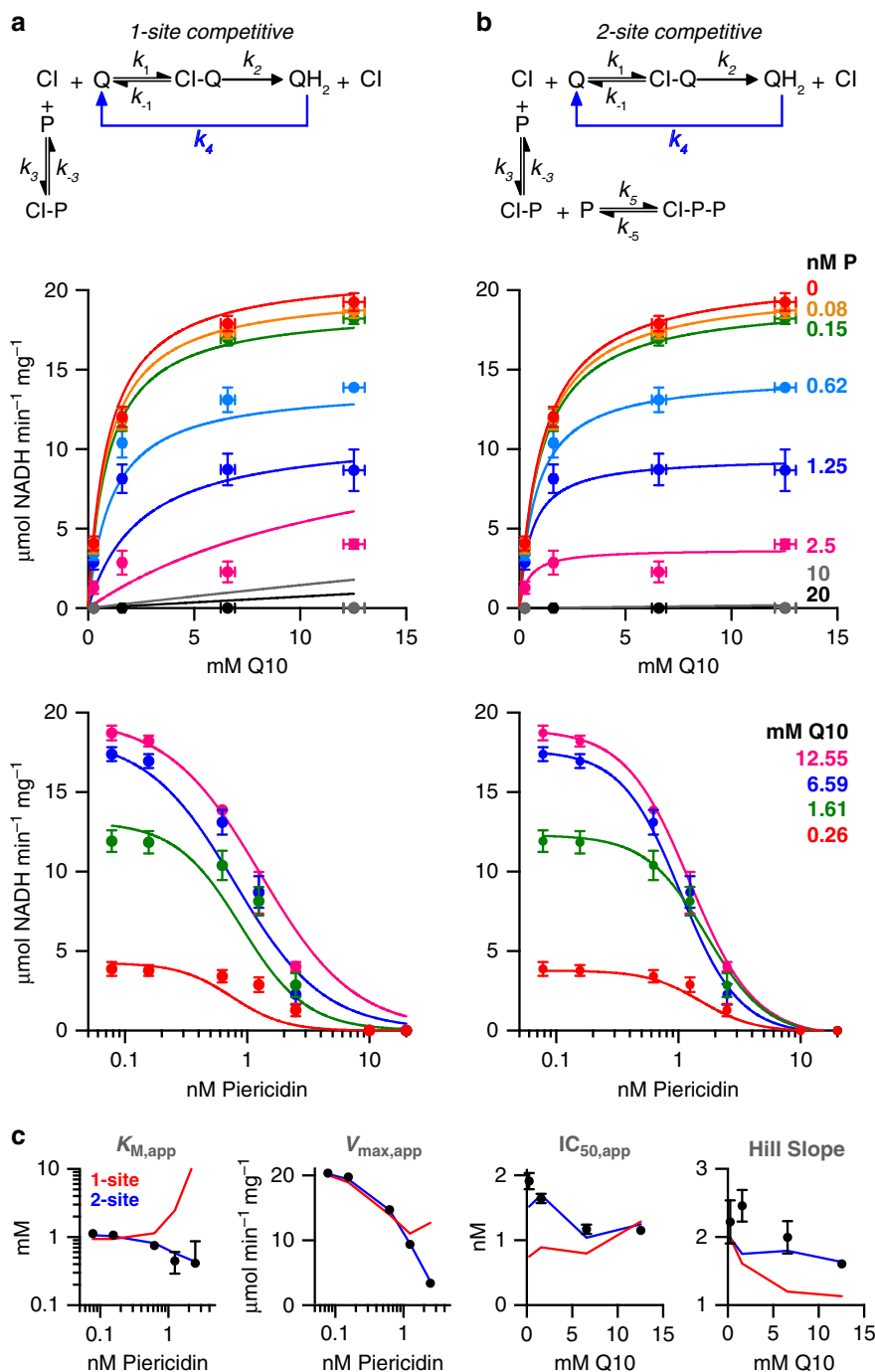
Fig. 5 Competitive models for piericidin inhibition of complex I in proteoliposomes. a One-site competitive inhibition mechanism. **b** Two-site competitive inhibition mechanism. In both cases, rapid reoxidation of ubiquinol by AOX (k_4) prevents appreciable levels accumulating. Experimentally-measured rates (average ± standard deviation, $n = 3$) are shown in K_M and IC_{50} plots, alongside the best-fit predictions from the models (see Supplementary Table 3 for parameters). **c** Plots of trends in $K_{M,app}$, $V_{max,app}$, IC_{50} and Hill slope produced by fitting K_M and IC_{50} curves to the experimental data (black points, error bars are 95% confidence intervals) and to the output data from the models. The SSR values are 201 and 46.7 for **a** and **b**, respectively.

10 was not considered. Importantly, piericidin is a tight-binding inhibitor, and the pool of ubiquinone/ubiquinol is small, so depletion of their free concentrations upon binding were considered. Although this precluded simple analytical solution of the rate equations and required kinetic rates to be calculated numerically, the numerical approach also enabled us to test essentially any reasonable model without having to make simplifying assumptions, such as about the identity of the rate determining step in complex I catalysis.

Figure 5 confirms that the simplest competitive inhibition model does not account for the data. The sum of the squared-residuals (SSR) for the overall fit (taking into account every data point) is high (Supplementary Table 3) and the observed trends in the apparent values for K_M, V_{max}, IC_{50}, and Hill slope are not replicated. However, by introducing a second binding site to the model, the fit improves substantially: the SSR decreases 7-fold and trends in K_M, V_{max}, and IC_{50} are predicted well. Early work on the inhibition of complex I by piericidin indicated a

stoichiometry of 2[24], and the Hill slopes of the IC_{50} curves determined here are around 2, suggestive of cooperative binding. Intriguingly, and consistent with the cooperativity, the best-fit K_I values (where K_I is the dissociation constant for the inhibitor) suggest the second inhibitor binds substantially more tightly than the first, with a broad minimum in SSR at $K_{I2}/K_{I1} = 10^{-5}$ to 10^{-1}, and the minimum point at 10^{-4}, with a 10,000× tighter binding of the second piericidin than the first (see below). Further analysis (Supplementary Table 3) showed that both a single inhibitor binding competitively to the ubiquinol-binding state, and a simple mixed (both competitive and uncompetitive) binding scheme also resulted in poor fits. A two-site uncompetitive binding model, where piericidin binds to the enzyme-substrate complex, was also able to explain the data satisfactorily, but this kinetic model is not supported by the structural data. Therefore, the kinetic data strongly support the two-site competitive inhibition model, implying that more than one piericidin enters the ubiquinone-binding channel.

Evaluation of densities in the distal section of the ubiquinone-binding channel. Broken, noisy densities are observed in the "distal" section of the ubiquinone-binding channel in both the piericidin-K2 and piericidin-FIII maps (Fig. 2a, b). However, they are insufficiently resolved for identification—and similar densities are also observed in the active mouse map[4]. To probe whether an additional piericidin would fit the extra densities in the piericidin-K2 map, we performed MD simulations with a second piericidin added. The fitted distal molecule adopts positions that adequately explain the density features observed (albeit with insufficient confidence for inclusion in the cryoEM model), and a hydrogen-bonding interaction was observed to form intermittently between the C10-hydroxyl on the proximal molecule and the 4′ carbonyl on the distal molecule (Fig. 3d). As expected for its location at the position of isoprenoid 4, the headgroup of the distal molecule is also surrounded by a number of charged residues[8]. The charged region, followed by a more hydrophobic region toward the channel entrance, may provide a favourable environment for a range of amphipathic species, and it also corresponds (Fig. 3e) to a recently proposed ubiquinone-binding subsite (site 2) identified in simulations on bacterial complex I[34]. However, the distal piericidin does not adopt a single, well-defined binding pose, and its fluctuations in the MD simulations, together with the unclear density, are better explained by a broad substrate recognition region, rather than a unique, well-defined distal site. The broad recognition region may lead to an entropic stabilisation for the set of subsites to provide a higher affinity for the site as a whole, whereas the interaction between the two piericidins may contribute to the cooperativity, explaining the tighter apparent binding of the second piericidin relative to the first.

Notably, early work to determine the number of piericidins bound to complex I showed there are two nonidentical specific binding sites per membrane-bound enzyme, but that this decreases to one in the delipidated enzyme[24,25]. Loss of high affinity piericidin binding in the detergent-solubilised enzyme, relative to in the membrane, has also been described[22]. Our data are consistent with two piericidins binding in our membrane-bound enzyme assay, and with lower occupancy of the distal site in the detergent-solubilised, structurally-characterised enzyme. Specific phospholipids are known to be essential for complex I activity[35] and their loss and/or the presence of detergent may destabilise the second piericidin, and affect the properties of the distal section of the ubiquinone-binding channel. Although our results thus present an example of how the structurally-characterised detergent-solubilized protein differs from the membrane-bound enzyme, our interpretation is consistent with

an important role for phospholipids, and cardiolipin in particular, in complex I catalysis[35,36].

Discussion

Our structures show that piericidin binds at the top of the ubiquinone-binding site between NDUFS2-Tyr108 and NDUFS2-His59, two highly conserved residues long known to be critical for catalysis[37,38]. Tyr108 hydrogen bonds with the piericidin 4′ carbonyl, and our MD simulations suggest it makes a substantial contribution to the binding energy (Fig. 3b and Supplementary Fig. 5). Intriguingly, in the complex I genes present in *S. mobaraensis*, from which piericidin was isolated, a tryptophan replaces Tyr108[39]. There are currently no structures available for ubiquinone bound to complex I, but simulations on the *T. thermophilus* enzyme[40] support an analogous hydrogen bond between Tyr108 and the ubiquinone 4′ carbonyl, poised to protonate the nascent quinol. In the yeast *Y. lipolytica*, all variants of the Tyr (Y144) rendered the enzyme essentially unable to reduce Q_9, while Q_1 and Q_2 supported near wild-type V_{max} values but with substantially increased K_M values[38]. IC_{50} values for the inhibitor 2-decyl-4-quinazolinyl amine (DQA), measured with Q_1, were similarly increased in Tyr mutants[37]—despite structural data showing DQA bound further along the binding channel[10].

Data on the role of His59 in ubiquinone binding and reduction is less conclusive. In piericidin, the 1′ carbonyl is replaced by a pyridone nitrogen hydrogen bond donor (N–H), so His59 must interact differently with piericidin than ubiquinone. Here, the His is observed to hydrogen bond to the 2′ piericidin methoxy. However, our MD simulations suggest piericidin samples two different binding poses, one stabilised by hydrogen bonding and one by van der Waals interactions. Confusingly, while the Ala, Met, or Arg variants of His59 (His95) in *Y. lipolytica* abolished catalysis with no deleterious effects on assembly[38], the analogous NuoD-His224 to Arg variant in *E. coli* gave an enzyme with near wild-type activity with all quinones tested, and near wild-type inhibitor-binding characteristics[41]. The exact interaction mode(s) of His59 with the ubiquinone headgroup and its role in catalysis thus remain unconfirmed.

Variants of NDUFS2-Thr156, NDUFS7-Met70, and NDUFS2-Met152, identified here as relevant to binding, have also been studied in *Y. lipolytica*. Rotenone and DQA inhibit more weakly in *Y. lipolytica* than in the mammalian enzyme[42]. Mutating *Y. lipolytica* Ser192, homologous to mouse Thr156, to Thr increased the affinities for both rotenone and DQA, whereas mutations to Ile, Arg, and Tyr were all detrimental to activity[7]. Variants of Met70 (*Yl* Met91) showed increased K_M values with Q_1 and Q_2, with V_{max} hardly affected, and resistance to DQA and rotenone[43]. Similarly, variants of Met152 (*Yl* Met188) exhibited varying amounts of activity with all quinones[42]. Finally, in *Rhodobacter capsulatus* residues on the C-terminal helix of NDUFS2, particularly Val424 (*Rc* Val407) were found to affect piericidin binding[44]. Val424 is close to the 3′ methoxy group on the piericidin headgroup, and is also identified in our energy decomposition analysis (Supplementary Fig. 5).

Our data demonstrate that piericidin competes with ubiquinone for its binding site, and that piericidin binds to an active-like state of mammalian complex I, with all elements of the ubiquinone-binding site defined in the density. However, strictly speaking, the structurally-characterised active state is an off-pathway state with oxidised FeS clusters, because during catalysis NADH oxidation outpaces ubiquinone reduction and cluster N2 is reduced. In contrast, our piericidin-bound structure contains a reduced cluster N2, which does not lead to observable structural changes. Charge delocalisation over the cluster core to minimise reorganisation and facilitate rapid electron transfer, is a feature of

3Fe-4S and 4Fe-4S cluster chemistry. For example, no substantial changes upon reduction were detected in high resolution structures of the 7Fe ferredoxin I from *Azotobacter vinelandii*[45]. In contrast, the NADH-bound, reduced structure of the hydrophilic domain of complex I from *T. thermophilus* was documented to show subtle movements in several helices at the hydrophilic/membrane domain interface[46]. Corresponding movements are not observed here suggesting they were not representative of the intact enzyme. Furthermore, our density shows no disconnection of either of the tandem cysteine residues that coordinate cluster N2 (Fig. 4b), as described for reduced N2 in the *T. thermophilus* hydrophilic arm, in which N2 is more highly solvent exposed[46].

Our data indicate that two piericidins can be accommodated in the ubiquinone binding channel in the membrane bound complex, with the distal molecule occupying a site that broadly resembles one of the additional binding sites for ubiquinone predicted by simulations on the structure of *T. thermophilus* complex I[34]. First, these sites may represent "staging posts" for the transit of quinone/quinol along the long channel, where the substrate pauses due to favourable interactions with its environment. This staging post concept may help to explain the relatively low V_{max} values of Q_1 and Q_2[8], since their small size renders them susceptible to becoming trapped at these local energy minima, and as multiple molecules may impede the progress of each other. High concentrations of Q_2, in particular, have been reported to inhibit catalysis[12]. The same principles[8] may help to explain why piericidin (which has an alkyl chain length similar to Q_3) is so tight binding. The distal site sits within a network of charged residues around the central region of the ubiquinone-binding channel, connected to the E-channel, a series of glutamate residues leading down into the membrane plane. It may therefore be functionally relevant in the (as-yet unknown) mechanism by which ubiquinone binding and reduction is coupled to proton translocation.

Methods

Preparation of mouse complex I inhibited by piericidin for cryoEM.
Piericidin-inhibited mouse complex I was prepared by adapting the method of Agip and coworkers for mouse complex I[4]. Six-week old C57BL/6 female mice were sacrificed by cervical dislocation in accordance with the UK Animals (Scientific Procedures) Act, 1986 (PPL: 70/7538, approved by the local ethics committees of the MRC Laboratory of Molecular Biology and the University of Cambridge and by the UK Home Office) and the University of Cambridge Animal Welfare Policy. Hearts were excised, immersed in ice-cold solution containing 10 mM Tris-HCl (pH 7.4 at 4 °C), 75 mM sucrose, 225 mM sorbitol, 1 mM EGTA and 0.1% (w/v) fatty acid-free bovine serum albumin (Sigma-Aldrich), then cut into 1-mm pieces and washed. The tissue was homogenised (10 mL buffer per gram tissue) by seven to ten strokes of a Potter–Elvehjem homogeniser fitted with a Teflon pestle at ~1000 r.p.m then the homogenate was centrifuged ($1000 \times g$, 5 min), and the supernatant recentrifuged ($9000 \times g$, 10 min) to collect the crude mitochondria. The mitochondria were washed several times by resuspension and centrifugation and then collected by centrifugation at $15,000 \times g$ for 2 min, resuspended in 20 mM Tris-HCl (pH 7.4 at 4 °C), 1 mM EDTA and 10% (v/v) glycerol to 10–20 mg protein mL^{-1} and frozen for storage. After thawing they were diluted to 5 mg protein mL^{-1}, then ruptured by three 5 s bursts of sonication (with 30 s intervals on ice) using a G700 Sonicator (Qsonica) at 65% amplitude setting and the membranes were collected by centrifugation ($75,000 \times g$, 1 h), resuspended to 5 mg protein mL^{-1} and frozen for storage.

Starting from 16 mg membrane protein, the membranes were solubilized by addition of 1% *n*-dodecyl β-D-maltoside (DDM, Glycon), stirred for 30 min on ice, and centrifuged ($48,000 \times g$, 30 min). The supernatant was loaded onto a 1 mL Hi-Trap Q HP anion exchange column (GE Healthcare) pre-equilibrated with buffer A (20 mM Tris-HCl (pH 7.14 at 20 °C), 1 mM EDTA, 0.1% DDM, 10% (v/v) ethylene glycol, 0.005% asolectin (Avanti) and 0.005% CHAPS (Calbiochem)). The column was washed at 0.3 mL min^{-1} with buffer A for 2 min, then with 20% buffer B (buffer A + 1 M NaCl) for 7 min, and then complex I was eluted in 35% buffer B for 10 min[4]. Complex I-containing fractions were pooled and concentrated to 100 μL using a 50 kDa MWCO spin column (Sartorius) and the following were then added: 200 μM NADH, 6 μM decylubiquinone (DQ), 0.15% asolectin (20% soy PC, Avanti) and 0.15% CHAPS (Calbiochem) to drive complex I catalysis; 150 μg mL^{-1} alcohol dehydrogenase from *Saccharomyces cerevisiae* (Sigma) and 1% ethanol to regenerate the NADH from NAD$^+$; 100 μg mL^{-1} alternative oxidase from

Trypanosoma brucei brucei (AOX[8]) to regenerate the ubiquinone from ubiquinol; 10 KU mL^{-1} catalase from *Corynebacterium glutamicum* (Sigma) and 400 U mL^{-1} superoxide dismutase from bovine erythrocytes (Sigma) to minimise oxidative damage. After ~5 min at 4 °C, 15 μL of the mixture was removed as a control, and the remaining sample added to a glass vial containing sufficient piericidin A (dried from an ethanolic stock solution to avoid addition of solvent) to give 200 μM. Then, each sample was applied to a Superose 6 Increase 5/150 column (GE Healthcare) and eluted in 20 mM Tris-Cl (pH 7.14 at 20 °C), 150 mM NaCl and 0.05% DDM[4]. The concentration of the peak piericidin-bound fraction (at ~1.65 mL) was estimated as 4.1 mg mL^{-1} using a nanodrop UV–vis spectrophotometer ($\varepsilon_{280} = 0.2$ (mg mL^{-1})$^{-1}$). It was found to be 88.7 ± 3.4% inhibited by comparing initial rates of catalysis by the control and inhibited samples in 20 mM Tris-HCl (pH 7.5 at 20 °C), 0.012% asolectin and 0.012% CHAPS, using 5 μM DQ and 10 μg mL^{-1} AOX, with catalysis initiated by 200 μM NADH and monitored at 340–380 nm ($\varepsilon = 4.81$ mM^{-1} cm^{-1}). The low DQ concentration was used to minimise competition with the piericidin. The specific activity of the control (measured in 0.075% asolectin, 0.075% CHAPS and 100 μM DQ) was 11.3 ± 0.6 μmol min^{-1} mg^{-1}.

CryoEM grid preparation, data acquisition, and processing.
UltrAuFoil® gold grids (0.6/1, Quantifoil Micro Tools GmbH) were glow discharged (20 mA, 60 s), incubated in a solution of 5 mM 11-mercaptoundecyl hexaethyleneglycol (SPT-0011P6, SensoPath Technologies) in ethanol for 2 days in an anaerobic glovebox, then washed with ethanol and air-dried just before use[11]. Using an FEI Vitrobot Mark IV, 2.5 μL of complex I solution (3–5 mg mL^{-1}) was applied to each grid at 4 °C in 100% relative humidity and blotted for 10–12 s at force setting –10, before the grid was frozen by plunging it into liquid ethane. Grids for the piericidin-K2 dataset were imaged using a Gatan K2 detector and GIF Quantum energy filter mounted on an FEI 300 keV Titan Krios microscope with a 100 μm objective aperture and EPU-1.9 software at the UK National Electron Bio-Imaging Centre (eBIC) at Diamond. The energy filter was operated in zero-energy-loss mode with a slit width of 20 eV. Data were imaged at 1.05 Å pixel^{-1} (magnification 47,600×) with a defocus range −2.2 to −3.4 μm and the autofocus routine run every 10 μm. The dose rate was 5 electrons Å^{-2} s^{-1} with 10 s exposures captured in 25 frames (total dose ~50 electrons Å^{-2}). Grids for the piericidin-FIII dataset were imaged using a Falcon III detector on an FEI 300 keV Titan Krios microscope with a 100 μm objective aperture and EPU-1.9 at the Astbury Centre for Structural Molecular Biology, University of Leeds. The detector was operated in counting mode and data were imaged at 1.06 Å pixel^{-1} (magnification 130,000×) with a defocus range −2.2 to −3.8 μm and the autofocus routine run every 10 μm. The dose rate was 0.64 electrons Å^{-2} s^{-1} with 71.5 s exposures captured in 40 frames (total dose 46 electrons Å^{-2}).

All data were processed first using RELION-2.1-patchb1[47]. First, beam-induced movement was corrected for using MotionCor2[48], both with and without dose weighting. CFT estimations were taken from nondose weighted micrographs using GCTF-1.06[49]. Following autopicking and manual curation 60,107 particles were extracted from 1200 K2 micrographs and 76,802 particles from 1454 Falcon III micrographs. The particles were extracted from dose-weighted micrographs and CTF corrected with an amplitude contrast of 0.1 for 2D classification, then CTF parameters were re-estimated with an amplitude contrast of 0.08 thereafter. For the K2 dataset, frames 1–12 were used for movie refinement (total dose 24 electrons Å^{-2}) with a running average window of 3, and polished to account for movement and radiation damage using a single frame average for B-factor estimation. For the Falcon III dataset, frames 122 were used with a total dose of 25 electrons Å^{-2}, with a running average window of 5 and polished using a 3-frame average for B-factor estimation. The particles were then subjected to 3D classification into five classes with angular sampling gradually incremented to 0.9°. For the K2 dataset, the major class containing 27,193 particles was taken forward to the final 3D refinement with solvent flattened FSC curves. For the Falcon III dataset, two classes containing similar numbers of particles at the same resolution were selected and combined in a single class of 36,759 particles for refinement. Subsequently, both datasets plus the previously-described data for the active state of mouse complex I[4] were re-analysed using RELION-3.0. Beam-induced movement was re-analysed using the in-built version of MotionCor2 and the final particles were re-extracted in RELION-3.0, refined, and subjected to Bayesian polishing and CTF refinement, including beam tilt and per-particle astigmatism correction, before the final refinement with solvent flattened FSC curves[50]. 12 frames were used for both K2 datasets, and all 40 frames for the Falcon III dataset. All three maps were post-processed using default parameters using a mask created with the mask creation tool in RELION, and a soft edge of 5, starting from a Molmap command on a pdb model of the piericidin-inhibited enzyme in UCSF Chimera-1.13.1[51]. Map resolution estimates are based on the FSC = 0.143 criterion[52]. Pixel sizes were adjusted to three decimal places; the scaling was determined by using the Chimera "fit in map" command to maximise the correlation between the maps. Example micrographs and 2D classes for all three datasets are shown in Supplementary Fig 2 and local resolution estimates, particle orientation distributions and FSC curves are shown in Supplementary Fig 3.

Model building, refinement, and validation.
The model for the active mouse (6G2J.PDB[4]) was fitted into the piericidin-K2 map using Chimera[51], then refined against the RELION-sharpened piericidin-K2 map by cycles of manual adjustment

in Coot-0.8.9.1/0.9-pre[53] and real-space refinement in Phenix-1.13-2998 or 1.16-3549[54] with secondary structure restraints. Model building in poorly resolved areas was aided by the unsharpened map and a blurred map. The piericidin A molecule was imported from ChemSpider and restraints generated using AceDRG[55]. The piericidin-K2 final model was then fitted into the active map using Chimera; the piericidin was removed, the B-factors reset to 20, and the model refinement and inspection carried out as above. Model-to-map FSC curves were generated by simulating a map from the model at Nyquist frequency created with Molmap in Chimera[51]. The created map was compared to a masked unfiltered, unsharpened experimental map from RELION using the Xmipp tool in SCIPION-1.2[56]. Final model statistics were produced by Phenix-1.16-3549, MolProbity-4.4[57] (Supplementary Table 1) and EMRinger[58] (scores 3.77 for piericidin K2, 3.33 for active, and 3.06 for the piericidin-K2 model in the piericidin-FIII map). Finally, each model was checked for overfitting by first "shaking" it using Phenix simple dynamics and resetting its B-factors to 20, then refining it against an unsharpened half map filtered to the FSC = 0.143 resolution of the combined map, and comparing the model-to-map FSC curves against each unfiltered, masked half map. Local resolution maps were produced in RELION-3.0 and visualised in Chimera. Small changes to the model (Supplementary Table 1), relative to PDB ID: 6G2J include: identification of a *cis*-proline, present in high resolution structures of the homologous NiFe hydrogenase enzymes such as from *Ralstonia eutropha* (PDB ID: 4IUC)[59], next to cluster N2 (NDUFS7 Pro160); replacement of the ADP bound to NDUFA10 with ATP, with π-stacking between its adenine ring and nearby Phe134; replacement of two PE molecules (M505 and M506 from 6G2J) by a single cardiolipin (N501); improvements to poorly resolved areas such as the N-terminal loops of subunits NDUFS2, NDUFA13, NDUFB7, and NDUFB10.

EPR measurements. Piericidin-bound bovine complex I was prepared by combining the standard method for bovine complex I[8,32] with the method for the piericidin-bound mouse enzyme. Starting from 70 mg membrane protein, the sample was concentrated to 500 μL after ion-exchange chromatography. Two hundred and fifty microliter (the nonturnover control) were applied immediately to a 10/300 Superose 6 Increase column; the other 250 μL were divided in two and treated like the mouse preparation to generate a 125 μL "turnover control" without piericidin and a 125 μL sample inhibited with 320 μM piericidin. Both were applied to a 5/150 Superose 6 increase column. All three resulting samples were concentrated, and ~ 8 μL per sample was placed in a 1.6 mm O.D. Suprasil quartz EPR tube and frozen immediately. Their concentrations were 18 (nonturnover), 13.9 (turnover) and 10 (piericidin-containing) mg mL^{-1} and the maximal activity of the nonturnover control was 16.4 ± 0.2 μmol min^{-1} mg^{-1}. All the procedures described were carried out in air (not anaerobically), to match the conditions of cryo-EM grid preparation. EPR measurements were performed using an X/Q-band Bruker Elexsys E580 spectrometer (Bruker BioSpin GmbH, Germany) equipped with a closed cycle cryostat (Cryogenic Ltd. UK) and Bruker Xepr software. Measurements were carried out at X-band (9.5 GHz) using a split-ring resonator module with 2 mm sample access (ER 4118X-MS2)[60]. The temperature and magnetic field were calibrated with an external Cernox thermometer and a Bruker strong pitch sample (g = 2.0028) at room temperature. Acquisition times for spectra were approximately 12.5 min per sample. All spectra have been baseline-subtracted using an oxidised complex I sample.

Preparation and characterisation of proteoliposomes. Proteoliposomes were produced using complex I from bovine heart and recombinant *Trypanosoma brucei brucei* AOX[8,32]. Briefly, liposomes were formed from 8 mg phosphatidylcholine, 1 mg phosphatidylethanolamine and 1 mg cardiolipin (bovine heart extracts from Avanti Polar Lipids) together with varying amounts of Q$_{10}$ (Sigma-Aldrich) in 10 mM Tris-SO$_4$ (pH 7.5 at 4 °C) and 50 mM KCl. Following extrusion with a 0.1 μm track etched membrane, they were gently solubilized with 1.5% octyl-glucoside (Anatrace). AOX (0.2 mg) and then complex I (0.2 mg) were added for reconstitution. Aliquots of Biobeads (Bio-Rad) were added hourly over 4 h to remove the detergent then the proteoliposomes collected by centrifugation, resuspended, and flash frozen for storage at –80 °C. Total protein concentrations were quantified using the amido black assay, and the orientation of complex I using the NADH:APAD$^+$ activity assay[8,32]. Total phospholipid contents were determined as follows[8,32]. Hundred microliter of sample or KH$_2$PO$_4$ as standard, 50 μL of methanol and 30 μL of 390 mM Mg(NO$_3$)$_2$ in ethanol were heated in boiling tubes over a roaring blue flame until no further brown fumes were formed to leave a white residue. After 5 min, 0.3 mL of 500 mM HCl was added, and the tubes were lightly stoppered, incubated at 99 °C for 15 min, cooled, and then 0.7 mL of an aqueous solution of 114 mM ascorbic acid, 2.72 mM (NH$_4$)$_6$Mo$_7$O$_{24}$ and 400 mM H$_2$SO$_4$ was added, and the tubes incubated at 37 °C for 1 h. The absorbances were measured at 820 nm. One milligram of phospholipid was assumed to occupy a volume of 1 μL[8,32]. Q$_{10}$ was quantified by HPLC. Five to ten microliter of sample were solubilised in 190 μL HPLC grade ethanol, then 50 μL loaded onto a Nucleosil 100-5C18 (Hichrom) column maintained at 30 °C on an Agilent 1100 series HPLC system. The mobile phase (run at 800 μL min^{-1}) contained 70% ethanol, 30% methanol, 0.7% NaClO$_4$, 0.07% HClO$_4$[8] with elution monitored using a Dionex Ultimate 3000 RS electrochemical detector. A conditioning electrode (6020RS omni Coulometric cell) placed before the sample injector was set to +1000 mV and the dual electrodes of the detecting Coulometric cell (6011RS Ultra Analytical cell)

were set to –500 and +450 mV. The ubiquinone content was calculated from its +450 mV-peak area, by comparison to a set of standards. Ubiquinone concentrations are expressed as nmol of Q$_{10}$ per mg of phospholipids, which equates to mM concentration units.

Piericidin inhibition kinetics of complex I-AOX proteoliposomes. Catalytic activity assays were conducted at 32 °C in 10 mM Tris-SO$_4$ (pH 7.5 at 32 °C) and 50 mM KCl in 96-well plates using a Molecular Devices Spectramax 384 plus platereader with Softmax Pro 5.4.42.1. Piericidin was added in ethanol and catalysis initiated by addition of 200 μM NADH and monitored at 340 and 380 nm (ε$_{340-380}$ = 4.81 mM^{-1} cm^{-1}). Linear rates were measured, typically after a 100 s period of equilibration and/or activation. Inhibitor insensitive rates (recorded using 20 nM piericidin) were subtracted. The data are presented as mean values with standard deviations propagated from the quantifications of enzyme, sidedness, membrane volume, quinone content as well as the activity, which was performed in triplicate. The data were modelled using the ordinary differential equation solver, ode15s, in MATLAB (Mathworks, R2018a) with all the reverse rate constants (k_{-1}, k_{-3}, etc.) set to 1. k_4 was set to 1000 s^{-1} to ensure reoxidation of ubiquinol by AOX was not rate limiting[8]. Rate calculations were terminated once they reached steady state (judged as a change in enzyme-substrate complex concentration of <1 fM s^{-1} upon each simulated timestep). The overall rate of NADH oxidation was calculated by multiplying the steady-state concentration of enzyme-substrate complex with the turnover rate (k_2). Fitting was initiated repeatedly with a random set of values for the variable parameters, and targeted on the parameter combination with the smallest error between the calculated and measured rates; the data were modelled simultaneously over all Q$_{10}$ and piericidin concentrations. The piericidin log$_{10}$P value (4.73) was calculated in ChemSpider (http://www.chemspider.com). For most final models, bootstrapping of the residuals ($n = 1000$) was conducted using the triplicate activity measurements (rather than the means) in order to derive fitting statistics (mean, median, and 95% confidence intervals) of model parameters. For poorly-fitting models, uncertainties in fit parameters are the standard errors of fitting from the nonlinear regression.

Molecular simulations. All-atom classical molecular dynamics (MD) simulations (see Supplementary Table 4) were performed using our resolved piericidin-K2 structure of mouse complex I. Chains A, B, C, D (residues 34–430), H, I, J, K, P, a, b, e, r, W, X (residues 1–150), and Z were embedded in a POPC membrane and solvated with TIP3P water and 100 mM NaCl. The entire system comprised ~396,000 atoms. Force field parameters for the cofactors were derived from DFT models[61]. Parameters for piericidin were estimated initially using CGenFF[62], and refined at B3LYP-D3/def2-TZVP level[63]. Protonation states of key titratable amino acids were determined by Poisson–Boltzmann electrostatic (PBE) calculations[64] and long-range electrostatics treated by the Particle Mesh Ewald method. All simulations were performed at 1 atm, 310 K, with a 2 fs integration timestep. After initial minimisation, the system was heated up to the final temperature using a set of harmonic restraints (2 kcal mol^{-1} Å^{-2}) on the entire protein structure and cofactors (0.5 ns), then in a second step restrains were applied only to the backbone atoms (1 ns). Simulations were performed using NAMD 2.9/2.13[65] and analysed using VMD[66].

The initial structure of piericidin near Tyr108 was based on the cryoEM model (simulation A1–A3, A5–A6), or docked in de novo (simulation A4) using the cryoEM density. The initial structure of the piericidin in the lower part of the cavity (simulation A1, A2, and A4) was based on the local quinone-binding substrates described previously, that correspond to local binding minima identified in simulations of *T. thermophilus* complex I[34]. The second binding site was also probed by molecular docking calculations. Although some docked models could be identified with qualitative resemblance to the MD poses, the best scoring models were qualitatively incorrect, and the approach was therefore not employed for further exploration of the site. Water clustering analysis was performed on the last 40 ns (2000 frames) of simulation A1 using the WATCLUST method[67], with standard clustering values, and a 10% threshold for cluster recognition. Water molecules ~3 Å from residues NDUFS2 Asp107, Thr156, Lys371, Asp422, and Val424 were considered in the clustering analysis.

All setups were optimised using the molecular dynamics flexible fitting (MDFF) procedure[27]. MDFF simulations of piericidin were carried out using the inner core of the quinone-binding cavity (subunits ND1, NDUFS2, and NDUFS7), and imposing secondary structure restraints on the backbone atoms of the protein, plus a distance restraint of 1.7 Å between the Tyr108 hydroxyl and the piericidin 1′ carbonyl. After initial minimisation, the system was equilibrated at 310 K for 1 ns, followed by minimisation with high restraints on the experimental density. Simulations based on the cryoEM model were subject only to the latter. Energy decomposition analyses comprised all residues within 5 Å, and included interactions with protein backbone atoms.

Coarse-grained molecular dynamics (CGMD) simulations (see Supplementary Table 4) were performed using the MARTINI 2.2 force field[68]. In total, 10×10 μs simulations using a 20 fs timestep, a semi-isotropic Parrinello–Rahman barostat with a coupling constant of $\tau_p = 12.0$ ps, and a compressibility of $\chi = 3.0 \times 10^{-4}$ bar^{-1} were carried out. The simulation temperature was set to $T = 310$ K with a thermostat coupling constant of $\tau_t = 1.0$ ps. Nonbonded interactions were treated using a cutoff distance of 11 Å and $\varepsilon = 15$. Piericidin was placed in different

starting positions across the quinone channel. The CGMD simulations performed using Gromacs 2016.3[69].

To obtain a parameter-free estimation of binding poses in the active site, QM/MM MD simulations (see Supplementary Table 4) were performed with the piericidin, Tyr108, His59, Thr156, and Asp160 in the QM region treated at the B3LYP-D3/def2-SVP level of theory[63]. Protein residues within 10 Å of the piericidin were allowed to move during the dynamics. The total QM/MM system was trimmed to include 9100 atoms. The classical region was modelled at the CHARMM36 level of theory in combination with in-house DFT parameters for the cofactors (see above). All QM/MM simulations were performed using the CHARMM c38b, TURBOMOLE 6.6–7.3 and the CHARMM/TURBOMOLE interface[70].

A complete list of all simulations performed are shown in Supplementary Table 4.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Data accession codes: EMD-11424, PDB ID: 6ZTQ (piericidin-bound structure), EMD-11425 (piericidin-bound map from the FIII detector), EMD-11377, PDB ID: 6ZR2 (active state structure). Other data supporting the findings of this manuscript are available from the corresponding authors upon reasonable request.

Code availability

The Matlab algorithms used for kinetics modelling are available from the corresponding authors upon request.

References

1. Hirst, J. Mitochondrial complex I. *Annu. Rev. Biochem.* **82**, 551–575 (2013).
2. Rodenburg, R. J. Mitochondrial complex I-linked disease. *Biochim. Biophys. Acta* **1857**, 938–945 (2016).
3. Zhu, J., Vinothkumar, K. R. & Hirst, J. Structure of mammalian respiratory complex I. *Nature* **536**, 354–358 (2016).
4. Agip, A.-N. A. et al. Cryo-EM structures of complex I from mouse heart mitochondria in two biochemically defined states. *Nat. Struct. Mol. Biol.* **25**, 548–556 (2018).
5. Parey, K. et al. High-resolution cryo-EM structures of respiratory complex I: mechanism, assembly, and disease. *Sci. Adv.* **5**, eaax9484 (2019).
6. Baradaran, R., Berrisford, J. M., Minhas, G. S. & Sazanov, L. A. Crystal structure of the entire respiratory complex I. *Nature* **494**, 443–448 (2013).
7. Tocilescu, M. A., Fendel, U., Zwicker, K., Kerscher, S. & Brandt, U. Exploring the ubiquinone binding cavity of respiratory complex I. *J. Biol. Chem.* **282**, 29514–29520 (2007).
8. Fedor, J. G., Jones, A. J. Y., Di Luca, A., Kaila, V. R. I. & Hirst, J. Correlating kinetic and structural data on ubiquinone binding and reduction by respiratory complex I. *Proc. Natl Acad. Sci. USA* **114**, 12737–12742 (2017).
9. Kaila, V. R. I. Long-range proton-coupled electron transfer in biological energy conversion: towards mechanistic understanding of respiratory complex I. *J. R. Soc. Interface* **15**, 20170916 (2018).
10. Zickermann, V. et al. Mechanistic insight from the crystal structure of mitochondrial complex I. *Science* **347**, 44–49 (2015).
11. Blaza, J. N., Vinothkumar, K. R. & Hirst, J. Structure of the deactive state of mammalian respiratory complex I. *Structure* **26**, 312–319 (2018).
12. Degli Esposti, M. Inhibitors of NADH-ubiquinone reductase: an overview. *Biochim. Biophys. Acta* **1364**, 222–235 (1998).
13. Miyoshi, H. Structure-activity relationships of some complex I inhibitors. *Biochim. Biophys. Acta* **1364**, 236–244 (1998).
14. Murai, M. & Miyoshi, H. Current topics on inhibitors of respiratory complex I. *Biochim. Biophys. Acta* **1857**, 884–891 (2016).
15. Zhou, X. & Fenical, W. The unique chemistry and biology of the piericidins. *J. Antibiot.* **69**, 582–593 (2016).
16. Molina, J. R. et al. An inhibitor of oxidative phosphorylation exploits cancer vulnerability. *Nat. Med.* **24**, 1036–1046 (2018).
17. Bridges, H. R., Jones, A. J. Y., Pollak, M. N. & Hirst, J. Effects of metformin and other biguanides on oxidative phosphorylation in mitochondria. *Biochem. J.* **462**, 475–487 (2014).
18. Tamura, S. et al. Isolation and physiological activities of piericidin A, a natural insecticide produced by Streptomyces. *Agr. Biol. Chem.* **27**, 576–582 (1963).
19. Schnermann, M. J. et al. Total synthesis of piericidin A1 and B1 and key analogues. *J. Am. Chem. Soc.* **128**, 11799–11807 (2006).
20. Horgan, D. J., Ohno, H., Singer, T. P. & Casida, J. E. Studies on the respiratory chain-linked reduced nicotinamide adenine dinucleotide dehydrogenase. XV. Interactions of piericidin with the mitochondrial respiratory chain. *J. Biol. Chem.* **243**, 5967–5976 (1968).
21. Okun, J. G., Lümmen, P. & Brandt, U. Three classes of inhibitors share a common binding domain in mitochondrial complex I (NADH:ubiquinone oxidoreductase). *J. Biol. Chem.* **274**, 2625–2630 (1999).
22. Friedrich, T. et al. Two binding sites of inhibitors in NADH:ubiquinone oxidoreductase (complex I): relationship of one site with the ubiquinone-binding site of bacterial glucose:ubiquinone oxidoreductase. *Eur. J. Biochem.* **219**, 691–698 (1994).
23. Pätsi, J. et al. LHON/MELAS overlap mutation in ND1 subunit of mitochondrial complex I affects ubiquinone binding as revealed by modeling in *Escherichia coli* NDH-1. *Biochim. Biophys. Acta* **1817**, 312–318 (2012).
24. Gutman, M., Singer, T. P. & Casida, J. E. Studies on the respiratory chain-linked reduced nicotinamide adenine dinucleotide dehydrogenase: XVII. Reaction sites of piericidin A and rotenone. *J. Biol. Chem.* **245**, 1992–1997 (1970).
25. Gutman, M. Electron flux through the mitochondrial ubiquinone. *Biochim. Biophys. Acta* **594**, 53–84 (1980).
26. Yoga, E. G. et al. Mutations in a conserved loop in the PSST subunit of respiratory complex I affect ubiquinone binding and dynamics. *Biochim. Biophys. Acta* **1860**, 573–581 (2019).
27. Trabuco, L. G., Villa, E., Mitra, K., Frank, J. & Schulten, K. Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. *Structure* **16**, 673–683 (2008).
28. Reda, T., Barker, C. D. & Hirst, J. Reduction of the iron-sulfur clusters in mitochondrial NADH:ubiquinone oxidoreductase (complex I) by Eu^{II}-DTPA, a very low potential reductant. *Biochemistry* **47**, 8885–8893 (2008).
29. Le Breton, N. et al. Using hyperfine electron paramagnetic resonance spectroscopy to define the proton-coupled electron transfer reaction at Fe−S cluster N2 in respiratory complex I. *J. Am. Chem. Soc.* **139**, 16319–16326 (2017).
30. Kussmaul, L. & Hirst, J. The mechanism of superoxide production by NADH: ubiquinone oxidoreductase (complex I) from bovine heart mitochondria. *Proc. Natl Acad. Sci. USA* **103**, 7607–7612 (2006).
31. Zwicker, K. et al. The redox-Bohr group associated with iron-sulfur cluster N2 of complex I. *J. Biol. Chem.* **281**, 23013–23017 (2006).
32. Jones, A. J. Y. et al. A self-assembled respiratory chain catalyzes NADH oxidation through ubiquinone-10 cycling between complex I and the alternative oxidase. *Angew. Chem. Int. Ed. Engl.* **55**, 728–731 (2016).
33. Fedor, J. G. & Hirst, J. Mitochondrial supercomplexes do not enhance catalysis by quinone channeling. *Cell Metab.* **28**, 525–531 (2018).
34. Warnau, J. et al. Redox-coupled quinone dynamics in the respiratory complex I. *Proc. Natl Acad. Sci. USA* **115**, E8413–E8420 (2018).
35. Sharpley, M. S., Shannon, R. J., Draghi, F. & Hirst, J. Interactions between phospholipids and NADH:ubiquinone oxidoreductase (complex I) from bovine mitochondria. *Biochemistry* **45**, 241–248 (2006).
36. Jussupow, A., Di Luca, A. & Kaila, V. R. I. How cardiolipin modulates the dynamics of respiratory complex I. *Sci. Adv.* **5**, 1850 (2019).
37. Tocilescu, M. A. et al. The role of a conserved tyrosine in the 49-kDa subunit of complex I for ubiquinone binding and reduction. *Biochim. Biophys. Acta* **1797**, 625–632 (2010).
38. Grgic, L., Zwicker, K., Kashani-Poor, N., Kerscher, S. & Brandt, U. Functional significance of conserved histidines and arginines in the 49-kDa subunit of mitochondrial complex I. *J. Biol. Chem.* **279**, 21193–21199 (2004).
39. Degli-Esposti, M. Genome analysis of structure–function relationships in respiratory complex I, an ancient bioenergetic enzyme. *Genome Biol. Evol.* **8**, 126–147 (2015).
40. Gamiz-Hernandez, A. P., Jussupow, A., Johansson, M. P. & Kaila, V. R. I. Terminal electron−proton transfer dynamics in the quinone reduction of respiratory complex I. *J. Am. Chem. Soc.* **139**, 16282–16288 (2017).
41. Sinha, P. K. et al. Conserved amino acid residues of the NuoD segment important for structure and function of *Escherichia coli* NDH-1 (complex I). *Biochemistry* **54**, 753–764 (2015).
42. Angerer, H. et al. Tracing the tail of ubiquinone in mitochondrial complex I. *Biochim. Biophys. Acta* **1817**, 1776–1884 (2012).
43. Fendel, U., Tocilescu, M. A., Kerscher, S. & Brandt, U. Exploring the inhibitor binding pocket of respiratory complex I. *Biochim Biophys. Acta* **1777**, 660–665 (2008).
44. Prieur, I., Lunardi, J. & Dupuis, A. Evidence for a quinone binding site close to the interface between NUOD and NUOB subunits of complex I. *Biochim. Biophys. Acta* **1504**, 173–178 (2001).
45. Schipke, C. G., Goodin, D. B., McRee, D. E. & Stout, C. D. Oxidized and reduced Azotobacter vinelandii ferredoxin I at 1.4 Å resolution:

conformational change of surface residues without significant change in the [3Fe-4S]$^{+/0}$ cluster. *Biochemistry* **38**, 8228–8239 (1999).

46. Berrisford, J. M. & Sazanov, L. A. Structural basis for the mechanism of respiratory complex I. *J. Biol. Chem.* **284**, 29773–29783 (2009).

47. Kimanius, D., Forsberg, B. O., Scheres, S. H. W. & Lindahl, E. Accelerated cryo-EM structure determination with parallelisation using GPUs in RELION-2. *Elife* **5**, e18722 (2016).

48. Zheng, S. Q. et al. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2017).

49. Zhang, K. Gctf: real-time CTF determination and correction. *J. Struct. Biol.* **193**, 1–12 (2016).

50. Zivanov, J. et al. New tools for automated high-resolution cryo-EM structure determination in RELION-3. *eLife* **7**, e42166 (2018).

51. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).

52. Rosenthal, P. B. & Henderson, R. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *J. Mol. Biol.* **333**, 721–745 (2003).

53. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Cryst.* **D66**, 486–501 (2010).

54. Afonine, P. V. et al. Real-space refinement in PHENIX for cryo-EM and crystallography. *Acta Crystallogr. D* **74**, 531–544 (2018).

55. Long, F. et al. AceDRG: a stereochemical description generator for ligands. *Acta Crystallogr. D* **73**, 112–122 (2017).

56. de la Rosa-Trevin, J. M. et al. Scipion: a software framework toward integration, reproducibility and validation in 3D electron microscopy. *J. Struct. Biol.* **195**, 93–99 (2016).

57. Williams, C. J. et al. MolProbity: more and better reference data for improved all-atom structure validation. *Protein Sci.* **27**, 293–315 (2018).

58. Barad, B. A. et al. EMringer: side chain-directed model and map validation for 3D cryo-electron microscopy. *Nat. Methods* **12**, 943–946 (2015).

59. Frielingsdorf, S. et al. Reversible [4Fe-3S] cluster morphing in an O$_2$-tolerant [NiFe] hydrogenase. *Nat. Chem. Biol.* **10**, 378–385 (2014).

60. Wright, J. J., Salvadori, E., Bridges, H. R., Hirst, J. & Roessler, M. M. Small-volume potentiometric titrations: EPR investigations of Fe-S cluster N2 in mitochondrial complex I. *J. Inorg. Biochem.* **162**, 201–206 (2016).

61. Sharma, V. et al. Redox-induced activation of the proton pump in the respiratory complex I. *Proc. Natl Acad. Sci. USA* **112**, 11571–11576 (2015).

62. Vanommeslaeghe, K. et al. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.* **31**, 671–690 (2009).

63. Becke, A. D. Density-functional thermochemistry. III. The role of exact change. *J. Chem. Phys.* **98**, 5648–5652 (1993).

64. Di Luca, A., Gamiz-Hernandez, A. P. & Kaila, V. R. I. Symmetry-related proton transfer pathways in respiratory complex I. *Proc. Natl Acad. Sci. USA* **114**, E6314–E6321 (2017).

65. Phillips, J. C. et al. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **26**, 1781–1802 (2005).

66. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph.* **14**, 33–38 (1996).

67. López, E. D. et al. WATCLUST: a tool for improving the design of drugs based on protein-water interactions. *Bioinformatics* **31**, 3697–3699 (2015).

68. Marrink, S. J., Risselada, J., Yefimov, S., Tieleman, D. P. & de Vries, A. H. The MARTINI force field: coarse grained model for biomolecular simulations. *J. Phys. Chem. B* **111**, 7812–7824 (2007).

69. Berendsen, H. J. C., van der Spoel, D. & van Drunen, R. GROMACS: a message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.* **91**, 43–56 (1995).

70. Riahi, S. & Rowley, C. N. The CHARMM–TURBOMOLE interface for efficient and accurate QM/MM molecular dynamics, free energies, and excited state properties. *J. Comp. Chem.* **35**, 2076–2086 (2014).

Author contributions

H.R.B., J.N.B., and A.N.A.A. prepared complex I and cryoEM grids and processed cryoEM data; H.R.B. and J.N.B. collected cryoEM data and built the models; H.R.B., J.G.F., J.N.B., V.R.I.K., and J.H. analysed and interpreted the models; A.D.L., A.J., A.P.G. H., and V.R.I.K. performed and analysed molecular simulations; J.G.F. performed kinetic modelling; J.J.W. and M.M.R. performed and analysed EPR spectroscopy; V.R.I.K. and J.H. directed the project; H.R.B., J.G.F., V.R.I.K., and J.H. wrote the manuscript with input from all authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-020-18950-3.

Correspondence and requests for materials should be addressed to V.R.I.K. or J.H.

Peer review information *Nature Communications* thanks Marcelo Martí, Tim Rasmussen and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

BIOPHYSICS

The dynamics of linear polyubiquitin

Alexander Jussupow[1], Ana C. Messias[2,3], Ralf Stehle[2,3], Arie Geerlof[2,3], Sara M. Ø. Solbak[4], Cristina Paissoni[5], Anders Bach[4], Michael Sattler[2,3]*, Carlo Camilloni[1,5]*

Polyubiquitin chains are flexible multidomain proteins, whose conformational dynamics enable them to regulate multiple biological pathways. Their dynamic is determined by the linkage between ubiquitins and by the number of ubiquitin units. Characterizing polyubiquitin behavior as a function of their length is hampered because of increasing system size and conformational variability. Here, we introduce a new approach to efficiently integrating small-angle x-ray scattering with simulations allowing us to accurately characterize the dynamics of linear di-, tri-, and tetraubiquitin in the free state as well as of diubiquitin in complex with NEMO, a central regulator in the NF-κB pathway. Our results show that the behavior of the diubiquitin subunits is independent of the presence of additional ubiquitin modules and that the dynamics of polyubiquitins with different lengths follow a simple model. Together with experimental data from multiple biophysical techniques, we then rationalize the 2:1 NEMO:polyubiquitin binding.

INTRODUCTION

Ubiquitination is a reversible posttranscriptional modification system that regulates key physiological processes, such as protein degradation, cell cycle, apoptosis, DNA repair, and signal transduction (*1–3*). Once a protein substrate is monoubiquitinated (e.g., a lysine of the substrate is conjugated through an isopeptide bond to the C terminus of a ubiquitin monomer), an additional ubiquitin may be conjugated to either one of the seven lysine residues of the first ubiquitin (K6, K11, K27, K29, K33, K48, and K63) (*4*) or its N-terminal methionine residue (M1) (*5–7*). This process can lead to the assembly of polyubiquitin chains of various lengths and topologies. The resulting polymeric chains are then associated with different cellular mechanisms (*8*). Since all these polymers are made of the same single unit, the highly conserved 76-residue-long ubiquitin domain, the ubiquitin code is an example of a conformation-based alphabet, where both the polymerization site (*8, 9*) and the chain length (*10*) regulate the recognition by different partners and thereby determine the cellular fate of the protein. The role of polyubiquitin length and dynamics in molecular recognition processes is poorly understood (*8, 10, 11*). An overall assessment of the typical length of different polyubiquitin chains in physiological conditions is missing, and only sporadic indications are available. For example, in the case of K48-linked polyubiquitin, a length of four is generally considered optimal for molecular recognition of the 26*S* proteasome (*12*), while the nuclear protein localization protein 4 is selective for K48-linked chains longer than six (*13*). It was reported that K48-linked tetraubiquitin (Ub₄) slows down further ubiquitination (*14–16*), while this is not the case for K63-linked Ub₄ (*16*).

Linear M1-linked polyubiquitin chains (Fig. 1), whose assembly is catalyzed by linear ubiquitin chain assembly complex (LUBAC) (*5*), have been shown to play a role in inflammation, immune responses,

and oncogenesis (*17–19*). Their most studied function is the involvement in the activation of the canonical nuclear factor κB (NF-κB) pathway (*6, 7, 17, 20–23*). In this pathway, the IKK complex [or IκB (inhibitor of the NF-κB proteins) kinase, formed by IKKα, IKKβ, and NEMO, also known as IKKγ, the NF-κB essential modulator] is activated by LUBAC upon activation by various stimuli (*22*). LUBAC preferentially recognizes and conjugates linear ubiquitin chains on NEMO. NEMO also has a specific linear diubiquitin-binding region referred to as the "ubiquitin binding in A20-binding inhibitor of NF-kappa-B activation (ABIN) and NEMO" (UBAN) motif (*24*), which forms a helical coiled-coil dimer in solution (*23*). Recognition of a linear polyubiquitin conjugated to NEMO by the UBAN domain of another NEMO may trigger the clustering of the IKK complex and conformational changes that subsequently activate IKK (*25, 26*). Once active, IKK can phosphorylate and inactivate the IκBs, leading to the release of NF-κB (*27*). It was recently shown that it is possible to inhibit NF-κB activation upon UBAN-dependent tumor necrosis factor–α and T cell receptor/CD28 stimulation by small molecules that inhibit the binding of linear polyubiquitins to the NEMO_UBAN domain (*23*). While the NEMO_UBAN domain can bind linear diubiquitin, it has been observed that full-length NEMO can only bind Ub₄ or longer, suggesting a length-dependent activation mechanism (*21*). Furthermore, another study suggested that the binding of NEMO to chains of 10 linear ubiquitins or longer induces a different conformation of NEMO compared to the binding of shorter chains (*20*).

Characterizing the conformational space of polyubiquitin chains as a function of length is critical to understand their physiological behavior. Such structural characterization is nonetheless very challenging. Polyubiquitins, from diubiquitin to longer chains, exhibit a very dynamic behavior (*28*) that requires determining a statistical ensemble of all the relevant configurations populated in solution. The combination of molecular dynamics (MD) with experimental small-angle x-ray scattering (SAXS) data is very well suitable to study dynamic protein systems (*29*) including polyubiquitin of varying chain size. SAXS does not provide high-resolution structural information. Conversely, MD simulations may be used to determine the statistical ensemble of configuration populated by a system in equilibrium condition, but a full modeling base on MD simulations is hampered by the size of the system (*30, 31*). This problem can, in principle, be alleviated by coarse-grain force fields (*32*), eventually

[1]Department of Chemistry and Institute for Advanced Study, Technical University of Munich, Garching 85747, Germany. [2]Institute of Structural Biology, Helmholtz Zentrum München, Neuherberg 85764, Germany. [3]Center for Integrated Protein Science Munich at Department of Chemistry, Technical University of Munich, Garching 85747, Germany. [4]Department of Drug Design and Pharmacology, Faculty of Health and Medical Sciences, University of Copenhagen, Universitetsparken 2, DK-2100 Copenhagen, Denmark. [5]Dipartimento di Bioscienze, Università degli studi di Milano, 20133 Milano, Italy.
*Corresponding author. Email: sattler@helmholtz-muenchen.de (M.S.); carlo.camilloni@unimi.it (C.C.)
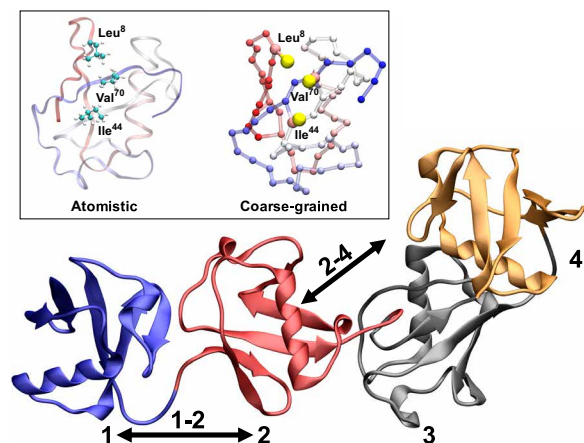
Fig. 1. Schematic illustrations linear polyubiquitins. Cartoon representation of linear tetraubiquitin; the ubiquitin domains are numbered from the N terminus to the C terminus from 1 to 4. The inset shows an atomistic and coarse-grained (CG) (Martini) representation of the hydrophobic patch (Ile^{44}, Val^{70}, and Leu^8) of the ubiquitin domain.

combined with enhanced sampling techniques (*33*), that can massively speed up MD simulations although potentially at the expense of the accuracy (*32*).

Here, we show that by integrating SAXS and MD simulations based on the Martini coarse-grain force field (*34*, *35*) by means of metainference (*36*), we can efficiently generate an ensemble of structures representing the dynamics of linear polyubiquitins (Fig. 1). The ensembles allow the description of the dynamics of these complex systems at the single-residue level. Our results show how polyubiquitins can populate multiple conformational states but unexpectedly indicate that linear polyubiquitin chains and potentially polyubiquitins, in general, can be described by a simple self-avoiding polymer model. Various biophysical experiments are used to characterize the stoichiometry, kinetics, and thermodynamic properties of the binding of polyubiquitin to the NEMO$_{UBAN}$ domain. Unexpectedly, our data demonstrate that NEMO$_{UBAN}$ binds to di-, tri-, and tetraubiquitin (Ub$_2$, Ub$_3$, and Ub$_4$) in all cases forming a 2:1 NEMO$_{UBAN}$:Ub$_N$ complex in solution. Notably, a conformational ensemble for the NEMO$_{UBAN}$:Ub$_2$ complex rationalizes the 2:1 binding. Combined with our proposed polyubiquitin polymer model, this suggests how longer polyubiquitin chains may modulate NEMO recognition as well as bind more than one NEMO dimer.

RESULTS

A simple Martini modification improves the simulation of linear diubiquitin

We first evaluated the ability of the Martini coarse-grain force field to describe the dynamics of a linear Ub$_2$. A metadynamics (*37*) simulation of Martini Ub$_2$ resulted in an extremely compact ensemble of structures (Fig. 2A), which does not reproduce the measured SAXS intensities (Fig. 2D and fig. S1). In Fig. 2A, we report a free energy landscape (in kilojoules per mole) as a function of the distance between the centers of the two ubiquitin domains and their relative orientation; the average distance between the two domains is very short, around 2.41 ± 0.02 nm, with a preferential orientation of the two ubiquitin's domain (measured as the torsion angle between two

axes defined using the first and second half of the sequence of each ubiquitin; cf. Methods). The average radius of gyration of 1.73 ± 0.01 nm strongly underestimates the value of 2.23 ± 0.02 nm derived from SAXS (fig. S1). The ensemble seems to be able only to capture compact Ub$_2$ configurations also when compared to the available crystal structures [PDB (Protein Data Bank) 2W9N (*38*) (open), 3AXC (*39*) (compact), and 4ZQS (*28*) (compact)]. This result is not unexpected for the Martini force field, and both weakening the protein-protein interactions (*40*, *41*) and increasing the protein-solvent interaction (*42*, *43*) have been suggested as possible solutions. A more complex water model, such as the Martini polarizable water (*44*), is also available, but at the expense of performance. Recent developments in atomistic force fields demonstrated the need for tuning solute-solvent interactions (*45*, *46*). Following recent approaches that have successfully improved atomistic and coarse-grained (CG) force fields, we repeated the same simulation after increasing by 5% the Martini water-protein Lennard-Jones interaction. This simple adjustment was sufficient to obtain a more expanded ensemble of structures as shown by the free energy landscape (Fig. 2B), without any additional computational cost (fig. S2). The new ensemble resulted in an improved, even if not yet quantitative, agreement with the SAXS data (Fig. 2D, blue curve, and fig. S1). The average distance between the domains increased to 3.10 ± 0.02 nm, and the protein can explore a much wider conformational space that now includes open and closed structures. In terms of the radius of gyration, the ensemble average resulted in 2.05 ± 0.01 nm to be compared with the 2.23 ± 0.02 nm derived from SAXS. Of note, comparing the free energy surface of the underdevelopment version of the Martini force field (Martini 3, currently in beta phase) with Martini 2.2 shows promising behavior by exploring more open conformation but may still benefit from increased protein-water (P-W) interaction (fig. S2). Nonetheless, our aim here is to obtain ensembles in quantitative agreement with the SAXS data without a large-scale force field reparameterization effort. This can be achieved at least, in principle, by integrating experimental information directly in the simulation by metainference (*36*). To show this, we run metadynamic metainference (M&M) (*47*) simulations (see Methods) with the K63-linked Ub$_2$ and compared them with published atomistic ensembles (fig. S3). In Paissoni *et al.* (*48*), an ensemble of K63-linked Ub$_2$ was generated through atomistic simulation with integrated SAXS data and validated against nuclear magnetic resonance (NMR) data. Overall, our energy landscape and the distribution of the radius of gyrations are comparable to the atomistic ones. The global minima region of the free energy landscape is correctly identified (red box in fig. S3), with a substantial improvement over an unrestraint atomistic ensemble (*48*). However, the modified Martini force field fails to correctly identify the conformation of very compact states (which are missing in the unrestrained atomistic ensemble). Comparing the contact maps shows that our approach still manages to identify the correct interdomain contact regions while being less specific. A weakening of the elastic network, which stabilizing the core of ubiquitin subunits (see Methods), or a further increase of the P-W interaction does not lead to improvement (fig. S4). A simple excluded volume model with integrated SAXS data is not sufficient to achieve qualitatively similar free energy surface (fig. S3). This shows that at least a qualitatively accurate description of the interdomain interactions is necessary to generate a precise SAXS ensemble.

Overall, our approach of coupling a modified Martini force field with SAXS manages to capture the overall balance between the
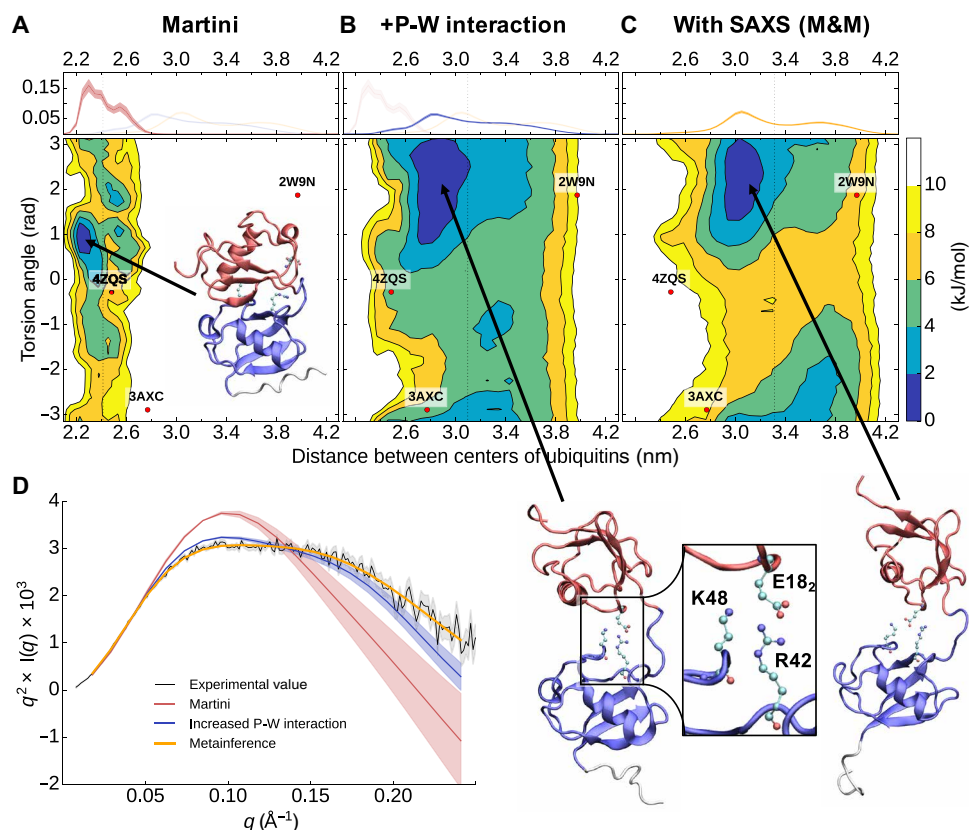
Fig. 2. Characterization of the dynamics of linear diubiquitin. (**A** to **C**) Free energy landscapes (in kilojoules per mole) as a function of the distance between the center of mass of the two ubiquitin domains and their relative orientation (measured as the torsion angle between two axes defined using the first and second half of the sequence of each ubiquitin; see Methods). The dots represent the coordinates associated with the available diubiquitin crystal structures. On top is shown the probability distribution of the distance between the centers of the two ubiquitin domains. (**D**) Experimental and from-simulation calculated Kratky plot. The shaded area represents the error range.

compact and open states. It allows us to identify the global minima conformations and interdomain contact regions, but it might fail to capture specific compact states accurately. While a restrained atomistic simulation can lead to more accurate results, having a few hundred times faster sampling (fig. S2E) is a substantial benefit, making the sampling of more extended ubiquitin chains systems accessible.

Metainference SAXS simulations of Martini linear diubiquitin quantitatively reproduce the experimental data

M&M (*47*) simulation (see Methods) for Martini linear Ub$_2$ (including our modified water) result in an ensemble of configurations characterized by a flatter and broader free energy landscape (Fig. 2C) and in quantitative agreement with the experimental SAXS (Fig. 2, D and E, and fig. S1). With respect to the unrestrained simulation, the average distance between the two domains increased from 3.10 ± 0.02 nm to 3.32 ± 0.02 nm. The radius of gyration of the ensemble of 2.23 ± 0.01 nm quantitatively agrees with that derived from SAXS of 2.23 ± 0.02 nm. Qualitatively, the topology of the free energy landscape is comparable to the unrestrained simulation but translated to larger relative distances. Overall, the free energy landscape is quite flat with relatively limited free energy differences indicating that the two ubiquitin domains are relatively free to move with respect to each other. Therefore, Ub$_2$ shows highly dynamical behavior, which cannot be described by a few individual structures. Instead, a full

ensemble is required in agreement with previous findings on linear as well as other diubiquitins.

From the performance point of view, the SAXS on-the-fly calculation used by metainference is computationally demanding, but the use of a CG representation makes it far more affordable with respect to the same simulation performed at full atomistic resolution (fig. S2). The loss of performance resulting from the use of SAXS is justified by the increased accuracy of the resulting simulations. Note that it is not required to calculate the metainference SAXS restraint at every step of the simulation. By calculating it every five steps, we obtained a quantitatively equivalent ensemble at a fraction of the computational cost (fig. S2). Notably, using metainference allows us also to sample the scaling value, which is necessary to compare the experimental and computed SAXS curves. For Ub$_2$, we observed a 3% higher scaling value for the simulation with increased P-W interaction and a 9% higher scaling value just with the Martini force field compared to the metainference solution (fig. S1).

Linear polyubiquitin chains are preferentially extended, do not show long-range correlations, and can be described as self-avoiding polymers

To investigate the dynamics of linear Ub$_3$ and Ub$_4$, we performed SAXS experiments on both proteins at different concentrations (fig. S5). The measured SAXS data were then used to perform M&M

simulations (cf. Methods and table S1). In addition, unrestrained simulations based only on Martini with our modified water were also performed. In Fig. 3 (see also fig. S1), we show the comparison of the back-calculated SAXS with respect to the experimental measures of Ub_3 and Ub_4. The effect of our improved water diminishes for the longer polyubiquitin chains. A comparison of the radius of gyration for the Ub_2, Ub_3, and Ub_4 ensembles show that while the unrestrained and restrained simulations sample a comparable range of compactness, the restrained simulations are shifted toward a more extended conformational space. The trend of the average radius of gyration (2.0, 2.7, and 3.3 nm for Ub_2, Ub_3, and Ub_4, respectively; cf. Methods and fig. S1A) suggests an almost linear increase of the size of the protein with the number of ubiquitin monomers. The analysis of the free energy landscape for the Ub_2 couples in Ub_3 and Ub_4 (fig. S6) qualitatively shows the same behavior, suggesting that the inter-domain interactions are essentially only those between neighbor domains (i.e., between Ub_2). For Ub_3 and Ub_4, this is confirmed by analyzing the free energy landscape of non-neighbor ubiquitin domains. Overall, the free energy landscape is flatter for larger polyubiquitin chains, indicating that the interaction with neighboring ubiquitin becomes less and less specific. Also, the distance (centers of the two ubiquitin domains) distribution shifts from a bimodal distribution for Ub_2 to a flatter one for Ub_3 and Ub_4. We also observe that Ub_3 samples more extended conformations for large distances >4.0 nm, while both Ub_4 and Ub_3 are forming more compact conformation below 2.6 nm. In fig. S7 (A to C), the free energy profiles are shown for the first-and-third ubiquitin $Ub_3(1-3)$ in Ub_3 as well as for the first-and-third $Ub_4(1-3)$ and second-and-fourth $Ub_4(2-4)$. These landscapes are all qualitatively similar showing that the interaction between two non-neighbor ubiquitins is quite rare. The average distance between a 1-3 and 2-4 ubiquitin pair is around 6 nm, with an average angle of around 140°. The first ubiquitin does not influence the relative orientation of the third ubiquitin. The

first-and-fourth $Ub_4(1-4)$ ubiquitin couple, as shown in fig. S7D, behaves similarly. Interactions between the first and fourth ubiquitin are also rare. In most cases, the distance between both ubiquitins is around 8.5 nm. There is also no strong preference for a specific torsion angle between all four ubiquitins.

To further assess the presence of short and long-range interactions between neighbor and non-neighbor ubiquitin couples, we estimated the fraction of compact configurations by analyzing the minimum distance between neighbor and non-neighbor ubiquitin couples (Fig. 4, A and B). For neighbor and non-neighbor couples, there is a peak in the distribution around 0.5 nm. As already indicated by the free energy profiles, compact neighbor ubiquitin pairs represent around 40 to 50% of the ensemble, while contacts between non-neighbor couples are only present in around 8% for Ub_3 and around 2% for Ub_4, indicating an overall lack of compact states in linear polyubiquitins. A contact analysis for the Ub_2 compact state indicates that this state is not structurally homogeneous. Even the most frequent contact is only present in 10 to 30% of all compact conformations, depending on the specific ubiquitin pair (Fig. 4C). On the other hand, even the 10th most frequent contact has still a probability between 5 and 15%, while the 100th most frequent one is still in the 1 to 5% range.

Nonetheless, all residues involved in the most frequent contacts belong to three distinct surfaces. These interactions define the preferred orientations between two adjacent ubiquitin pairs. All residues involved in the most frequent contacts from the first ubiquitin are on the same surface as the hydrophobic patch I44 (Fig. 4D), which is known to be also essential for interactions of ubiquitin with other proteins. The hydrophobic patch is also the main contributor for the interdomain contacts in the K63 Ub_2 (fig. S3). This result is consistent between atomistic and coarse-grain simulation. The I44 surface interacts either with the surface around $E18_2$ or $I13_2$ ($E18$ or $I13$ would be the analog residues of the first ubiquitin). The $E18_2$ surface is
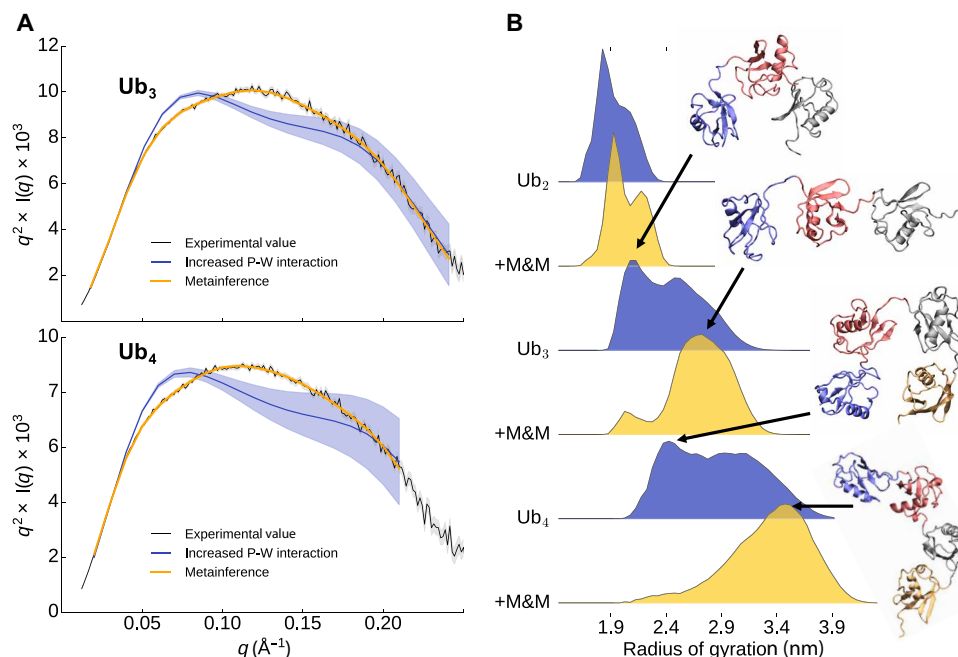

Fig. 3. Characterization of the dynamics of linear tri- and tetraubiquitin. (**A**) Experimental and from-simulations calculated Kratky plot for tri- and tetraubiquitin. The shaded area represents the error range. (**B**) Distribution of the radius of gyration from the ensembles with and without M&M.
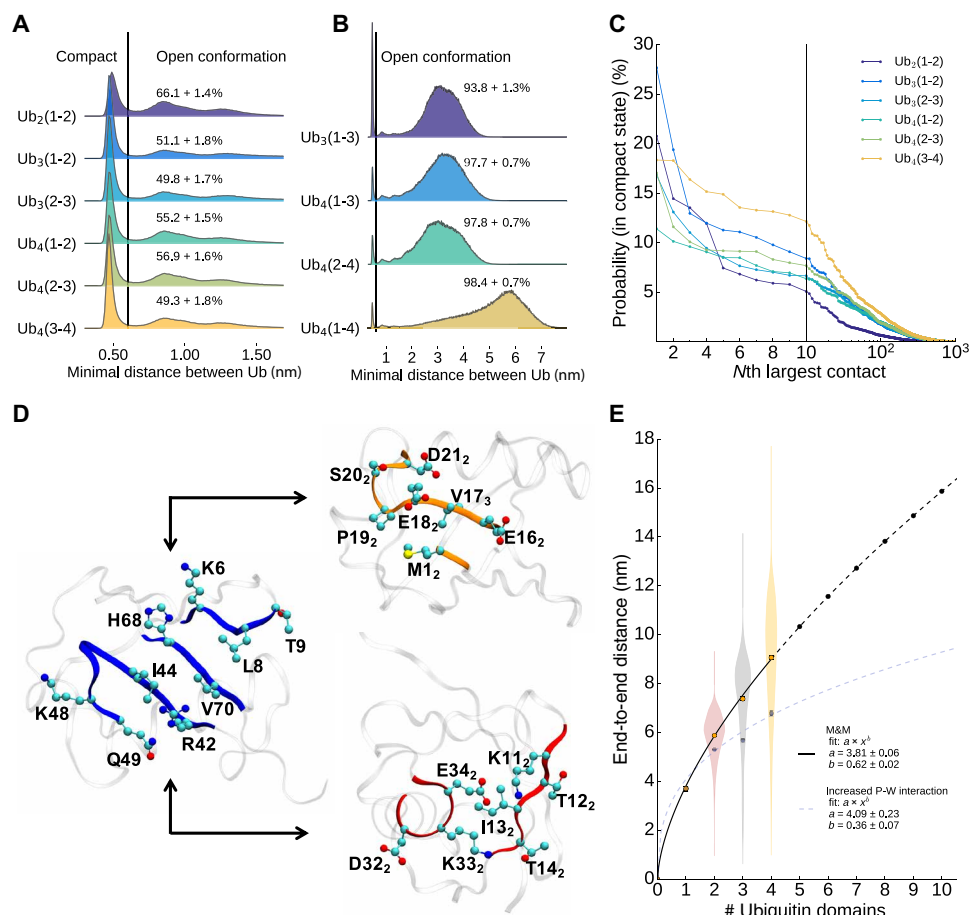
Fig. 4. Intramolecular interactions of polyubiquitin. (**A**) Minimum distance distribution between two neighboring ubiquitin cores (residues 1 to 70, residues 77 to 146, and so forth). Structures with a minimal distance larger than 0.6 nm are defined as open. (**B**) Minimum distance distribution between two non-neighboring ubiquitin cores. Structures with a minimal distance larger than 0.6 nm are defined as open. (**C**) Probability of finding contacts between two amino acids of neighboring ubiquitin cores. (**D**) Interaction surface of two neighboring ubiquitins. Residues from the blue marked surface (first ubiquitin, left) are interacting with residues of the orange marked surface (middle) or red marked surface (right) of the second ubiquitin. (**E**) Average end-to-end distance of a linear polyubiquitin chain.

located opposite to the I44 surface, while the $I13_2$ surface is roughly 90° rotated to the I44 and $E18_2$ surface. In fig. S6 (H to M), the 10 most frequent contacts between ubiquitin cores for all adjacent ubiquitin pairs are illustrated. Ub_2 is predominantly stabilized by salt bridges between the positive charged R42 and K48 and the negative charged $E16_2$ and $E18_2$. However, going to Ub_3 and Ub_4, electrostatic interactions become less important compared to van der Waals (vdW) interactions (fig. S7F). For Ub_2, the Coulomb interaction between charged amino acids is responsible for 27% of the total interaction energy between two ubiquitin cores. This value goes as low as 9% for the Ub (3-4) pair of Ub_4. Since the Martini force field has limitations in terms of electrostatics, the absolute ratios between vdW and Coulomb interactions are likely not meaningful. However, since our modified Martini force field with SAXS data manages to capture the important interaction regions, the relative changes between different ubiquitin pairs are likely qualitatively correct. The lower electrostatic interactions may also explain the flatter free energy surfaces of Ub_3 and Ub_4 neighbor pairs (fig. S4). On the other hand, the increased role of vdW interactions in Ub_3 and Ub_4 is compatible with the increase in the compact population of diubiquitin couples. Last, while in Ub_2 the I44 surface prefers to interact with the $E18_2$

surface, interactions between the I44 and the $I13_2$ surface are more important for the last pair of Ub_4, causing a shift of the preferred orientation between both ubiquitins.

Overall, our linear Ub_2, Ub_3, and Ub_4 ensembles indicate that linear polyubiquitins are extended polymers, whose dynamics are mostly uncorrelated over a distance of more than one ubiquitin domain. In Fig. 4E, this behavior is further highlighted by plotting the end-to-end distance as a function of the number N of ubiquitin domain, $e2e(N)$. Fitting the data with a power law (including the end-to-end distance for Ub_1) resulted in $e2e(N) = 3.81N^{0.62}$ in remarkable agreement with Flory theory for self-avoiding polymers (*49*). This is remarkable, since generally, proteins do not behave as self-avoiding chain, showing also less entropy in the denatured state. This behavior is not shown in simulations just with increased P-W interactions, which have a substantially lower exponent of 0.36. It is tempting to speculate that all polyubiquitins may be described as self-avoiding polymers following the same relationship for the end-to-end distance but with a different prefactor (i.e., characteristic length) associated with the distance between the C-terminal glycine and the specific linkage side chain. For K63-linked polyubiquitins, we could test this, even if only with the monoubiquitin and diubiquitin,

making use of a SAXS-based ensemble that we recently published (*48*). By fixing the exponent to 0.6 and setting the prefactor to the average distance between the C terminus and K63, the expected *e2e* distance of Ub_2 fits the ensemble. This suggests that the fit can be used to predict the behavior for longer K63-linked Ub_N chains (fig. S6). Extrapolating for other linkages, we observe that K11- and K48-linked polyubiquitin, which are known to populate more compact states in solution (*50, 51*), have a shorter distance between the C terminus and the lysine and would consequently show smaller prefactors and populate systematically more compact states than K63- and M1-linked.

The conformational entropy of long linear polyubiquitins modulates NEMO binding

To study the dynamics and changes of linear polyubiquitin dynamics upon binding to cognate proteins, we used simulations and experiments to characterize the interaction of the NEMO UBAN domain ($NEMO_{258-350}$) to the linear polyubiquitins Ub_2, Ub_3, and Ub_4. The NEMO UBAN domain is a dimer in solution (*23*). Previous studies have shown two different binding stoichiometries in solution and crystalline state for Ub_2: with either two NEMO monomers bound to one Ub_2(2:1) or two NEMO monomers bound to two Ub_2 (2:2) (*23, 24*). Polyubiquitin chains longer than Ub_2 harbor potentially more than one binding site and could thus bind more than one NEMO dimer with a theoretical stoichiometry for Ub_N of $2(N − 1){:}1$ ($NEMO{:}Ub_N$). A crystal structure (PDB 5H07) shows the binding of two linear Ub_3 to four ABIN monomers (a homolog of NEMO that also forms a dimer in solution) (*52*). The observed binding mode requires Ub_3 to be in a relatively compact and univocally oriented configuration to avoid steric hindrances between the two ABIN dimers. This would markedly decrease the entropy not only of each diubiquitin couple but also that of the overall chain and thus should be entropically disfavored in solution (fig. S7). Isothermal titration calorimetry (ITC) experiments show that in solution, only one ABIN dimer binds to Ub_3 (*52*), arguing that higher stoichiometries are artifactual, induced by crystal packing, and do not reflect the solution assembly.

To characterize the binding in solution of NEMO to Ub_3 and Ub_4, we performed SAXS (Fig. 5 and fig. S5), ITC (Fig. 5 and fig. S9), size exclusion chromatography (SEC) coupled with static light scattering (SLS) (Fig. 5, fig. S9, and table S2), and surface plasmon resonance (SPR) (fig. S9). Extending Ub_2 to Ub_3 and Ub_4 does not affect the (2:1) binding stoichiometry, with either polyubiquitin protein binding two NEMO monomers (one NEMO dimer) (Fig. 5). In SEC experiments with SLS, we detected $NEMO{:}Ub_3$ complexes with molecular weights (MWs) ranging from 41 to 45 kDa, which is similar to the MW of the calculated 2:1 $NEMO{:}Ub_3$ complex (47.7 kDa), while for $NEMO{:}Ub_4$ complexes only a single peak is found with MW between 53 and 56 kDa (calculated MW 2:1 $NEMO{:}Ub_4$, 56.2 kDa). SAXS measurements confirmed the stoichiometry observed by SLS-SEC. ITC indicates that NEMO binds to Ub_3 and Ub_4 with very similar enthalpy (ΔH of −17.9 and −18.8 kJ/mol, respectively), suggesting that the molecular interactions and binding interfaces between NEMO and the different polyubiquitins are similar to the one described for $NEMO{:}Ub_2$ (*23*) (ΔH of −16.9 kJ/mol). Affinities for Ub_3 and Ub_4 are 1.6 and 4.1 µM close to 1.8 µM obtained for Ub_2. ITC data comparing shorter and longer polyubiquitins may suggest that longer polyubiquitins can form long-range, flanking interactions with NEMO, resulting in a gain of enthalpy and loss of entropy with

respect to shorter ones. SPR confirmed the binding between NEMO and Ub_3 and Ub_4 with equilibrium dissociation constants (K_D) of 9.6 and 6.4 µM for Ub_3 and Ub_4, respectively.

To explain the contradictory observation of the stoichiometry in solution and in crystal and to better understand the molecular recognition between linear polyubiquitins and NEMO, we characterized the dynamics of a $NEMO_{258-350}{:}Ub_2$ complex. An M&M Martini simulation was performed including SAXS data for the complex previously measured (Fig. 6A and fig. S10). The resulting ensemble of structures highlights how binding to NEMO strongly decreases the conformational freedom of linear Ub_2 (Figs. 2 and 6A). Neither the ensemble nor the crystal structures of other bound Ub_2 are located close to the minima of the free Ub_2 ensemble, which has a different distance and orientation between both ubiquitins. Ub_2 residues building up the $NEMO_{258-350}{:}Ub_2$ interface overlap with those involved in the interdomain interactions (fig. S10), in particular residues around the hydrophobic patch I44 and the previously mentioned $E18_2$ surface. The observed interaction sites are in agreement with observed NMR chemical shifts perturbations reported in Vincendeau *et al.* (*23*). The ensemble also provides a possible explanation for the different binding stoichiometry observed in solution ($NEMO{:}Ub_2$ 2:1) and crystalline state ($NEMO{:}Ub_2$ 2:2). A detailed analysis of the NEMO-binding sites indicates that almost all the residues of the NEMO unoccupied site are less exposed to the solvent than those on the occupied one (fig. S10H). While the solvent accessible surface area of the occupied site is 13 nm^2, the one of the unoccupied is 11.9 nm^2, showing that both binding sites are not equal after the binding of Ub_2. Together, these observations indicate that, in solution, binding of linear Ub_2 to $NEMO_{UBAN}$ induces allosteric effects that modulate the overall structure and dynamics of the NEMO dimer. These observations also suggest that the 2:2 highly symmetric binding mode observed in the dense and ordered crystalline state becomes entropically unfavorable to the more flexible and far less dense solution state of the complex.

DISCUSSION

Structural biology investigations on polyubiquitins have mostly focused on diubiquitins, observing that different protein linkages correspond to different protein dynamics leading to different exposed regions for the binding with partners (*48, 50, 51, 53–59*). Ubiquitin signaling has been found associated not only to the linkage type but also to the length of the ubiquitin chains (*8–10*). Here, we first develop an efficient and accurate integrative approach to characterize the conformational ensembles of linear polyubiquitin by combining the Martini coarse-grain force field with SAXS experiments in the framework of metainference. We then use our method to try to rationalize the length-dependent behavior of linear polyubiquitins and the consequence for the interaction with their partner NEMO. Figure 6 rationalizes the observed differences in binding by comparing our free Ub_N ensembles with our NEMO-bound ensemble. The fraction of bound-like configurations in the Ub_2 ensemble is a small fraction of the total ensemble, suggesting a large conformational entropy loss upon binding. This is likely compensated by a release of a large number of water molecules from the binding interfaces upon binding to result in a final entropy gain as indicated by ITC (Fig. 5). The probability of finding at least one diubiquitin pair in a bound-like configuration in the Ub_3 and Ub_4 ensemble increases slightly more than linearly (3.2, 6.7, and 12.0% for Ub_2, Ub_3, and
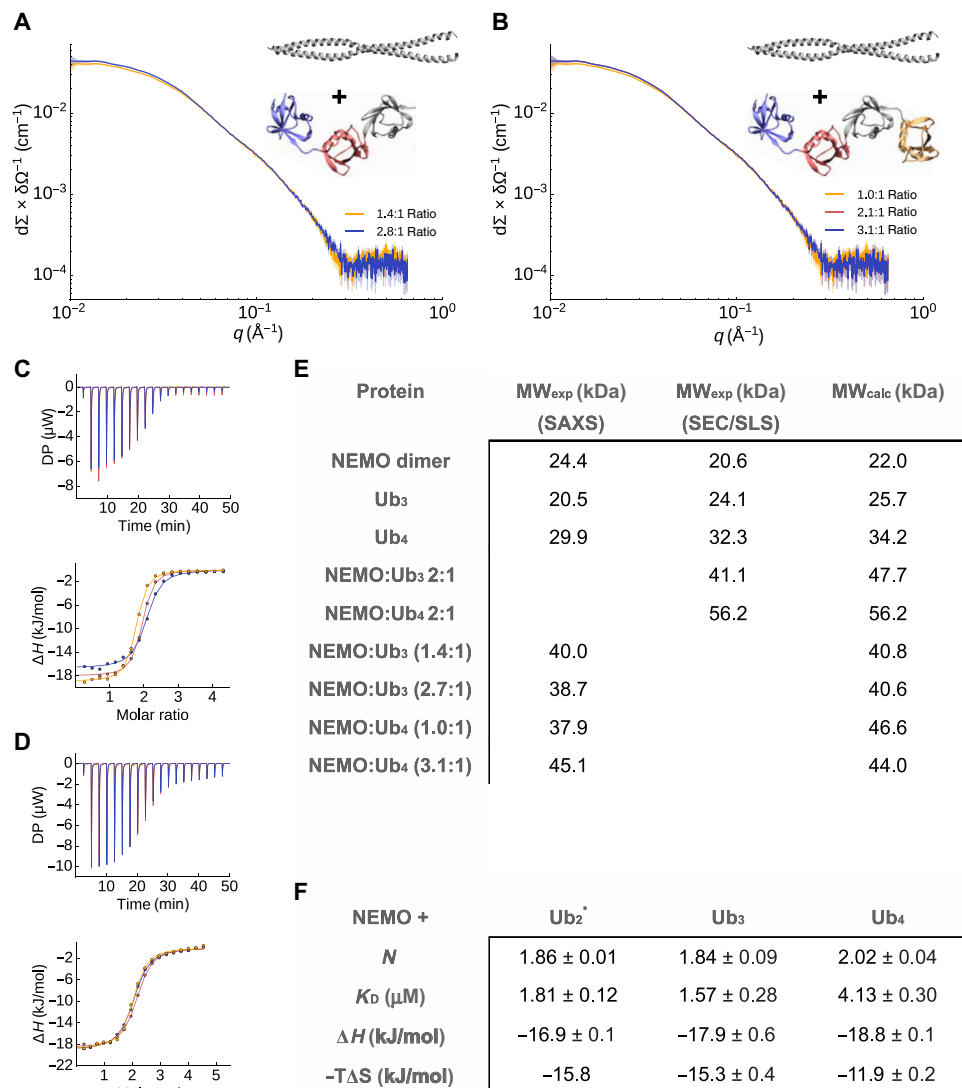
Fig. 5. Effect of chain length on the binding of NEMO. (**A**) and (**B**) SAXS experiments for different ratios of NEMO and Ub_3 (A) and Ub_4 (B). (**C** and **D**) ITC measurement of the interaction of NEMO with Ub_3 (C) and Ub_4 (D). NEMO was titrated into the polyubiquitin solutions. The experiment was repeated three times. DP, differential power. (**E**) MW determination. SAXS and SEC in combination with SLS were used to determine the MW of NEMO, Ub_3, Ub_4, NEMO:Ub_3, and NEMO:Ub_4. The conditions were 50 mM tris HCl (pH 8) and 300 mM NaCl. (**F**) ITC measurement of the NEMO interaction with Ub_2, Ub_3, and Ub_4. NEMO was titrated into the ubiquitin solutions in 50 mM sodium phosphate (pH 7) and 50 mM NaCl. Values are averages ± SEs from three measurements. The individual ITC curves are shown in fig. S7. *Experiments taken from Vincendeau *et al.* (*23*). A stoichiometry of $N = 2$ corresponds to one NEMO dimer binding to one polyubiquitin protein.

(E)

Protein	MW$_{exp}$ (kDa) (SAXS)	MW$_{exp}$ (kDa) (SEC/SLS)	MW$_{calc}$ (kDa)
NEMO dimer	24.4	20.6	22.0
Ub$_3$	20.5	24.1	25.7
Ub$_4$	29.9	32.3	34.2
NEMO:Ub$_3$ 2:1		41.1	47.7
NEMO:Ub$_4$ 2:1		56.2	56.2
NEMO:Ub$_3$ (1.4:1)	40.0		40.8
NEMO:Ub$_3$ (2.7:1)	38.7		40.6
NEMO:Ub$_4$ (1.0:1)	37.9		46.6
NEMO:Ub$_4$ (3.1:1)	45.1		44.0

(F)

NEMO +	Ub$_2$*	Ub$_3$	Ub$_4$
N	1.86 ± 0.01	1.84 ± 0.09	2.02 ± 0.04
K_D (μM)	1.81 ± 0.12	1.57 ± 0.28	4.13 ± 0.30
ΔH (kJ/mol)	−16.9 ± 0.1	−17.9 ± 0.6	−18.8 ± 0.1
$-T\Delta S$ (kJ/mol)	−15.8	−15.3 ± 0.4	−11.9 ± 0.2

Ub_4, respectively), suggesting that longer polyubiquitins are likely to more favorably bind NEMO with respect to shorter ones. The SEC-SLS experiments show that NEMO bound to Ub_4 eluted as a single bound peak in comparison with Ub_3 (table S2 and fig. S9, D and E). Since both NEMO:Ub_4 and NEMO:Ub_3 have similar K_Ds, this can indicate a difference in kinetic stability. To provide a structural interpretation for this hypothesis, we calculated the actual probability of finding the full polyubiquitin in a configuration compatible with the binding (and thus avoiding configurations that would lead to a steric clash with NEMO (Fig. 6, B to D) from our free polyubiquitin ensembles. The probability decreases from Ub_2 to Ub_4 (3.2, 1.5, and 1.4% for Ub_2, Ub_3, and Ub_4, respectively), which can lead to entropy loss. At the same time, nonspecific flanking interactions between polyubiquitin and NEMO far from the binding site can increase the

enthalpy. This is also in agreement with previous measures where, using a longer NEMO construct (NEMO$_{242–419}$) that could provide more surface for interactions, affinities of 3 and 0.3 μM were reported for Ub_2 and Ub_4, respectively (*60*). This principle is also common for intrinsically disordered proteins possibly modulating the lifetime of complexes (*61*). These long-range effects would be less pronounced for a less entropic chain and can play a length-dependent role in the overall interaction.

The probability of finding two diubiquitin pairs in a bound-like configuration is essentially negligible for both Ub_3 and Ub_4 providing a rationale why a 2:1 NEMO$_{258–350}$:Ub_4 interaction is favored by entropy with respect to the 4:1. Making use of our polymer model, we can also speculate that a long-enough polyubiquitin may be able to bind two NEMO dimers with a higher-order (i.e., 4:1) stoichiometry
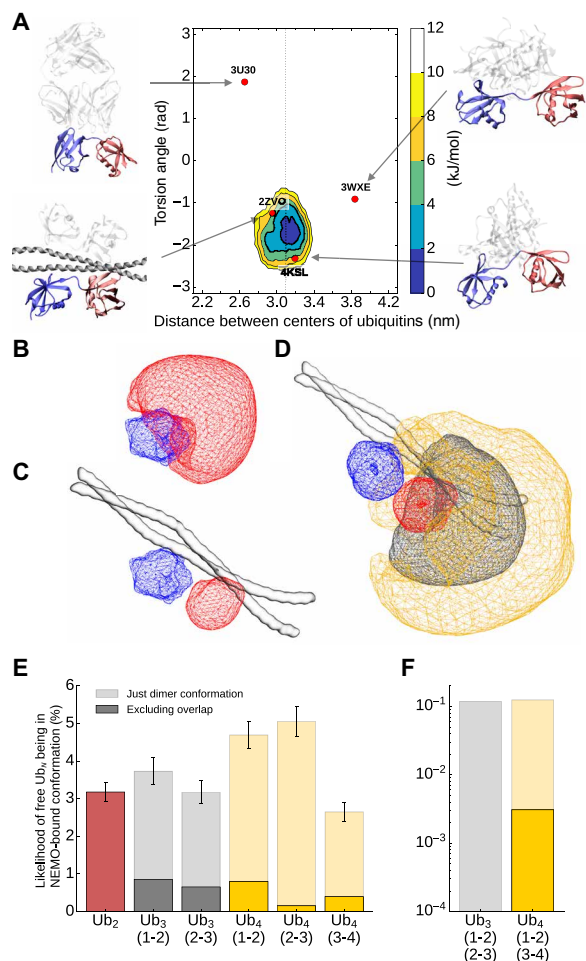
Fig. 6. Comparison between free and NEMO-bound polyubiquitin ensembles. (**A**) Free energy landscapes (in kJ/mol) as a function of the distance between the centers of the two ubiquitin domains and their relative orientation for Ub_2 bound to NEMO. The dots represent the coordinates associated with the available crystal structures with Ub_2 bound to different proteins. (**B** and **C**) Conformational space of free (B) and NEMO-bound (C) ubiquitin pairs in Ub_2. The blue area represents the first ubiquitin, while the red area shows the conformation of the second ubiquitin relative to the first one. (**D**) Conformational space of third (gray area) and forth (orange area) ubiquitin in Ub_4 with the first Ub pair being in a NEMO-bound conformation. (**E** and **F**) Probability of free Ub_2, Ub_3, and Ub_4 of being in a NEMO-bound conformation for one (E) or two (F) NEMO dimers. The transparent bars show the likelihood of the individual pairs being in the NEMO-bound conformation (root mean square deviation <6 Å compared to the average Ub_2 structure in the NEMO-bound simulation). The dark bars show the probability of being in the NEMO-bound conformation, excluding structures with an overlap between the nonbound ubiquitins and NEMO.

with respect to the 2:1 observed for the Ub_2 to Ub_4 range. Given that the end-to-end distance of Ub_2 corresponds to one-half of that of our NEMO construct (5.8 nm for Ub_2 and 11.4 nm for NEMO), and that the two Ub_2 units that bind the two NEMO should be allowed to be flexible, one can estimate the length of this polyubiquitin to be such that $e2e(N-2) \geq 11.4$ nm. This results in a minimum length of $N = 8$ ubiquitins. While this result will not be quantitative when considering a full-length NEMO, it suggests a possible need for these long chains in the assembly of the IKK complex.

Conclusions

The combined use of experiments and MD simulation is a powerful tool to investigate the structure and dynamics of biomolecules and provide a ground for the functional interpretation of protein dynamics. Here, we combined SAXS and CG Martini simulations to accurately and efficiently study the conformational dynamics of linear polyubiquitins and their binding to NEMO. The resulting conformational ensembles allowed us to propose that linear polyubiquitin behave as self-avoiding polymer chains. This might also apply for polyubiquitins in general (with different characteristic lengths). Combining structural studies with multiple biophysical experiments, we provide a systematic assessment of the effect of the polyubiquitin chain length in the molecular recognition of cognate proteins, suggesting that polyubiquitin may modulate the binding with their partners in a length-dependent manner.

METHODS
CG MD simulations

CG MD (CG-MD) simulations were applied to investigate the dynamic of linear di-, tri-, and tetraubiquitin (Ub_2, Ub_3, and Ub_4), as well as Ub_2 with bound NEMO. In total, 11 different simulations have been performed with a total simulation length of 780 µs. An overview of all simulations can be found in table S1.

All CG simulations were run using Gromacs 2016.3 (*62*) and the Martini force field (*34, 35*). In addition, an elastic network model with a force constant of 500 kJ mol^{-1} nm^{-2} was used to conserve the secondary and tertiary structure (*34, 63*). In the case of polyubiquitin, the elastic network inside was only defined for the backbone beads of the core region (from residues 1 to 70, 77 to 146, 153 to 222, or 229 to 298) and not between different domains nor for the linker region. All simulations were performed with periodic boundary conditions, and the systems were solvated with a 0.1 M NaCl solution and run as an isothermal-isobaric (NpT) with a temperature of 300 K and a pressure of 1 bar using 20-fs time steps. To control the temperature and pressure, the v-rescale thermostat (*64*) was used with a coupling constant $\tau_t = 1.0$ ps together with the Parrinello-Rahman barostat (*65*) and a coupling constant of $\tau_p = 20.0$ ps and compressibility of $\chi = 3.0 \times 10^{-4}$ bar^{-1}. Nonbonded interactions were treated with a dielectric constant of 15 and using a cutoff distance of 1.1 nm. Visual Molecular Dynamics (VMD) was used for visualization (*66*).

For many simulations, the Martini 2.2. force field was modified to increase the P-W interaction, which was achieved by giving water beads their own atom type with a 5% larger C6 in interactions with all other atom types, resulting in around 5% higher P-W interactions.

Parallel-biased metadynamics (*37, 67*), as implemented in PLUMED2 (*68*), was used to enhance the sampling of the conformational space, together with the multiple-walker approach (*69*) with 112 replicas for free polyubiquitin or 64 replicas for free NEMO as well as bound NEMO with Ub_2, where each replica had a different starting conformation. The used collective variables were the distances between the centers of the different ubiquitin cores, a torsion angle between the centers of residues 1 to 36 and 37 to 70 of two different ubiquitins, the radius of gyration (calculated only with backbone atoms) and the alphabeta collective variable describing the torsional angles for linkers between the ubiquitin pairs. In total, 4 collective variables for Ub_2, 9 for Ub_3, and 16 for Ub_4 were used. In the case of simulations

with NEMO, additional collective variables were used for the shape of NEMO and distances between Ub_2 and NEMO. The bias factor of the well-tempered metadynamics (70) was set to 10, the frequency for the hill addition was 200 (every 4 ps), the height of Gaussian hills was 0.1 kJ/mol for simulations with Ub_2 and Ub_3, 0.075 kJ/mol for simulations with tetraubiquitin and 0.02 kJ/mol for the simulation with NEMO bound to Ub_2. The flexible Gaussian approach (71) was used to determine the Gaussian width during the simulation.

Metainference (36), a method based on Bayesian inference, was used to integrate experimental SAXS data into simulations and was coupled with metadynamics (M&M) (47). The calculation of the SAXS intensities from a CG Martini representation is implemented in the PLUMED-ISDB module (72, 73) using the parameters derived by Niebling *et al.* (74) and the Debye equation. The SAXS data of the different systems were fitted with a 16th-degree polynomial to calculate points used for restraints. Twenty-one equidistant points for q between 0.017 and 0.24 nm^{-1} were used for Ub_2 and Ub_3 and for free and bound NEMO, 19 points for q between 0.025 and 0.19 nm^{-1} for Ub_4. The rage depends on the quality of the experimental data. An initial scaling value was determent by comparing the calculated and experimental SAXS intensities for the lowest q value. Metainference was used with the outlier noise model (36) for each data point, and the restraints were applied every fifth step. A scaling faction and offset for the experimental data were sampled using a flat prior between 0.9 and 1.1 or −1 and 1. The error for calculating an average quantity σ_{mean} was determined automatically (75) from the maximum SE over 2 ps of simulations.

Six different simulations with Ub_2 were performed using: Martini 2.2 with metadynamics, with increased P-W interaction and metadynamics, with M&M, with M&M applied every step, and with Martini 3 beta. Notably, Martini 3 beta was not stable with metadynamics; therefore, 112 replicas were run on a longer time scale. The SAXS data of Ub_2 were taken from Vincendeau *et al.* (23). Notably, the profile was measured with Ub_2 containing a His-tag on the N terminus that was also modeled. Ub_3 and Ub_4 simulations were run with increased P-W interaction, metadynamics with and without metainference, and SAXS. All polyubiquitin simulations were run for at least 500 ns per replica. The simulations with free NEMO and NEMO bound to Ub_2 were performed with increased P-W interaction and M&M with SAXS for at least 100 ns per replica. The SAXS data of NEMO bound to Ub_2 were taken from Vincendeau *et al.* (23).

Five simulations were performed with a K63 Ub_2 construct, using Martini 2.2 with +5% increased P-W interaction and metadynamics and with +5% increased P-W interactions and M&M, as well as two additional tests with +10% increased P-W interactions and a weaker elastic network with a force constant of 500 kJ mol^{-1} nm^{-2}, using the same SAXS data and constructs as described in Paissoni *et al.* (48). In addition, two control simulations with K63 Ub_2 were performed using a Martini bead–based excluding volume. The Martini nonbonded interactions were replaced with repulsion term. The plumed input files, as well as the modified Martini topology files, are deposited in PLUMED-NEST (76) as plumID:20.009.

Protein expression and purification
Human $NEMO_{258-350}$ C347S was expressed and purified as described in Vincendeau *et al.* (23). Protein concentration was determined by measuring the absorbance at 205 nm using specific absorbance for $NEMO_{258-350}$ C347S of 300,990 M^{-1} cm^{-1}, respectively (77).

The constructs for the expression of Ub_3 and Ub_4 were a gift of P. Elliott and D. Komander [Medical Research Council (MRC) Laboratory of Molecular Biology, Cambridge, UK]. The constructs were transformed into *Escherichia coli* strain BL21 (DE3) and cultured at 20°C in 2-liter flasks containing 500 ml of ZYM 5052 autoinduction medium (78) and carbenicillin (100 μg/ml). Cells were harvested by centrifugation after reaching saturation, resuspended in 60-ml lysis buffer [50 mM tris-HCl, 300 mM NaCl, 10 mM $MgCl_2$, deoxyribonuclease I (10 μg/ml), 1 mM AEBSF.HCl (4-(2-Aminoethyl) benzenesulfonyl fluoride hydrochloride), 0.2% (v/v) NP-40, and lysozyme (1 mg/ml; pH 8.0)], and lysed by sonication. The lysate was clarified by centrifugation (40,000g) and filtration (0.2 μM). The supernatant was heated in a water bath for 10 to 15 min at 60°C and the precipitate removed by centrifugation. The supernatant was dialyzed overnight against 2 liters of buffer A [50 mM sodium acetate (pH 4.5)], clarified by centrifugation and applied to a 5-ml HiTrap SP HP column (GE Healthcare), and equilibrated in buffer A. Bound proteins were eluted using a linear gradient (10 column volumes) from 0 to 1 M NaCl in buffer A using an Äkta Purifier (GE Healthcare). Elution fractions (1.6 ml) were collected in wells containing 250 μl of 1 M tris-HCl (pH 9.0). Fractions containing Ub_3 or Ub_4 were pooled, concentrated and applied to a HiLoad 16/600 Superdex 75 column (GE Healthcare), and equilibrated in buffer B [50 mM tris-HCl and 100 mM NaCl (pH 7.4)]. The main elution peak containing Ub_3 or Ub_4 was collected and concentrated to approximately 3 to 6 mg/ml, flash-frozen, and stored at −80°C. Protein concentrations were determined by measuring the absorbance at 205 nm using specific absorbance for Ub_3 and Ub_4 of 747,790 and 997,980 M^{-1} cm^{-1}, respectively (77). The Ub_4 concentration values used on interaction studies of Ub_4 with NEMO were corrected by 30% on the basis of the SEC with SLS results.

Small-angle x-ray scattering measurements
SAXS measurements were performed on a Rigaku BioSAXS-1000 instrument with an HF007 microfocus generator equipped with a Cu target at 40 kV and 30 mA. Transmissions were measured with a photodiode beamstop, and q calibration was made by an Ag-behenate measurement. Absolute calibration was done with calibrated glassy carbon (79). Measurements were done in four 900-s frames, which were averaged. Under these conditions, no radiation damage was detected. Circular averaging and background subtraction were done with the Rigaku SAXSLab software v 3.0.1r1.

Radii of gyration were calculated with the ATSAS package v 2.8.0 (80). Fits for the MW determination were made in Origin v 9. SAXS measurements were made at 293 K using a buffer containing 300 mM NaCl and 50 mM tris-HCl at pH 8.0. Experiments on the free proteins were performed at the following concentrations (fig. S3, I and J): NEMO at 2.34, 4.62, and 7.72 mg/ml; Ub_3 at 3.41, 6.72, and 11.17 mg/ml; and Ub_4 at 4.5, 9.1 and 15.1 mg/ml. Experiments with NEMO and Ub_3 and Ub_4 at different ratios were performed at two concentrations (between 3 and 4 and 7 and 8 mg/ml) at the following ratios: $NEMO:Ub_3$ at 1.4:1 and 2.7:1 ratios and $NEMO:Ub_4$ at 1.0:1, 2.1:1, and 3.1:1 ratios. No concentration-dependent effects were detected.

Isothermal titration calorimetry
ITC measurements were carried out at 298 K using a PEAQ-ITC titration microcalorimeter (MicroCal, Malvern). The NEMO-to-Ub_3 calorimetric titration consisted of 19 injections of 2 μl of a 2.13 mM NEMO solution, into the reaction cell containing 300 μl of 94.71 μM

Ub$_3$, at a stirring speed of 750 rpm. The NEMO-to-Ub$_4$ calorimetric titration consisted of 19 injections of 2 μl of a 2.84 mM NEMO solution, into the reaction cell containing 300 μl of 120.6 μM Ub$_4$, at a stirring speed of 750 rpm. Sample conditions were 50 mM sodium phosphate (pH 7.0) and 50 mM NaCl. The heat of dilution was obtained by titrating NEMO into the sample cell containing only buffer. Experiments were done in triplicate. The ITC data were analyzed using the software MicroCal PEAQ-ITC analysis software. Parameters are presented as averages ± SEs.

SEC with SLS

SLS experiments were performed of NEMO mutant (C347S) in complex with tri- and tetraubiquitin at 30°C using a Viscotek TDA 305 triple array detector (Malvern Instruments) downstream to an Äkta Purifier (GE Healthcare) equipped with an analytical size exclusion column (Superdex 200 10/300 GL, GE Healthcare) at 4°C. The samples were run at approximately 8 mg/ml at a flow rate of 0.5 ml/min. The experiments were performed using a tris buffer [50 mM tris-HCl and 300 mM NaCl (pH 8.0)] and a phosphate buffer [50 mM sodium phosphate and 50 mM NaCl (pH 7.0)]. The molecular masses of the samples were calculated from the refractive index and right-angle light-scattering signals using Omnisec (Malvern Instruments). The SLS detector was calibrated with a bovine serum albumin (BSA) solution (4 mg/ml) using 66.4 kDa for the BSA monomer and a dn/dc value of 0.185 ml/g for all protein samples.

SPR measurements

SPR measurements were performed at 25°C using a Pioneer FE instrument (FortéBio, Molecular Devices). Ub$_3$ and Ub$_4$ were covalently immobilized onto two different flow cell channels on a biosensor chip by amine coupling to 456 and 721 response unit, respectively, using a 10 mM NaOAc (pH 5) immobilization buffer. NEMO was injected in a twofold concentration series over immobilized ubiquitins at a flow rate of 30 μl/min using a phosphate-buffered saline running buffer [50 mM sodium phosphate, 50 mM NaCl, and 0.005% Tween 20 (pH 7)]. The data were analyzed using Qdat Data Analysis Tool version 2.6.3.0 (FortéBio). The sensorgrams were corrected for buffer effects and unspecific binding to the chip matrix by subtraction of blank and reference surface (a blank flow cell channel activated by injection of EDC/NHS (N-ethyl-N′-(3-dimethylaminopropyl)carbodiimide hydrochloride/N-hydroxysuccinimide) and inactivated by injection of ethanolamine). The equilibrium dissociation constants (K_D) were estimated by plotting responses at equilibrium (Req) against the injected concentration and curve fitted to a Langmuir (1:1) binding isotherm.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/6/42/eabc3786/DC1

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. K. Haglund, I. Dikic, Ubiquitylation and cell signaling. *EMBO J.* **24**, 3353–3359 (2005).
2. R. J. Deshaies, C. A. Joazeiro, RING domain E3 ubiquitin ligases. *Annu. Rev. Biochem.* **78**, 399–434 (2009).
3. W. Li, Y. Ye, Polyubiquitin chains: Functions, structures, and mechanisms. *Cell. Mol. Life Sci.* **65**, 2397–2406 (2008).
4. J. Peng, D. Schwartz, J. E. Elias, C. C. Thoreen, D. Cheng, G. Marsischky, J. Roelofs, D. Finley, S. P. Gygi, A proteomics approach to understanding protein ubiquitination. *Nat. Biotechnol.* **21**, 921–926 (2003).
5. T. Kirisako, K. Kamei, S. Murata, M. Kato, H. Fukumoto, M. Kanie, S. Sano, F. Tokunaga, K. Tanaka, K. Iwai, A ubiquitin ligase complex assembles linear poly-ubiquitin chains. *EMBO J.* **25**, 4877–4887 (2006).
6. K. Iwai, F. Tokunaga, Linear polyubiquitination: A new regulator of NF-kappaB activation. *EMBO Rep.* **10**, 706–713 (2009).
7. K. Iwai, H. Fujita, Y. Sasaki, Linear ubiquitin chains: NF-kappaB signalling, cell death and beyond. *Nat. Rev. Mol. Cell Biol.* **15**, 503–508 (2014).
8. D. Komander, M. Rape, The ubiquitin code. *Annu. Rev. Biochem.* **81**, 203–229 (2012).
9. S. Zhao, H. D. Ulrich, Distinct consequences of posttranslational modification by linear versus K63-linked poly-ubiquitin chains. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 7704–7709 (2010).
10. D. Fushman, K. D. Wilkinson, Structure and recognition of poly-ubiquitin chains of different lengths and linkage. *F1000 Biol. Rep.* **3**, 26 (2011).
11. M. Akutsu, I. Dikic, A. Bremm, Ubiquitin chain diversity at a glance. *J. Cell Sci.* **129**, 875–880 (2016).
12. J. S. Thrower, L. Hoffman, M. Rechsteiner, C. M. Pickart, Recognition of the poly-ubiquitin proteolytic signal. *EMBO J.* **19**, 94–102 (2000).
13. H. Tsuchiya, F. Ohtake, N. Arai, A. Kaiho, S. Yasuda, K. Tanaka, Y. Saeki, In vivo ubiquitin linkage-type analysis reveals that the Cdc48-Rad23/Dsk2 axis contributes to K48-linked chain specificity of the proteasome. *Mol. Cell* **66**, 488–502.e7 (2017).
14. M. von Delbruck, A. Kniss, V. V. Rogov, L. Pluska, K. Bagola, F. Lohr, P. Guntert, T. Sommer, V. Dötsch, The CUE domain of Cue1 aligns growing ubiquitin chains with Ubc7 for rapid elongation. *Mol. Cell* **62**, 918–928 (2016).
15. N. W. Pierce, G. Kleiger, S. O. Shan, R. J. Deshaies, Detection of sequential polyubiquitylation on a millisecond timescale. *Nature* **462**, 615–619 (2009).
16. J. Kovacev, K. Wu, D. E. Spratt, R. A. Chong, C. Lee, J. Nayak, G. S. Shaw, Z. Q. Pan, A snapshot of ubiquitin chain elongation: Lysine 48-tetra-ubiquitin slows down ubiquitination. *J. Biol. Chem.* **289**, 7068–7081 (2014).
17. B. Gerlach, S. M. Cordier, A. C. Schmukle, C. H. Emmerich, E. Rieser, T. L. Haas, A. I. Webb, J. A. Rickard, H. Anderton, W. W. Wong, U. Nachbur, L. Gangoda, U. Warnken, A. W. Purcell, J. Silke, H. Walczak, Linear ubiquitination prevents inflammation and regulates immune signalling. *Nature* **471**, 591–596 (2011).
18. I. E. Wertz, K. Newton, D. Seshasayee, S. Kusam, C. Lam, J. Zhang, N. Popovych, E. Helgason, A. Schoeffler, S. Jeet, N. Ramamoorthi, L. Kategaya, R. J. Newman, K. Horikawa, D. Dugger, W. Sandoval, S. Mukund, A. Zindal, F. Martin, C. Quan, J. Tom, W. J. Fairbrother, S. Townsend, S. Warming, J. DeVoss, J. Liu, E. Dueber, P. Caplazi, W. P. Lee, C. C. Goodnow, M. Balazs, K. Yu, G. Kolumam, V. M. Dixit, Phosphorylation and linear ubiquitin direct A20 inhibition of inflammation. *Nature* **528**, 370–375 (2015).
19. D. Senft, J. Qi, Z. A. Ronai, Ubiquitin ligases in oncogenic transformation and cancer therapy. *Nat. Rev. Cancer* **18**, 69–88 (2018).
20. D. A. Catici, J. E. Horne, G. E. Cooper, C. R. Pudney, Polyubiquitin drives the molecular interactions of the NF-kappaB essential modulator (NEMO) by allosteric regulation. *J. Biol. Chem.* **290**, 14130–14139 (2015).
21. A. V. Hauenstein, G. Xu, V. Kabaleeswaran, H. Wu, Evidence for M1-Linked polyubiquitin-mediated conformational change in NEMO. *J. Mol. Biol.* **429**, 3793–3800 (2017).
22. M. Hinz, C. Scheidereit, The IκB kinase complex in NF-κB regulation and beyond. *EMBO Rep.* **15**, 46–61 (2014).
23. M. Vincendeau, K. Hadian, A. C. Messias, J. K. Brenke, J. Halander, R. Griesbach, U. Greczmiel, A. Bertossi, R. Stehle, D. Nagel, K. Demski, H. Velvarska, D. Niessing, A. Geerlof, M. Sattler, D. Krappmann, Inhibition of canonical NF-kappaB signaling by a small molecule targeting NEMO-ubiquitin interaction. *Sci. Rep.* **6**, 18934 (2016).
24. S. Rahighi, F. Ikeda, M. Kawasaki, M. Akutsu, N. Suzuki, R. Kato, T. Kensche, T. Uejima, S. Bloor, D. Komander, F. Randow, S. Wakatsuki, I. Dikic, Specific recognition of linear ubiquitin chains by NEMO is important for NF-kappaB activation. *Cell* **136**, 1098–1109 (2009).
25. F. J. Ivins, M. G. Montgomery, S. J. Smith, A. C. Morris-Davies, I. A. Taylor, K. Rittinger, NEMO oligomerization and its ubiquitin-binding properties. *Biochem. J.* **421**, 243–251 (2009).
26. S. Polley, D. B. Huang, A. V. Hauenstein, A. J. Fusco, X. Zhong, D. Vu, B. Schrofelbauer, Y. Kim, A. Hoffmann, I. M. Verma, G. Ghosh, T. Huxford, A structural basis for IkappaB kinase 2 activation via oligomerization-dependent trans auto-phosphorylation. *PLOS Biol.* **11**, e1001581 (2013).
27. J. A. DiDonato, M. Hayakawa, D. M. Rothwarf, E. Zandi, M. Karin, A cytokine-responsive IkappaB kinase that activates the transcription factor NF-kappaB. *Nature* **388**, 548–554 (1997).
28. T. T. Thach, D. Shin, S. Han, S. Lee, New conformations of linear poly-ubiquitin chains from crystallographic and solution-scattering studies expand the conformational space of poly-ubiquitin. *Acta Crystallogr. D Struct. Biol.* **72**, 524–535 (2016).
29. L. Boldon, F. Laliberte, L. Liu, Review of the fundamental theories behind small angle x-ray scattering, molecular dynamics simulations, and relevant integrated application. *Nanotechnol. Rev.* **6**, 25661 (2015).

30. D. M. Zuckerman, Equilibrium sampling in biomolecular simulations. *Annu. Rev. Biophys.* **40**, 41–62 (2011).

31. C. Camilloni, F. Pietrucci, Advanced simulation techniques for the thermodynamic and kinetic characterization of biological systems. *Adv. Phys. X* **3**, 1477531 (2018).

32. S. Kmiecik, D. Gront, M. Kolinski, L. Wieteska, A. E. Dawid, A. Kolinski, Coarse-grained protein models and their applications. *Chem. Rev.* **116**, 7898–7936 (2016).

33. M. Bonomi, G. T. Heller, C. Camilloni, M. Vendruscolo, Principles of protein structural ensemble determination. *Curr. Opin. Struct. Biol.* **42**, 106–116 (2017).

34. S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, A. H. de Vries, The MARTINI force field: Coarse grained model for biomolecular simulations. *J. Phys. Chem. B* **111**, 7812–7824 (2007).

35. D. H. de Jong, G. Singh, W. F. Bennett, C. Arnarez, T. A. Wassenaar, L. V. Schafer, X. Periole, D. P. Tieleman, S. J. Marrink, Improved parameters for the Martini coarse-grained protein force field. *J. Chem. Theory Comput.* **9**, 687–697 (2013).

36. M. Bonomi, C. Camilloni, A. Cavalli, M. Vendruscolo, Metainference: A Bayesian inference method for heterogeneous systems. *Sci. Adv.* **2**, e1501177 (2016).

37. A. Laio, M. Parrinello, Escaping free-energy minima. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 12562–12566 (2002).

38. D. Komander, F. Reyes-Turcu, J. D. Licchesi, P. Odenwaelder, K. D. Wilkinson, D. Barford, Molecular discrimination of structurally equivalent Lys 63-linked and linear poly-ubiquitin chains. *EMBO Rep.* **10**, 466–473 (2009).

39. A. Rohaim, M. Kawasaki, R. Kato, I. Dikic, S. Wakatsuki, Structure of a compact conformation of linear diubiquitin. *Acta Crystallogr. D Biol. Crystallogr.* **68**, 102–108 (2012).

40. A. C. Stark, C. T. Andrews, A. H. Elcock, Toward optimized potential functions for protein-protein interactions in aqueous solutions: Osmotic second virial coefficient calculations using the MARTINI coarse-grained force field. *J. Chem. Theory Comput.* **9**, 4176–4185 (2013).

41. M. Javanainen, H. Martinez-Seara, I. Vattulainen, Excessive aggregation of membrane proteins in the Martini model. *PLOS ONE* **12**, e0187936 (2017).

42. A. Berg, O. Kukharenko, M. Scheffner, C. Peter, Towards a molecular basis of ubiquitin signaling: A dual-scale simulation study of ubiquitin dimers. *PLOS Comput. Biol.* **14**, e1006589 (2018).

43. A. H. Larsen, Y. Wang, S. Bottaro, S. Grudinin, L. Arleth, K. Lindorff-Larsen, Combining molecular dynamics simulations with small-angle x-ray and neutron scattering data to study multi-domain proteins in solution. *PLoS Comput. Biol.* **16**, e1007870 (2020).

44. S. O. Yesylevskyy, L. V. Schafer, D. Sengupta, S. J. Marrink, Polarizable water model for the coarse-grained MARTINI force field. *PLOS Comput. Biol.* **6**, e1000810 (2010).

45. A. W. Gotz, D. Bucher, S. Lindert, J. A. McCammon, Dipeptide aggregation in aqueous solution from fixed point-charge force fields. *J. Chem. Theory Comput.* **10**, 1631–1637 (2014).

46. P. Robustelli, S. Piana, D. E. Shaw, Developing a molecular dynamics force field for both folded and disordered protein states. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E4758–E4766 (2018).

47. M. Bonomi, C. Camilloni, M. Vendruscolo, Metadynamic metainference: Enhanced sampling of the metainference ensemble using metadynamics. *Sci. Rep.* **6**, 31232 (2016).

48. C. Paissoni, A. Jussupow, C. Camilloni, Determination of protein structural ensembles by hybrid-resolution SAXS restrained molecular dynamics. *J. Chem. Theory Comput.* **16**, 2825–2834 (2020).

49. S. M. Bhattacharjee, A. Giacometti, A. Maritan, Flory theory for polymers. *J. Phys. Condens. Matter* **25**, 503101 (2013).

50. C. A. Castaneda, T. R. Kashyap, M. A. Nakasone, S. Krueger, D. Fushman, Unique structural, dynamical, and functional properties of k11-linked poly-ubiquitin chains. *Structure* **21**, 1168–1181 (2013).

51. K. Berlin, C. A. Castaneda, D. Schneidman-Duhovny, A. Sali, A. Nava-Tudela, D. Fushman, Recovering a representative conformational ensemble from underdetermined macromolecular structural data. *J. Am. Chem. Soc.* **135**, 16595–16609 (2013).

52. S. M. Lin, S. C. Lin, J. Y. Hong, T. W. Su, B. J. Kuo, W. H. Chang, Y. F. Tu, Y. C. Lo, Structural insights into linear tri-ubiquitin recognition by A20-binding inhibitor of NF-kappaB, ABIN-2. *Structure* **25**, 66–78 (2017).

53. Z. Liu, C. Tang, Ensemble structure description of Lys63-linked diubiquitin. *Data Brief* **7**, 81–88 (2016).

54. Z. Liu, Z. Gong, W. X. Jiang, J. Yang, W. K. Zhu, D. C. Guo, W. P. Zhang, M. L. Liu, C. Tang, Lys63-linked ubiquitin chain adopts multiple conformational states for specific target recognition. *eLife* **4**, e05767 (2015).

55. Y. Wang, C. Tang, E. Wang, J. Wang, PolyUbiquitin chain linkage topology selects the functions from the underlying binding landscape. *PLOS Comput. Biol.* **10**, e1003691 (2014).

56. Y. Ye, G. Blaser, M. H. Horrocks, M. J. Ruedas-Rama, S. Ibrahim, A. A. Zhukov, A. Orte, D. Klenerman, S. E. Jackson, D. Komander, Ubiquitin chain conformation regulates recognition and activity of interacting proteins. *Nature* **492**, 266–270 (2012).

57. A. Bremm, S. M. Freund, D. Komander, Lys11-linked ubiquitin chains adopt compact conformations and are preferentially hydrolyzed by the deubiquitinase Cezanne. *Nat. Struct. Mol. Biol.* **17**, 939–947 (2010).

58. T. Tenno, K. Fujiwara, H. Tochio, K. Iwai, E. H. Morita, H. Hayashi, S. Murata, H. Hiroaki, M. Sato, K. Tanaka, M. Shirakawa, Structural basis for distinct roles of Lys63- and Lys48-linked poly-ubiquitin chains. *Genes Cells* **9**, 865–875 (2004).

59. A. Berg, L. Franke, M. Scheffner, C. Peter, Machine learning driven analysis of large scale simulations reveals conformational characteristics of ubiquitin chains. *J. Chem. Theory Comput.* **16**, 3205–3220 (2020).

60. K. Hadian, R. A. Griesbach, S. Dornauer, T. M. Wanger, D. Nagel, M. Metlitzky, W. Beisker, M. Schmidt-Supprian, D. Krappmann, NF-κB essential modulator (NEMO) interaction with linear and lys-63 ubiquitin chains contributes to NF-κB activation. *J. Biol. Chem.* **286**, 26107–26117 (2011).

61. M. D. Crabtree, W. Borcherds, A. Poosapati, S. L. Shammas, G. W. Daughdrill, J. Clarke, Conserved helix-flanking prolines modulate intrinsically disordered protein: Target affinity by altering the lifetime of the bound complex. *Biochemistry* **56**, 2379–2384 (2017).

62. M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, E. Lindahl, GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1-2**, 19–25 (2015).

63. X. Periole, M. Cavalli, S. J. Marrink, M. A. Ceruso, Combining an elastic network with a coarse-grained molecular force field: Structure, dynamics, and intermolecular recognition. *J. Chem. Theory Comput.* **5**, 2531–2543 (2009).

64. G. Bussi, D. Donadio, M. Parrinello, Canonical sampling through velocity rescaling. *J. Chem. Phys.* **126**, 014101 (2007).

65. M. Parrinello, A. Rahman, Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **52**, 7182–7190 (1981).

66. W. Humphrey, A. Dalke, K. Schulten, VMD: Visual molecular dynamics. *J. Mol. Graph.* **14**, 33–38 (1996).

67. J. Pfaendtner, M. Bonomi, Efficient sampling of high-dimensional free-energy landscapes with parallel Bias Metadynamics. *J. Chem. Theory Comput.* **11**, 5062–5067 (2015).

68. G. A. Tribello, M. Bonomi, D. Branduardi, C. Camilloni, G. Bussi, PLUMED 2: New feathers for an old bird. *Comput. Phys. Commun.* **185**, 604–613 (2014).

69. P. Raiteri, A. Laio, F. L. Gervasio, C. Micheletti, M. Parrinello, Efficient reconstruction of complex free energy landscapes by multiple walkers metadynamics. *J. Phys. Chem. B* **110**, 3533–3539 (2006).

70. A. Barducci, G. Bussi, M. Parrinello, Well-tempered metadynamics: a smoothly converging and tunable free-energy method. *Phys. Rev. Lett.* **100**, 020603 (2008).

71. D. Branduardi, G. Bussi, M. Parrinello, Metadynamics with adaptive Gaussians. *J. Chem. Theory Comput.* **8**, 2247–2254 (2012).

72. M. Bonomi, C. Camilloni, Integrative structural and dynamical biology with PLUMED-ISDB. *Bioinformatics* **33**, 3999–4000 (2017).

73. C. Paissoni, A. Jussupow, C. Camilloni, Martini bead form factors for nucleic acids and their application in the refinement of protein–nucleic acid complexes against SAXS data. *J. Appl. Crystallogr.* **52**, 394–402 (2019).

74. S. Niebling, A. Bjorling, S. Westenhoff, MARTINI bead form factors for the analysis of time-resolved x-ray scattering of proteins. *J. Appl. Crystallogr.* **47**, 1190–1198 (2014).

75. T. Löhr, A. Jussupow, C. Camilloni, Metadynamic metainference: Convergence towards force field independent structural ensembles of a disordered peptide. *J. Chem. Phys.* **146**, 165102 (2017).

76. The PLUMED consortium, Promoting transparency and reproducibility in enhanced molecular simulations. *Nat. Methods* **16**, 670–673 (2019).

77. N. J. Anthis, G. M. Clore, Sequence-specific determination of protein and peptide concentrations by absorbance at 205 nm. *Protein Sci.* **22**, 851–858 (2013).

78. F. W. Studier, Protein production by auto-induction in high-density shaking cultures. *Protein Expr. Purif.* **41**, 207–234 (2005).

79. L. Fan, M. Degen, S. Bendle, N. Grupido, J. Ilavsky, The absolute calibration of a small-angle scattering instrument with a laboratory x-ray source. *J. Phys. Conf. Ser.* **247**, 012005 (2010).

80. M. V. Petoukhov, D. Franke, A. V. Shkumatov, G. Tria, A. G. Kikhney, M. Gajda, C. Gorba, H. D. Mertens, P. V. Konarev, D. I. Svergun, New developments in the ATSAS program package for small-angle scattering data analysis. *J. Appl. Crystallogr.* **45**, 342–350 (2012).

Citation: A. Jussupow, A. C. Messias, R. Stehle, A. Geerlof, S. M. Ø. Solbak, C. Paissoni, A. Bach, M. Sattler, C. Camilloni, The dynamics of linear polyubiquitin. *Sci. Adv.* **6**, eabc3786 (2020).

The dynamics of linear polyubiquitin

Alexander Jussupow, Ana C. Messias, Ralf Stehle, Arie Geerlof, Sara M. Ø. Solbak, Cristina Paissoni, Anders Bach, Michael Sattler and Carlo Camilloni

ARTICLE TOOLS	http://advances.sciencemag.org/content/6/42/eabc3786
SUPPLEMENTARY MATERIALS	http://advances.sciencemag.org/content/suppl/2020/10/09/6.42.eabc3786.DC1
REFERENCES	This article cites 80 articles, 15 of which you can access for free http://advances.sciencemag.org/content/6/42/eabc3786#BIBL
PERMISSIONS	http://www.sciencemag.org/help/reprints-and-permissions

Use of this article is subject to the Terms of Service

Martini bead form factors for nucleic acids and their application in the refinement of protein–nucleic acid complexes against SAXS data

Cristina Paissoni,[a]* Alexander Jussupow[b] and Carlo Camilloni[a]*

[a]Dipartimento di Bioscienze, Università degli Studi di Milano, Via Celoria 26, Milano 20133, Italy, and [b]Department of Chemistry and Institute of Advanced Study, Technical University of Munich, Garching 85747, Germany.
*Correspondence e-mail: cristina.paissoni@unimi.it, carlo.camilloni@unimi.it

The use of small-angle X-ray scattering (SAXS) in combination with molecular dynamics simulation is hampered by its heavy computational cost. The calculation of SAXS from atomic structures can be speeded up by using a coarse-grain representation of the structure. Following the work of Niebling, Björling & Westenhoff [*J. Appl. Cryst.* (2014), **47**, 1190–1198], the Martini bead form factors for nucleic acids have been derived and then implemented, together with those previously determined for proteins, in the publicly available *PLUMED* library. A hybrid multi-resolution strategy has also been implemented to perform SAXS restrained simulations at atomic resolution by calculating the virtual positions of the Martini beads on the fly and using them for the calculation of SAXS. The accuracy and efficiency of the method are demonstrated by refining the structure of two protein–nucleic acid complexes. Instrumental for this result is the use of metainference, which allows the consideration and alleviation of the approximations at play in the present SAXS calculations.

1. Introduction

Small-angle X-ray scattering (SAXS) is a powerful structural technique to study biomolecules in a solution environment. Even if it does not reach atomic resolution, SAXS can complement and be integrated with other structural techniques providing information on the size, shape, global dynamics and intermolecular interactions of a system (Tuukkanen *et al.*, 2017; Koch *et al.*, 2003). Furthermore, time-resolved SAXS measures can be employed to study conformational changes over multiple time scales (Levantino *et al.*, 2015). Crystal structures can provide a detailed interpretation of the measured SAXS curves. If large discrepancies between experimental and theoretical scattering intensities are observed then these may indicate differences between the crystal and solution state of the system. In this context, a more dynamic view of the system and methods of exploring possible conformations consistent with experimental data are needed. To tackle this problem multiple strategies have been implemented, from fast techniques based on normal model analysis (Gorba *et al.*, 2008; Panjkovich & Svergun, 2016; Zheng & Tekpinar, 2011) to more accurate, but time-consuming, approaches based on the combination of experimental data and molecular dynamics (MD) simulations (Hub, 2018).

From a computational perspective, calculating SAXS given a structure of N atoms is an $O(N^2)$ problem. In comparison, the calculation of NMR observables like chemical shifts, ^3J couplings or residual dipolar couplings requires only a few

$\chi^2 = 2.65$

Intensity

10^0
10^{-1}
10^{-2}

Residuals

10
0
−10

0.2   0.4

q (Å^{-1})
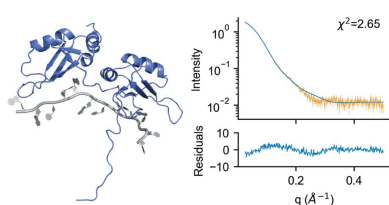
atoms (Schwieters *et al.*, 2003). This places some limitation on the use of SAXS as a restraint or as a scoring function for large systems and large numbers of conformers.

Multiple strategies have been adopted to calculate SAXS efficiently and reduce the complexity of the problem. For example, *CRYSOL* adopted a spherical harmonics expansion (Svergun *et al.*, 1995), which reduces the complexity from $O(N^2)$ to $O(q^2D^2N)$, with D indicating the maximum dimension of the sample and q the magnitude of the scattering vector. Other approaches include hierarchical algorithms (Berlin *et al.*, 2014) displaying $O[(qD)^3\log(qD) + N\log(N)]$ complexity, or the $O[N\log(N)]$ particle mesh Ewald summation (Marchi, 2016). Alternatively, a possible strategy consists of adopting the standard Debye summation but coarse-graining the structure representation using M beads, each comprising a variable number of atoms (Yang *et al.*, 2009; Ravikumar *et al.*, 2013; Stovgaard *et al.*, 2010; Zheng & Tekpinar, 2011; Niebling *et al.*, 2014), resulting in an $O(M^2)$ complexity with $M < N$. Recently, Niebling *et al.* (2014) derived the Martini bead form factors for proteins, making use of the single-bead approximation (SBA) (Yang *et al.*, 2009), and showed how this approach can be almost 50 times faster than SAXS calculation performed with Debye summation and standard atomistic form factors.

Here, we first build on the work of Niebling *et al.* by deriving the Martini bead (Uusitalo *et al.*, 2015, 2017; Marrink & Tieleman, 2013) form factors for nucleic acids (DNA and RNA), and then we implement a hybrid multi-resolution strategy where the positions of the Martini beads are calculated on the fly in a full atomistic MD simulation and employed in combination with the above-mentioned form factors to calculate SAXS. We demonstrate the accuracy and efficiency of this approach by refining the structures of a protein–DNA and a protein–RNA complex using the measured SAXS as a restraint. Importantly, in our strategy, the strength of the restraint is determined by metainference, a Bayesian inference approach that allows the consideration of multiple sources of error, in such a way that the approximations at play are considered (Bonomi *et al.*, 2016). The presented approach, including the form factors for Martini beads of proteins and nucleic acids, is implemented in the *PLUMED-ISDB* module (Bonomi & Camilloni, 2017) of the *PLUMED* library (Tribello *et al.*, 2014), making it readily available to all the codes compatible with *PLUMED* as well as a post-processing tool.

2. Theory and methods

2.1. Computing scattering intensities of biomolecules in solution

Given a molecule of N atoms, the total scattering amplitude at vector **q** is described by

$$A(\mathbf{q}) = \sum_{i=1}^{N} f_i(q) \exp\left(i\mathbf{q} \cdot \mathbf{r}_i\right), \tag{1}$$

where $\mathbf{r}_i$ and f_i denote the position and atomic scattering factor of atom i, respectively.

If the molecules are randomly oriented, it is possible to use the Debye equation to compute the scattered intensity:

$$\begin{aligned} I(q) = \langle I(\mathbf{q}) \rangle &= \left\langle \sum_{i=1}^{N} \sum_{j=1}^{N} f_i(q) f_j(q) \exp\left(i\mathbf{q} \cdot \mathbf{r}_{ij}\right) \right\rangle \\ &= \sum_{i=1}^{N} \sum_{j=1}^{N} f_i(q) f_j(q) \frac{\sin\left(qr_{ij}\right)}{qr_{ij}}, \end{aligned} \tag{2}$$

where $q = |\mathbf{q}| = 4\pi\sin\theta/\lambda$, with 2θ the scattering angle and λ the X-ray wavelength, $\mathbf{r}_{ij}$ is the vector distance from particle i to particle j and $\langle \ldots \rangle$ indicates the spherical average over all the orientations. Here, the atomic scattering factor $f_i(q)$ can be computed using the Cromer–Mann analytic function:

$$f_i(q) = \sum_{k=1}^{4} a_k \exp\left[-b_k(q/4\pi)^2\right] + c, \tag{3}$$

with the parameters a_k, b_k and c available in *International Tables for Crystallography* (Cromer & Waber, 1965; Brown *et al.*, 2006).

While equation (2), in combination with these form factors, is effective in computing the scattering intensities of biomolecules *in vacuo*, additional effects must be considered for a realistic representation of scattering in solution: (i) the electron density of the solvent displaced by the molecule; (ii) the excess of electron density in the hydration shell; and (iii) the conformational averaging of the molecules.

In SAXS measurements, the background buffer scattering is subtracted from the sample scattering to remove unwanted solvent signal. To take the displaced solvent effect into account, an approach commonly adopted consists of using reduced atomic scattering factors, according to Fraser *et al.* (1978):

$$f_i'(q) = f_i(q) - v_i\,\rho_b \exp\left(-\frac{q^2 v_i^{2/3}}{4\pi}\right), \tag{4}$$

where v_i is the tabulated displaced solvent volume of atom i and ρ_b is the electron density of bulk water (*e.g.* 0.334 e Å^{-3}). These modified form factors can be used in place of $f_i(q)$ in the Debye equation to include the effect of the displaced solvent.

Further, it should be considered that the density of the solvent layer around a molecule can be different from the density of the bulk solvent owing to solute–solvent interactions, thus resulting in additional scattering terms. The inclusion of this effect in theoretical calculations of scattering intensities can be performed either by modelling the hydration shell implicitly, as in *CRYSOL* (Svergun *et al.*, 1995) and *FoXS* (Schneidman-Duhovny *et al.*, 2013, 2010) among others, or via more computationally expensive explicit solvent approaches (Chen & Hub, 2015, 2014; Knight & Hub, 2015; Köfinger & Hummer, 2013; Park *et al.*, 2009). Most of these techniques require the adjustment of one fitting parameter against the experimental data, to tune the level of contrast in the hydration shell. Interestingly, it has been shown (Björling *et al.*, 2015; Niebling *et al.*, 2014) that, if data are recorded as

differences between two states, which is the typical case in time-resolved scattering experiments, the contribution of the solvation layer can be neglected.

Finally, conformational averaging can be included by averaging over multiple configurations of the system generated, for example, by MD simulations (Yang *et al.*, 2009).

2.2. Coarse-grain form factors

The Debye equation (2) requires the evaluation of pairwise distances between all the atoms in a biomolecule. This is an $O(N^2)$ problem, where N is the number of atoms, that becomes more and more computationally expensive as the dimension of the system increases. This is particularly serious when multiple evaluations of the scattering profile are required, as in the case of MD simulations driven by SAXS data, in iterative refinement and modelling, or when several trial structures must be tested. Several approaches circumvent this problem by adopting a coarse-grain representation of the biomolecule to reduce the cost of the Debye summation (Yang *et al.*, 2009; Ravikumar *et al.*, 2013; Stovgaard *et al.*, 2010; Zheng & Tekpinar, 2011; Niebling *et al.*, 2014). According to this strategy, which is well justified by the low resolution of SAXS data, the molecule of interest is represented as a collection of beads, each comprising a variable number of atoms. The dimension of the beads can be tuned to find a proper balance between accuracy and computational efficiency (Niebling *et al.*, 2014). The beads can also be placed in different positions: examples include the atom's centre of mass, the centre of electron-density distribution or, in the case of protein residues, the Cα atom (Tong *et al.*, 2016). Given M beads and their associated scattering factors $F(q)$, the Debye equation becomes

$$I(q) \cong \sum_{i=1}^{M} \sum_{j=1}^{M} F_i(q) F_j(q) \frac{\sin(qR_{ij})}{qR_{ij}}, \qquad (5)$$

where the indices i and j run over the beads and R_{ij} is their relative distance. Computing accurate coarse-grain scattering factors $F(q)$ is a non-trivial task and diverse strategies, with different degrees of accuracy, can be employed. A review and a comparison of some of these possibilities, accompanied by a description of the approximations used in each case, is given by Niebling *et al.* (2014). Among these, the SBA method proposed by Yang *et al.* (2009) has emerged as a reliable and fast method to calculate effective form factors. Herein, $F_i(q)$ is calculated to reproduce the scattering intensity of an isolated bead i according to

$$F_i(q) = \left[\sum_{k \in i} \sum_{l \in i} f'_k(q) f'_l(q) \frac{\sin(qr_{kl})}{qr_{kl}} \right]^{1/2}, \qquad (6)$$

where the atomic scattering factors f' are the ones corrected for the excluded volume in equation (4). The use of the reduced form factors in equation (6) could cause the violation of the condition

$$F_i(q = 0) = \sum_{k \in i} f'_k(q = 0), \qquad (7)$$

which guarantees computation of the correct scattering intensities at $q = 0$. This is because equation (6) always produces positive coarse-grain form factors, while the sum in equation (7) can be negative (*e.g.* in the case of small beads containing several hydrogen atoms, whose reduced atomic scattering factor at $q = 0$ corresponds to -0.72 electron units). To overcome this problem, Niebling *et al.* (2014) proposed to correct the form factors $F_i(q)$ not satisfying equation (7) by fitting a sixth-order polynomial to data with q larger than the high-q inflection point and imposing a constraint at $q = 0$. The resulting curve is then used as the corrected coarse-grain form factor.

2.3. Coarse-grain nucleic acid representation with Martini

Here, we applied the SBA approach to compute coarse-grained form factors for DNA and RNA, using the Martini force field (Uusitalo *et al.*, 2015, 2017; Marrink & Tieleman, 2013) as the mapping scheme and following the work done by Niebling *et al.* (2014) for proteins. In the Martini force field, each nucleotide is represented by six or seven beads. The backbone is mapped with three beads (one including the phosphate group and two comprising the atoms of the sugar ring), while the nitrogen bases are modelled with three (cytosine and thymine) or four beads (adenine and guanine).

We computed coarse-grain SBA form factors for each Martini bead in the eight nucleotides, averaging over several crystal structures. The correction described in the previous section was applied for those beads not satisfying equation (7) (*i.e.* the backbone beads BB3 of DNA nucleotides; see Table S1 in the supporting information). Furthermore, we added two terminal beads for each nucleotide, denoted TE5 and TE3 (for the 5′ and 3′ termini, respectively), which are obtained by including the terminal hydroxyl groups in the backbone beads BB2 and BB3, respectively. In the calculation of scattering intensities, the position of each bead was placed at the centre of mass of the non-hydrogen atoms belonging to the bead, according to the Martini model. The only exception is represented by the positioning of the termini, for which the terminal oxygen atom was not considered to be coherent with the Martini representation of the nucleotide. This implies that Martini structures can be directly used as input for the evaluation of scattering intensities, facilitating the use of SAXS data as restraints in Martini MD simulations.

2.4. DNA and RNA data sets

To compute accurate coarse-grain form factors considering the internal details of each bead, the summation in equation (6) is expected to be averaged over several different conformations. To achieve this, a set of non-redundant molecular structures from the Protein Data Bank (PDB) was employed. This includes (i) a manually curated library of 77 X-ray crystal structures for RNA (Bernauer *et al.*, 2011), selected to be nonredundant and with resolution higher than 3.5 Å; and (ii) 175 crystal structures, selected from the data set used by Svozil *et al.* (2008), containing non-complexed DNA structures adopting a wide variety of conformations (45 in A form, 72 B

form, 39 Z form and 19 quadruplexes). Hydrogen atoms were added with the *Reduce* software (Word *et al.*, 1999). All the listed structures were used for the calculation of the averaged form factors. For the validation, we considered only a subset of these crystal structures, for which neither missing non-hydrogen atoms nor missing or modified residues were present. This resulted in a validation set comprising 44 PDB structures for RNA and 121 for DNA. A complete list of the PDB codes included in the data sets is given in Table S2.

2.5. Hybrid all-atom/coarse-grain SAXS calculation and *PLUMED-ISDB* implementation

In order to facilitate the adoption of the computed coarse-grain form factors, we implemented the Martini form factors for both proteins (Niebling *et al.*, 2014) and nucleic acids (as computed in this work; Table S3) in *PLUMED-ISDB* (Bonomi & Camilloni, 2017), an open-source module of the *PLUMED* library (Tribello *et al.*, 2014) thought to be compatible with several MD codes.

PLUMED-ISDB is a module specifically designed to allow the integration of diverse experimental data (including SAXS, NMR observables, Förster resonance energy transfer and cryo-electron microscopy) with prior information and can be exploited with different aims: refinement of single structures against experimental data; determination of conformational ensembles via experimental-driven multi-replica MD simulations; or post-processing of existing models and trajectories via the *PLUMED* driver tool. Importantly, in *PLUMED-ISDB* the integration of experimental data in simulations is achieved via metainference (Bonomi *et al.*, 2016), a Bayesian framework which is able to account for the approximations involved in the SAXS calculation (*i.e.* the coarse-grain representation does not consider the excess of electron density in the hydration shell, and the coarse graining itself), as well as for the noise in the data. In Section S1 of the supporting information we give a brief review of this approach, showing the explicit form of the metainference energy (Löhr *et al.*, 2017) in the case in which the global error is modelled by a Gaussian per data point (which is the case for properly collected and processed SAXS data; Franke *et al.*, 2015).
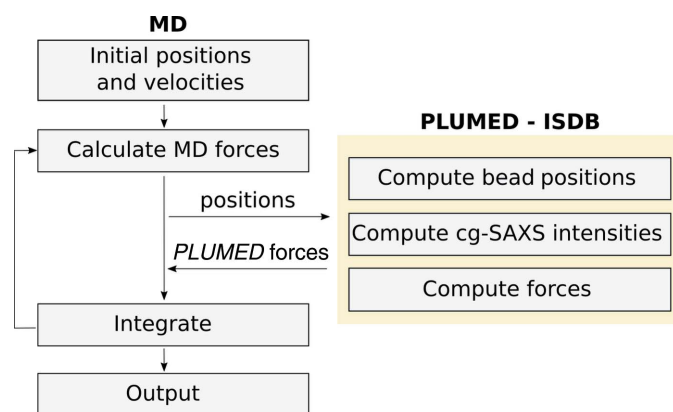

Figure 1
A flowchart for SAXS-driven MD simulations in the hybrid multi-resolution mode.

Concerning the SAXS experimental observables in *PLUMED-ISDB*, in addition to the implementation of the Martini form factors for both protein and nucleic acids, we also provide an implementation of the atomic scattering factors, corrected by the excluded volume, and the possibility of defining new custom form factors using a polynomial expansion of any order. This allows high customizability of SAXS-driven simulations, which can be run in different modes: (i) the atomistic mode, using both an all-atom force field and an atomistic forward model; (ii) the coarse-grain mode, using both the Martini force field and the form factor; and (iii) the hybrid multi-resolution mode, where the simulations are run with an atomistic force field and the Martini or other user-defined form factors are used for the forward model.

In this work we present two cases of structure refinement in which the hybrid multi-resolution mode has been adopted. This approach allows us to perform accurate simulations with full atomistic details while significantly speeding up the calculation of SAXS intensities. Practically, given an atomic resolution structure, which is sampled during the simulation, the positions of the Martini beads are calculated on the fly by *PLUMED*. The beads are then associated with virtual atoms and subsequently used in combination with the appropriate form factors to calculate the SAXS intensities (Fig. 1). In Section S2 of the supporting information we provide more details of how to set up metainference simulations with this hybrid multi-resolution method. Further, we provide a tutorial (https://plumed.github.io/doc-master/user-doc/html/_i_s_d_b_tutorial.html) with step-by-step instructions to (i) calculate all-atom and coarse-grained SAXS profiles from a structure and/or a pre-existing trajectory; and (ii) run a SAXS-driven metainference simulation with the hybrid approach.

2.6. Computational details of the simulations

MD simulations were performed on two protein/nucleic acid complexes. The first is the ComE–*comcde* DNA–protein complex, for which both experimental SAXS data and a calculated model are deposited in the SASBDB (Valentini *et al.*, 2015), entry SASDAB7 (Sanchez *et al.*, 2015). The second is the complex of a single-stranded 12-*mer* oligonucleotide with a region of the heterogeneous nuclear ribonucleoprotein A1 (Kooshapur *et al.*, 2018), which we previously refined using standard atomistic scattering factors. All MD simulations were performed with *GROMACS 2018*, *PLUMED 2* and the *PLUMED-ISDB* module (Tribello *et al.*, 2014; Abraham *et al.*, 2015; Bonomi & Camilloni, 2017), using the amber14sb force field for proteins (Maier *et al.*, 2015) with *parmbsc1* (Ivani *et al.*, 2016) and the TIP3P water model (Jorgensen *et al.*, 1983).

In the case of the ComE–*comcde* DNA–protein complex, two runs were performed: (i) a metainference simulation on a single replica in which SAXS restraints were used; and (ii) an unrestrained simulation, to be used as reference, without the inclusion of experimental information. Both simulations were evolved for a total of 5 ns through a series of 20 simulated annealing cycles, with the temperature varying between 300

and 400 K. To avoid the opening of DNA in the high-temperature intervals, in both simulations we added hydrogen-bond restraints involving the first and last two pairs of nucleotides. For the metainference simulation, a set of 15 representative SAXS intensities at different scattering vector magnitudes, ranging between 0.02 and 0.3 Å^{-1}, were included as restraints.

In the case of the RNA–protein complex, a metainference simulation was evolved for 5 ns, maintaining the temperature at 300 K. Restraints in the form of harmonic upper-wall potentials were applied as described by Kooshapur *et al.* (2018) to maintain critical protein–RNA interface contacts, salt bridges and protein secondary structures, as found in the related crystal structure (PDB code 6dcl). A total of 43 representative SAXS intensities were used as restraints in the metainference, corresponding to scattering vectors between 0.03 and 0.45 Å^{-1}.

Further details of the simulation are provided in Section S3 of the supporting information. Additionally, complete example files to run metainference simulations with the hybrid multi-resolution mode are provided in our GitHub repository, https://github.com/carlocamilloni/papers-data.

For each run, one reference model was selected, clustering [based on the root-mean-square deviation of the position (RMSD)] the structures sampled at 300 K in the second half of the run and choosing the centre of the most populated cluster.

3. Results and discussion

3.1. Form factors of DNA and RNA nucleotides

According to the Martini mapping scheme, the RNA and DNA nucleotides are represented with six or seven beads.

Here we also considered two additional beads per nucleotide, representing the 5′ and 3′ terminal beads and being a simple modification (*i.e.* addition of a terminal hydroxyl group) of the backbone beads BB2 and BB3. While non-differentiating the terminal beads in protein is acceptable, this approximation in nucleic acids could result in computed scattering intensities with large deviations from the atomistic ones, especially when short oligomers are considered and for scattering vectors close to $q = 0$. In total, this results in the computation of 34 coarse-grain form factors for DNA and 34 for RNA, listed in Table S1.

For each bead, the effective form factor is calculated using equation (6) as described in the *Theory and methods* section, averaging over several different structures taken from the PDB (Table S2). The resulting coarse-grain form factors are shown in Fig. 2 and listed in Table S3. We observed that beads with the same chemical composition in different nucleotides (*i.e.* all the backbone and terminal beads, purine beads SC1–SC4, and pyrimidine bead SC1) have perfectly superimposable Martini form factors. This also holds true when comparing corresponding beads in DNA and RNA, with the obvious exceptions of the SC3 bead for thymine/uracil and the backbone BB3/TE3 beads, which in RNA contain the additional oxygen atom in position 2′. It is worthy of note that the variability observed between the individual form factors to be averaged (Fig. S1) is smaller than that observed in proteins (Niebling *et al.*, 2014). This can be explained by the highly conserved structural arrangement of the different atoms within each bead and it is a promising indication of reliability for coarse-grain scattering calculations. The only exceptions to this behaviour are represented by the terminal beads, which display a larger variation due to the different orientations that

Figure 2
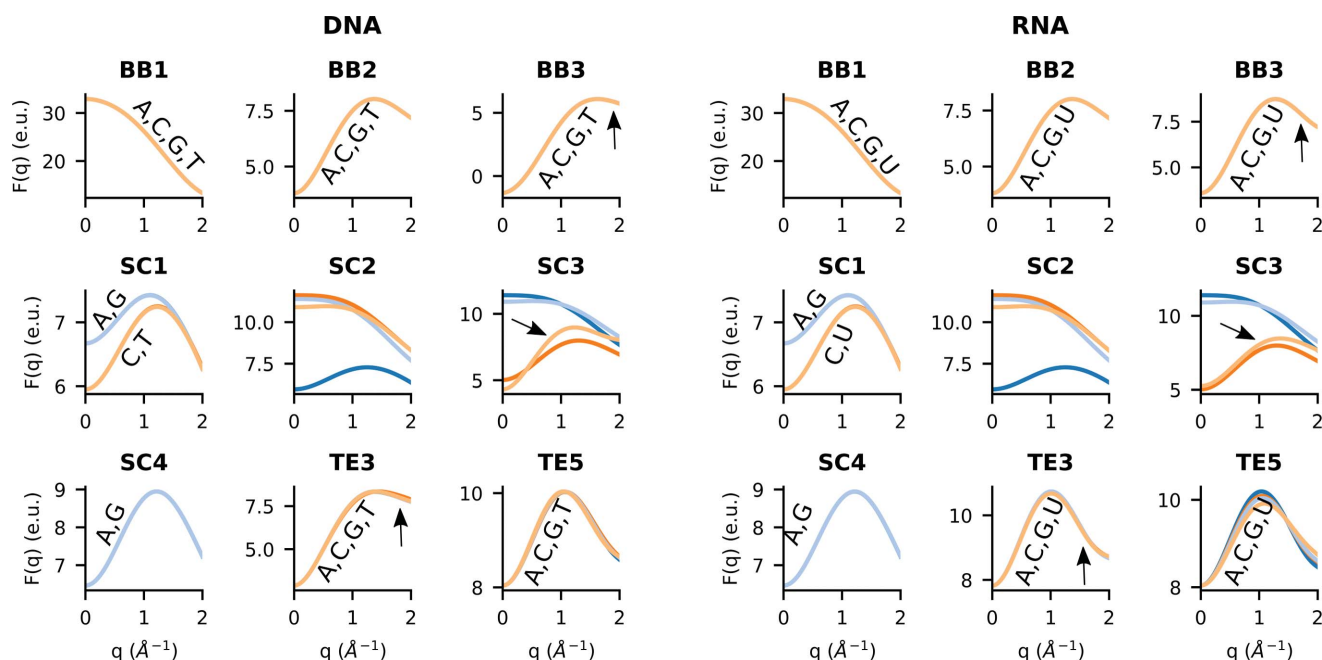Calculated coarse-grain form factors for (left) DNA and (right) RNA. The form factors are represented for all Martini beads of the nucleotides: adenine (blue), cytosine (dark orange), guanine (light blue) and thymine/uracil (light orange). When two or more structure factors are superimposed, labels are added for clarity. Differences between corresponding DNA and RNA beads are highlighted with black arrows.

the terminal hydroxyl group can assume with respect to the other atoms of the bead.

3.2. Comparison of scattering intensities computed with all-atom or coarse-grain form factors

To evaluate the accuracy of Martini form factors in computing scattering intensities for nucleic acids, we compared coarse-grain SAXS profiles with the atomistic ones for a library of crystal structures, comprising 44 RNA and 121 DNA structures (Table S2). The differences between the Martini and atomistic curves were measured by calculating the average relative squared error over different q values:

$$\text{Err} = \frac{1}{N} \sum_{q=0}^{q(N)} \frac{\left[I_{\text{AA}}(q) - I_{\text{CG}}(q)\right]^2}{I_{\text{AA}}(q)^2}, \tag{8}$$

where $I_{\text{AA/CG}}$ are the intensities computed with all-atom or Martini form factors, respectively, N is the number of data points, and q varies between 0 and $q(N)$, with the step between $q(i)$ and $q(i-1)$ equal to 0.01 Å^{-1}. To compare our results with those obtained for proteins by Niebling *et al.* (2014), we identified the maximum q value for which the error of equation (8) is smaller than 0.2%. This value, denoted $q_{\text{threshold}}$, was computed for each of the considered RNA and DNA structures and its distribution is reported in Fig. 3. The average $q_{\text{threshold}}$ values for RNA and DNA (0.47 and 0.88 Å^{-1}, respectively) are comparable to that previously found for proteins (0.53 Å^{-1}) and suggest that scattering intensities up to $q \simeq 0.45$ Å^{-1} can be reliably calculated using the coarse-grain approximation. Fig. 3 highlights a distinct behaviour for DNA and RNA: while the $q_{\text{threshold}}$ values display small deviations in the RNA structures, they are spread considerably in the case of DNA. We found that different DNA conformations are associated with diverse values of $q_{\text{threshold}}$,
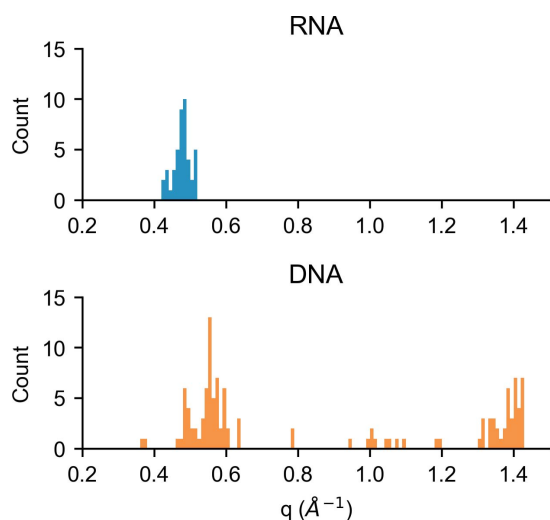

Figure 3
Distribution of $q_{\text{threshold}}$ values for 44 RNA and 121 DNA crystallographic structures. Each $q_{\text{threshold}}$ value has been computed by comparing atomistic and coarse-grain scattering curves and represents the maximum q value for which the error defined in equation (8) is smaller than 0.2%.

where DNA in the A or Z forms mainly displays $q_{\text{threshold}}$ between 0.4 and 0.6 Å^{-1}, while B-form DNA structures often reach values greater than 1.0 Å^{-1}, conceivably because of their less compact structure (see Fig. S2). The range of q values (up to 0.45 Å^{-1}) over which coarse-grained intensities can be considered a reliable approximation of the atomistic ones was confirmed by further tests (see Section S4 and Figs. S3–S5).

We can therefore conclude that the form factors derived herein for nucleic acids can be seamlessly combined with those for proteins and then used for an efficient back-calculation of SAXS curves in protein–nucleic acid complexes for scattering vectors up to 0.45 Å^{-1}. The computational efficiency gained using these coarse-grain form factors is very important in applications where repeated evaluations of scattering intensities are requested. They could be exploited to drive MD simulations to match SAXS data, allowing an extension of the system size and the simulation length with respect to previous applications. The Martini form factors can naturally be included in simulations run with the coarse-grain Martini force field. Moreover, we propose a hybrid coarse-grain/all-atom approach, where the simulations are run with full atomistic details, while the Martini form factors are used for the SAXS calculation, thus allowing a faster back-calculation of the scattering intensities (*cf. Theory and methods* section).

3.3. Refinement of protein–nucleic acid complexes against SAXS data

To demonstrate the efficiency and reasonable accuracy of the Martini form factors in experimentally driven MD simulations, we exploited them in the refinement of protein–nucleic acid complexes against SAXS data. To this end, we took advantage of the metainference technique, which allows the introduction of noisy and ensemble-averaged experimental data in MD simulations. Importantly, metainference also takes errors resulting from the forward model into account. This is particularly relevant here, since the approximations resulting from the coarse-grain representation do not consider the excess of electron density in the hydration shell.

3.3.1. ComE–*comcde* DNA–protein complex. ComE is a two-domain protein, part of the ComD–ComE two-component signalling system, which dimerizes in solution via its REC domain when activated by ComD-induced phosphorylation. In the work by Sanchez *et al.* (2015), SAXS data were used to show that the ComE$^{\text{D58E}}$ active mimic mutant is found in dimeric form when bound to the promoter region *comcde*. Furthermore, it induces an extra bending of DNA. In that work, a model comprising two ComE bound to the 38-*mer comcde* duplex was built to fit SAXS data exploiting the available crystallographic structure of the ComE dimer (PDB code 4cbv; Boudes *et al.*, 2014). The model proposed by Sanchez and co-workers displays a good agreement with the SAXS data and provides interesting structural insights into the ComE–*comcde* binding mode. Here, we show that our metainference-based approach could be exploited to improve the quality of the model, solving steric clashes and other

Table 1
Evaluation of representative protein–DNA structures in terms of agreement with SAXS data and model quality.

The following structures are considered: (i) the initial model; (ii) the initial model after energy minimization; (iii) a refined model extracted from metainference simulation; and (iv) a representative structure extracted from the unrestrained simulation. The agreement with SAXS data was measured with *CRYSOL* (Svergun *et al.*, 1995) using the maximum order of harmonics available and 18 points for the Fibonacci grid. The contrast of the solvation shell was fixed to 0.01. The values for the radius of the atomic group and the excluded volume were the default ones computed by *CRYSOL*. The model quality was assessed using the *Molprobity* validation implemented in *Phenix* (Adams *et al.*, 2010; Davis *et al.*, 2007).

	Parameter	Initial model	Minimized	Refined model (metainference)	Unrestrained
Agreement with SAXS	*CRYSOL* χ^2	4.03	2.43	1.58	7.21
	CRYSOL χ^2 ($q < 0.3$)	4.76	2.58	1.75	9.40
Model quality	*Molprobity* score	3.41	2.26	1.75	1.85
	Clash score	32.1	1.85	0.74	1.11
	Ramachandran favoured (%)	90	90	93	91
	Ramachandran outliers (%)	1.0	1.4	1.0	1.4

Table 2
A summary of the performance of protein–DNA and protein–RNA simulations, achieved using the metainference approach (computing the scattering intensities with either atomistic or Martini form factors) or without the integration of SAXS data (unrestrained).

Data for metainference simulations in which the bias is applied every ten time steps are reported. Performance was estimated on an Intel Xeon E5 3.5 GHz using four cores. The number of protein or nucleic acid (NA) atoms and beads for each system is also indicated.

	Performance (ns per day)					No. of atoms		No. of beads	
	Atomistic		Martini						
	Step = 1	Step = 10	Step = 1	Step = 10	Unrestrained	Protein	NA	Protein	NA
Protein–DNA	0.01	0.08	0.44	2.26	4.77	8687	2417	1251	492
Protein–RNA	0.04	0.35	1.13	4.37	5.77	3124	387	449	79

defects created during the modelling phase, while further improving the agreement with SAXS data.

After a short energy minimization of the system, whose initial coordinates were taken from SASBDB entry SASDAB7, we performed a 5 ns-long simulated annealing (SA) simulation in which metainference was used to introduce SAXS restraints (see *Theory and methods* and Section S3). We selected the refined model clustering the structures sampled at 300 K in the last ten SA cycles (out of 20) based on geometric similarity and choosing the centre of the most populated cluster. To check the importance of using SAXS data, we also performed an additional run using the same SA protocol without metainference (*i.e.* without introducing experimental restraints).

The agreement with the SAXS data and the model quality were assessed using *CRYSOL* and *Molprobity* (Adams *et al.*, 2010; Davis *et al.*, 2007), comparing four different structures: the initial model, the initial model after energy minimization, the refined structure extracted from the metainference simulation and the structure selected from the unrestrained simulation. While the energy minimization can solve most of the steric clashes, metainference simulations are useful for further improving both the agreement with the experimental data and the quality of the model in terms of Ramachandran and clash scores (Table 1).

The refined model (Fig. 4) displays a χ^2 of 1.58, slightly better than the initial one, maintains the known critical interactions between ComE and the DNA recognition sites (mainly involving residues H168, K203 and K235), and shows only small deviations from the crystal structures of the ComE

dimer (backbone RMSD of 2.5 Å with respect to the reference structure in PDB code 4cbv). Conversely, the structure extracted from the unrestrained simulation, even though it has a good *Molprobity* score, misses most of the DNA–ComE contacts and significantly alters the ComE dimer

Figure 4
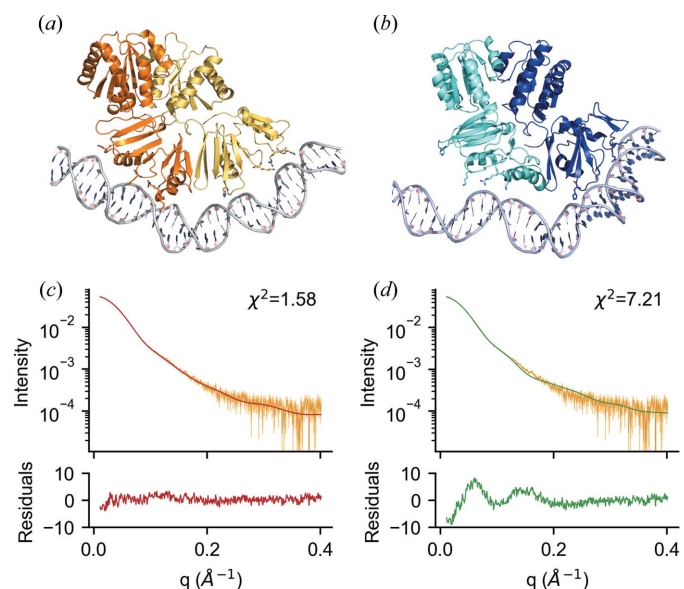The representative ComE–DNA structures extracted from (*a*), (*c*) the metainference and (*b*), (*d*) the unrestrained simulations, along with their fits to the experimental data and residuals, according to *CRYSOL*. In the *CRYSOL* calculations the contrast of the solvation shell was fixed to 0.01, while the default values for the radius of the atomic group and the excluded volume were used.

conformation, showing a backbone RMSD of 6.3 Å with respect to the crystal structure. Overall, this results in a poor agreement with the experimental data, confirmed by a significantly higher χ^2 value.

Note that all the χ^2 values reporting the distance from experimental SAXS data were calculated with *CRYSOL* (with the contrast of the solvation shell fixed to 0.01 and default values for the radius of the atomic group and excluded volume) and are therefore independent of the strategy used to back-calculate SAXS intensities and consequently restrain the simulations.

Importantly, we observed that the integration of SAXS data in simulation is prohibitive if atomistic scattering factors are used for the back-calculations of the intensities; however, its impact on MD performance is significantly reduced by exploiting the Martini form factors and adopting the multiple time-step strategy, where the metainference bias is applied only every ten time steps (Table 2).

3.3.2. The UP1–miRNA complex. As a second test system, we used a protein–RNA complex that we previously refined against SAXS and NMR data (Kooshapur *et al.*, 2018). This complex involves the binding of a single-stranded 12-*mer* oligonucleotide, derived from the micro RNA 18*a* primary transcript, and the unwinding protein 1 (UP1), comprising two tandem domains that constitute the RNA-binding region of the heterogeneous nuclear ribonucleoprotein A1. The previously solved crystal structure of UP1–12-*mer* RNA (PDB code 6dcl) showed a 2:2 stoichiometry that was found to be not representative of the 1:1 stoichiometry measured in solution. Therefore, a refinement against experimental data in solution was performed, where the key features of the crystallographic binding interface were retained, leading to a model of UP1–12-*mer* RNA with the correct 1:1 stoichiometry (Kooshapur *et al.*, 2018). Here we reproduce this same refinement procedure, using an analogous approach where the main protein–RNA and protein–protein interaction sites are restrained (see *Theory and methods*). The resulting UP1–12-*mer* RNA model, extracted from our metainference simulation as described in the methods section, is compatible with the reference one (Table S4). The representative metainference-derived model and the fit with experimental data, according to *CRYSOL*, are shown in Fig. 5. Note that the adoption of Martini form factors

in combination with a multiple time-step scheme allows us to approach the performance of the unrestrained simulation (4.4 ns per day versus 5.8 ns per day), outperforming the simulations relying on atomistic scattering factors (0.35 ns per day also when applying metainference bias every ten time steps; see Table 2).

4. Conclusions

The cost of computing scattering intensities from atomic structures is a limiting factor for the integration of SAXS experimental data in MD simulations and for other applications where multiple evaluations of scattering curves are required. Here, we have extended the work of Niebling *et al.* (2014) to nucleic acids, computing Martini bead form factors for RNA and DNA and showing that they can be exploited for the accurate evaluation of SAXS intensities for scattering vectors up to 0.45 Å^{-1}. We have implemented these coarse-grain form factors in *PLUMED* and shown how they could be used for the structure refinement of molecular systems against SAXS data by adopting a hybrid atomistic/coarse-grain approach.

Overall, our results clearly indicate that Martini form factors, for both proteins and nucleic acids, can safely be used to restrain atomistic simulations against SAXS intensities, reproducing the experimental data with an accuracy comparable to that achieved in atomistic mode and improving the performance by up to a factor of 50.

We anticipate that our protocol, using metainference, could be used in combination with other experimental data, and also extended to run multiple-replica simulations taking molecular conformational averaging into account. Lastly, we note that the applicability of the Martini form factors is not limited to their use in SAXS-driven MD simulations, but they could also be used for the analysis of single structures or trajectories exploiting the *PLUMED* driver utility.

5. Related literature

The following additional references are cited in the supporting information: Berendsen *et al.* (1984), Bussi *et al.* (2007), Darden *et al.* (1993), Ferrarotti *et al.* (2015), Hess *et al.* (1998), Kimanius *et al.* (2015) and Rieping *et al.* (2005).

Figure 5
(*a*) A representative UP1–miRNA structure extracted from metainference, and (*b*) its fit to the experimental data and residuals, according to *CRYSOL*.

References

Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B. & Lindahl, E. (2015). *SoftwareX*, **1–2**, 19–25.

Adams, P. D., Afonine, P. V., Bunkóczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L.-W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C. & Zwart, P. H. (2010). *Acta Cryst.* D**66**, 213–221.
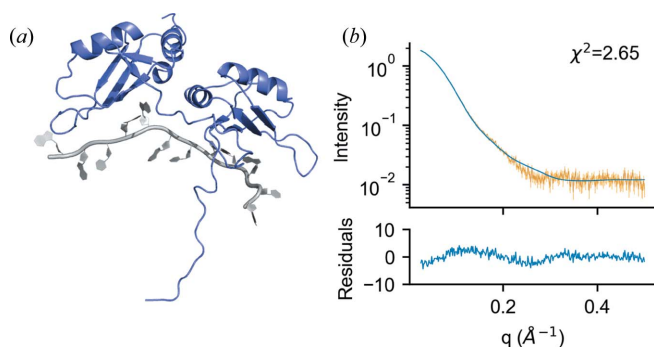
Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A. & Haak, J. R. (1984). *J. Chem. Phys.* **81**, 3684–3690.

Berlin, K., Gumerov, N. A., Fushman, D. & Duraiswami, R. (2014). *J. Appl. Cryst.* **47**, 755–761.

Bernauer, J., Huang, X., Sim, A. Y. L. & Levitt, M. (2011). *RNA*, **17**, 1066–1075

Björling, A., Niebling, S., Marcellini, M., van der Spoel, D. & Westenhoff, S. (2015). *J. Chem. Theory Comput.* **11**, 780–787.

Bonomi, M. & Camilloni, C. (2017). *Bioinformatics*, **33**, 3999–4000.

Bonomi, M., Camilloni, C., Cavalli, A. & Vendruscolo, M. (2016). *Sci. Adv.* **2**, e1501177.

Boudes, M., Sanchez, D., Graille, M., van Tilbeurgh, H., Durand, D. & Quevillon-Cheruel, S. (2014). *Nucleic Acids Res.* **42**, 5302–5313.

Brown, P. J., Fox, A. G., Maslen, E. N., O'Keefe, M. A. & Willis, B. T. M. (2006). *International Tables for Crystallography*, Vol. C, *Mathematical, Physical and Chemical Tables*, 1st online ed., edited by E. Prince, pp. 554–595. Chester: International Union of Crystallography.

Bussi, G., Donadio, D. & Parrinello, M. (2007). *J. Chem. Phys.* **126**, 14101.

Chen, P. C. & Hub, J. S. (2014). *Biophys. J.* **107**, 435–447.

Chen, P. C. & Hub, J. S. (2015). *Biophys. J.* **108**, 2573–2584.

Cromer, D. T. & Waber, J. T. (1965). *Acta Cryst.* **18**, 104–109.

Darden, T., York, D. & Pedersen, L. (1993). *J. Chem. Phys.* **98**, 10089–10092.

Davis, I. W., Leaver-Fay, A., Chen, V. B., Block, J. N., Kapral, G. J., Wang, X., Murray, L. W., Arendall, W. B., Snoeyink, J., Richardson, J. S. & Richardson, D. C. (2007). *Nucleic Acids Res.* **35**, W375–W383.

Ferrarotti, M. J., Bottaro, S., Pérez-Villa, A. & Bussi, G. (2015). *J. Chem. Theory Comput.* **11**, 139–146.

Franke, D., Jeffries, C. M. & Svergun, D. I. (2015). *Nat. Methods*, **12**, 419–422.

Fraser, R. D. B., MacRae, T. P. & Suzuki, E. (1978). *J. Appl. Cryst.* **11**, 693–694.

Gorba, C., Miyashita, O. & Tama, F. (2008). *Biophys. J.* **94**, 1589–1599.

Hess, B., Bekker, H., Berendsen, H. J. C. & Fraaije, J. G. E. M. (1998). *J. Comput. Chem.* **18**, 1463–1472.

Hub, J. S. (2018). *Curr. Opin. Struct. Biol.* **49**, 18–26.

Ivani, I., Dans, P. D., Noy, A., Pérez, A., Faustino, I., Hospital, A., Walther, J., Andrio, P., Goñi, R., Balaceanu, A., Portella, G., Battistini, F., Gelpí, J. L., González, C., Vendruscolo, M., Laughton, C. A., Harris, S. A., Case, D. A. & Orozco, M. (2016). *Nat. Methods*, **13**, 55–58.

Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. (1983). *J. Chem. Phys.* **79**, 926–935.

Kimanius, D., Pettersson, I., Schluckebier, G., Lindahl, E. & Andersson, M. (2015). *J. Chem. Theory Comput.* **11**, 3491–3498.

Knight, C. J. & Hub, J. S. (2015). *Nucleic Acids Res.* **43**, W225–W230.

Koch, M. H. J., Vachette, P. & Svergun, D. I. (2003). *Q. Rev. Biophys.* **36**, 147–227.

Köfinger, J. & Hummer, G. (2013). *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **87**, 052712.

Kooshapur, H., Choudhury, N. R., Simon, B., Mühlbauer, M., Jussupow, A., Fernandez, N., Jones, A. N., Dallmann, A., Gabel, F., Camilloni, C., Michlewski, G., Caceres, J. F. & Sattler, M. (2018). *Nat. Commun.* **9**, 2479.

Levantino, M., Yorke, B. A., Monteiro, D. C. F., Cammarata, M. & Pearson, A. R. (2015). *Curr. Opin. Struct. Biol.* **35**, 41–48.

Löhr, T., Jussupow, A. & Camilloni, C. (2017). *J. Chem. Phys.* **146**, 165102.

Maier, J. A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K. E. & Simmerling, C. (2015). *J. Chem. Theory Comput.* **11**, 3696–3713.

Marchi, M. (2016). *J. Chem. Phys.* **145**, 045101.

Marrink, S. J. & Tieleman, D. P. (2013). *Chem. Soc. Rev.* **42**, 6801–6822.

Niebling, S., Björling, A. & Westenhoff, S. (2014). *J. Appl. Cryst.* **47**, 1190–1198.

Panjkovich, A. & Svergun, D. I. (2016). *Phys. Chem. Chem. Phys.* **18**, 5707–5719.

Park, S., Bardhan, J. P., Roux, B. & Makowski, L. (2009). *J. Chem. Phys.* **130**, 134114.

Ravikumar, K. M., Huang, W. & Yang, S. (2013). *J. Chem. Phys.* **138**, 024112.

Rieping, W., Habeck, M. & Nilges, M. (2005). *Science*, **309**, 303–306.

Sanchez, D., Boudes, M., van Tilbeurgh, H., Durand, D. & Quevillon-Cheruel, S. (2015). *FEBS J.* **282**, 1538–1553.

Schneidman-Duhovny, D., Hammel, M. & Sali, A. (2010). *Nucleic Acids Res.* **38**, W540–W544.

Schneidman-Duhovny, D., Hammel, M., Tainer, J. A. & Sali, A. (2013). *Biophys. J.* **105**, 962–974.

Schwieters, C. D., Kuszewski, J. J., Tjandra, N. & Marius Clore, G. (2003). *J. Magn. Reson.* **160**, 65–73.

Stovgaard, K., Andreetta, C., Ferkinghoff-Borg, J. & Hamelryck, T. (2010). *BMC Bioinformatics*, **11**, 429.

Svergun, D., Barberato, C. & Koch, M. H. J. (1995). *J. Appl. Cryst.* **28**, 768–773.

Svozil, D., Kalina, J., Omelka, M. & Schneider, B. (2008). *Nucleic Acids Res.* **36**, 3690–3706.

Tong, D., Yang, S. & Lu, L. (2016). *J. Appl. Cryst.* **49**, 1148–1161.

Tribello, G. A., Bonomi, M., Branduardi, D., Camilloni, C. & Bussi, G. (2014). *Comput. Phys. Commun.* **185**, 604–613.

Tuukkanen, A. T., Spilotros, A. & Svergun, D. I. (2017). *IUCrJ*, **4**, 518–528.

Uusitalo, J. J., Ingólfsson, H. I., Akhshi, P., Tieleman, D. P. & Marrink, S. J. (2015). *J. Chem. Theory Comput.* **11**, 3932–3945.

Uusitalo, J. J., Ingólfsson, H. I., Marrink, S. J. & Faustino, I. (2017). *Biophys. J.* **113**, 246–256.

Valentini, E., Kikhney, A. G., Previtali, G., Jeffries, C. M. & Svergun, D. I. (2015). *Nucleic Acids Res.* **43**, D357–D363.

Word, J. M., Lovell, S. C., Richardson, J. S. & Richardson, D. C. (1999). *J. Mol. Biol.* **285**, 1735–1747.

Yang, S., Park, S., Makowski, L. & Roux, B. (2009). *Biophys. J.* **96**, 4449–4463.

Zheng, W. & Tekpinar, M. (2011). *Biophys. J.* **101**, 2981–2991.

JCTC
Journal of Chemical Theory and Computation

Determination of Protein Structural Ensembles by Hybrid-Resolution SAXS Restrained Molecular Dynamics

Cristina Paissoni,* Alexander Jussupow, and Carlo Camilloni*

Read Online

ACCESS | Metrics & More | Article Recommendations | SI Supporting Information

ABSTRACT: Small-angle X-ray scattering (SAXS) experiments provide low-resolution but valuable information about the dynamics of biomolecular systems, which could be ideally integrated into molecular dynamics (MD) simulations to accurately determine conformational ensembles of flexible proteins. The applicability of this strategy is hampered by the high computational cost required to calculate scattering intensities from three-dimensional structures. We previously presented a hybrid resolution method that makes atomistic SAXS-restrained MD simulation feasible by adopting a coarse-grained approach to efficiently back-calculate scattering intensities; here, we extend this technique, applying it in the framework of metainference with the aim to investigate the dynamical behavior of flexible biomolecules. The efficacy of the method is assessed on the K63-diubiquitin, showing that the inclusion of SAXS restraints is effective in generating a reliable conformational ensemble, improving the agreement with independent experimental data.
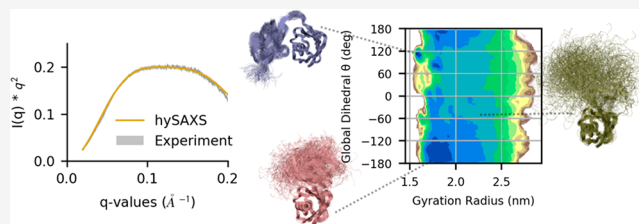
1. INTRODUCTION

Biomolecules in solution can be characterized by a different extent of conformational dynamics, depending on the specific system and experimental conditions.[1−3] While the dynamics of single-domain proteins under native conditions is generally limited to fluctuations around a well-defined structure, fully disordered proteins can only be described as statistical ensembles of conformations. Between these cases, multi-domain proteins connected by linker regions can populate multiple states generally characterized by a different size.[4]

Experimentally, the characterization of conformational heterogeneity can be achieved by employing multiple solution techniques, such as nuclear magnetic resonance (NMR), Förster resonance energy transfer (FRET), and small-angle X-ray scattering (SAXS).[1,2] The latter has the advantages of being label-free, as well as having the ability to work with systems of any size and operate under essentially all experimental conditions.[5] An atomistic interpretation of scattering data could benefit from its combination with computational techniques, such as molecular dynamics (MD) simulations, which could provide an accurate physical model to generate reliable conformational ensembles, in agreement with SAXS data.[6,7] Common approaches employ SAXS to reweight conformational ensembles a posteriori, making use of statistically founded theoretical frameworks.[8−14] Recently, few methods in which SAXS experimental data are integrated into MD to drive conformational sampling have been proposed; nevertheless, their application is hindered by the high computational cost required to calculate scattering intensities.[15−18]

In a previous work,[19] we developed a MD-based multi-resolution strategy to efficiently refine protein-DNA and protein-RNA complexes integrating SAXS experimental data with metainference.[20] According to this strategy, MD is performed with full atomistic details, using standard atomistic force fields, while the back-calculation of SAXS intensities is performed in a coarse-grain fashion,[21] based on the Martini force field.[22] In the refinement protocol, conformational averaging was not considered, under the assumption that a single structure, representing the most populated state of the system, could reliably reproduce all of the measured experimental data used as restraints.

In this work, we aim to further extend our MD-based multiresolution approach for the integration of SAXS data in the case of biomolecules that can adopt multiple conformations in solution. To this aim, we propose to take advantage of metainference technique,[20] which allows one to account for conformational flexibility by restraining the average over multiple simulations (i.e., replica) to fit with input experimental data. Importantly, multireplica metainference simulations combined with metadynamics (M&M)[23,24] have been previously exploited to integrate NMR data, showing that the inclusion of experimental restraints allows one to correct

force-field limitations, leading to well-converged conformational ensembles, independent from the employed force field.[25]

Here, we applied our multiresolution strategy to investigate the conformational ensemble of K63-linked diubiquitin (K63-Ub$_2$). Diubiquitins represent an ideal test system, since they are known to populate multiple conformational states, because of the presence of a highly flexible linker connecting the C-terminal of the distal ubiquitin with either a lysine or the N-terminus methionine of the proximal domain (Figure 1A).[26−32] In particular, the heterogeneity of K63-Ub$_2$ conformational space is supported by the presence of numerous crystallographic structures of this protein, free or in complex with diverse targets, displaying different degrees of opening and arrangements of the two subunits.[33−39] Furthermore, studies based on different biophysical techniques, including SAXS, NMR, cross-linking, and FRET, support the hypothesis that K63-Ub$_2$ in solution populates a dynamic ensemble, including both extended and compact states.[31,32,40] This equilibrium between multiple states is considered critical in modulating the affinity of diubiquitin toward its biological partners.[31]

In the following, we present our SAXS-restrained all-atom M&M[1,23,41] simulation of K63-Ub$_2$, performed with the hybrid resolution approach (hySAXS simulation), in comparison with an unrestrained reference simulation, in which the same setting was used except for the inclusion of experimental data. Both conformational ensembles indicate an equilibrium between extended and compact conformations, but their assessment with independent experimental NMR paramagnetic relaxation enhancement (PRE) experiments[32] reveals that only the hySAXS restrained simulations can accurately describe the specific contacts responsible for the formation of compact states. All of the methods described in this paper are freely available in the PLUMED-ISDB module[42] of the PLUMED library;[43] furthermore, all of the input files used are available on the PLUMED-NEST repository,[44] as plumID:19.057.

2. THEORY AND METHODS

2.1. Metainference. Metainference, combining Bayesian inference and replica-averaging modeling, allows one to integrate experimental data with prior information (generally represented by a molecular mechanic force field), taking into account the effect of conformational averaging.[20] Following the replica-averaging modeling strategies, multiple replicas of the system are simulated in parallel and the quantities to be restrained against experimental data are back-calculated as an average over the replica. Importantly, the combination of this technique with the statistical basis provided by Bayesian inference allows to tune the strength of the restraints dealing with diverse sources of errors, including random and systematic errors, as well as inaccuracies of the forward model. This is particularly important when using SAXS intensities as restraints, to account for both the noise in the data and for the possible approximations of the forward model (e.g., the coarse-grain representation and the lack of hydration layer). In the case of Gaussian noise, the metainference energy, representing the optimal balance between force field energy (E_{FF}) and experimental data, can be written as[25]

$$E_{MI} = E_{FF} + \frac{k_B T}{2} \sum_{i=1}^{N_d} \sum_{r=1}^{N_r} \frac{[d_i - \lambda f_i(\mathbf{X})]^2}{(\sigma_{r,i}^B)^2 + (\sigma_{r,i}^{SEM})^2} + E_\sigma$$

where k_B is the Boltzmann constant, T the temperature, and d_i the set of N_d experimental data. The term $f_i(\mathbf{X})$, which is given as $f_i(\mathbf{X}) = \frac{1}{N_r} \sum_{r=1}^{N_r} f_i(X_r)$, is averaged over the N_r replicas; $f_i(X_r)$ is the forward model used to predict observable i from conformation X_r, $\sigma_{r,i}^{Bias}$ is an uncertainty parameter that describes random and systematic errors, $\sigma_{r,i}^{SEM}$ is the standard error of the mean related to the conformational averaging and E_σ is an energy term that accounts for normalization of the data likelihood and error priors. Monte Carlo sampling is used to sample the uncertainty $\sigma_{r,i}^{Bias}$ and, optionally, a scaling parameter λ that relates experimental and back-calculated data (as in the case of SAXS experiment): these parameters are inferred during the simulation, along with the model of the system. Importantly, if only one replica is considered, metainference becomes equivalent to the Inferential Structure Determination approach,[45] in which Bayesian inference is exploited to combine experimental data (with proper statistical treatment of errors) and physical properties, eventually determining the probability distribution of an unknown structure and its precision. Conversely, if $\sigma_{r,i}^{Bias} = 0$, metainference is equivalent to the replica-averaged MaxEnt modeling,[46] in which errors are defined a priori to be as small as possible, rather than being determined during the simulation, along with the model of the system, thus not guaranteeing a proper statistical treatment of data and forward model errors.

Metainference can be combined with metadynamics to accelerate the exploration of the conformational space.[23,47] In particular, it was proposed to apply it in combination with parallel bias metadynamics[48] (PBMetaD), which allows the use of many collective variables (CVs) applying multiple low-dimensional bias potentials and therefore reducing the risk of missing slow degrees of freedom. In M&M, multiple copies of the simulation are run in parallel, where all the replicas use the same conditions and force field and share the bias potential, as in the case of the multiple-walkers method.[49] The coupling of metainference and metadynamics is given by the calculation of the average forward model $f_i(\mathbf{X})$, where each replica contributes differently to the average with a weight $w(X_r)$, depending on the bias potential V_{PB}, according to $w(X_r) = \exp[V_{PB}(CV(X_r),t)/(k_B T)]$.

2.2. Hybrid-Resolution SAXS-Driven Metainference Simulations. Given a coarse-grain representation of a molecule of N atoms as a collection of M beads, each comprising a variable number of atoms, if the form factors $F(q)$ of the beads are known, the scattering intensities can be approximated as

$$I(q) = \sum_{i=1}^{M} \sum_{j=1}^{M} F_i(q) F_j(q) \frac{\sin(qR_{ij})}{qR_{ij}}$$

where R_{ij} indicates the distance between the center of mass of beads ij and with the sum running over the number of beads. Therefore, the complexity is reduced from $O(N^2)$ to $O(M^2)$. The form factors $F(q)$ for custom beads can be computed by adopting the Single Bead Approximation averaging over multiple structures.[21] The level of beads coarse-graining is critical to determine the accuracy and the efficiency of the SAXS back-calculation, where large beads representing many atoms are more efficient but less accurate than smaller ones, containing only few atoms. A thoughtful comparison between schemes with different resolution (from atomistic to one-bead per amino-acid) was previously reported.[50] While one bead per
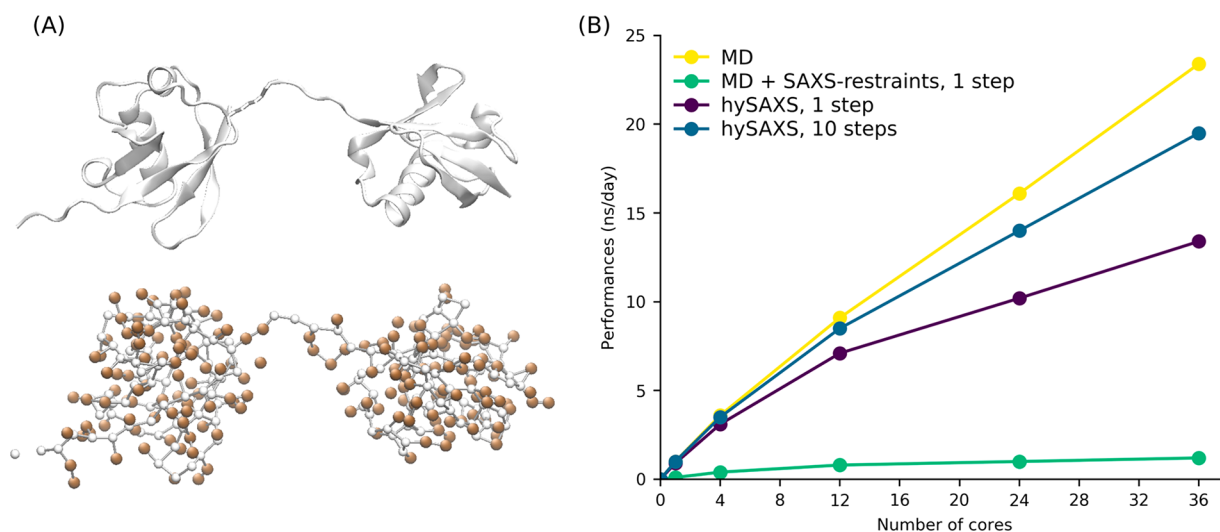
Figure 1. (A) K63-Ub$_2$ (2473 atoms), shown as a cartoon representation (top) or highlighting the centers of the 328 Martini beads (bottom), colored in white and orange for backbone and side chain, respectively. Figures were created with VMD software.[66] (B) Performances, as a function of the number of cores, estimated on Intel Xeon E5-2697 2.30 GHz for a single replica of K63-Ub$_2$ in water.

amino acid or coarser schemes could be useful to efficiently simulate large biological system, at the cost of a resolution loss and strongly limiting the range of scattering angles, biomolecular system containing few hundreds of amino acids clearly benefit from the use of less-approximated approaches. In this work, we chose a bead scheme based on Martini force field[22] (with beads containing ~4 non-hydrogen atoms), which was previously shown to represent an optimal compromise between efficiency and accuracy, allowing scattering calculations 50 times faster than the atomistic ones and with good accuracy for the range of interest of SAXS profiles.[50] Importantly, form factors for Martini beads are available and were previously implemented in the PLUMED-ISDB module.[19,42,50]

Recently, we have implemented a hybrid multiresolution strategy to perform full atomistic MD simulations in which SAXS intensities, computed at a coarse-grain level based on the Martini force field, are used as restraints within the metainference framework[19] (see Figure 1A). The virtual positions of the Martini beads are computed on-the-fly and are used in combination with Martini form factors[50] for SAXS calculations. The computational efficiency of this strategy can be further improved using a multiple time-step protocol, where the metainference bias is applied only every few time steps.[51] In our previous work, we demonstrated the reliability of the hybrid resolution approach for single-replica simulations in which two protein-nucleic acids complexes were refined against SAXS data. Here, we extended the described approach to multireplica M&M simulations, with the aim to exhaustively explore the conformational space of flexible biomolecules that are able to populate multiple conformational states.

2.3. Computational Details of the Simulations. K63-Ub$_2$, for which both SAXS and PRE experimental data are available,[32,40] was used as a test system. As a starting model for the simulations, we used the B and C chains of PDB 2ZNV:[37] since the SAXS data we used to restrain the simulation were acquired on K63-Ub$_2$ with distal-K63R mutation and proximal-D77 addition, we introduced these same modifications in our model. MD simulations were performed with GROMACS 2018,[52] PLUMED 2,[43] and the PLUMED-ISDB[42] module, using the Amber ff03w force field[53] with TIP4P/2005 water

model[54] and scaled protein−water Lennard-Jones parameters (amber03ws).[55] The choice of this force field, which was specifically designed to increase molecules solvation, avoiding collapsed states and nonspecific protein−protein interactions, was guided by the fact that we expect an equilibrium between open and compact states of K63-Ub$_2$ with only transient interdomain contacts. The system was solvated in a periodic dodecahedron box, initially 1.2 nm larger than the protein in each direction, and neutralized. After an initial energy minimization to a maximum force of 100 kJ/mol/nm, the solute was equilibrated under NVT conditions at a temperature of 300 K for 50 ps using the Berendsen thermostat;[56] then, the Berendsen barostat was used to equilibrate the system in the NPT ensemble to a target pressure of 1 atm for 200 ps. The equilibration phase was followed by an initial MD simulation of 100 ns, from which a pool of well-equilibrated conformations was extracted to be used as starting models for the subsequent runs. During the production runs in the NPT ensemble, the MD integrator was employed with a time step of 2 fs; the temperature was maintained at 300 K using the Bussi thermostat,[57] and the pressure was controlled with Parrinello−Rahman barostat.[58] Bonds were constrained with the LINCS algorithm,[59] using a matrix expansion of order 6 and 2 iterations per step. The electrostatic interaction was treated by using the particle mesh Ewald scheme[60] with a short-range cutoff of 0.9 nm and a Fourier grid spacing of 0.12 nm; the van der Waals interaction cutoff was set to 0.9 nm.

Two metadynamics multireplica simulations were performed: (1) a metainference simulation, consisting of 32 replicas, in which metainference was used to enforce the agreement with SAXS data according to the hybrid approach (hySAXS); and (2) an unrestrained simulation, consisting of 8 replicas, in which similar settings of simulation (1) were used but without the inclusion of experimental restraints. PBMetaD was performed in combination with well-tempered metadynamics[61] and the multiple-walker scheme,[49] where Gaussians with an initial height of 1.0 kJ/mol were deposited every 0.4 ps, using a bias factor of 30. Four CVs were biased: two of them (hydContacts and polContacts) count the number of hydrophobic and polar contacts between the two ubiquitin domains, and the other two (TICAcv1 and TICAcv2) are the results of
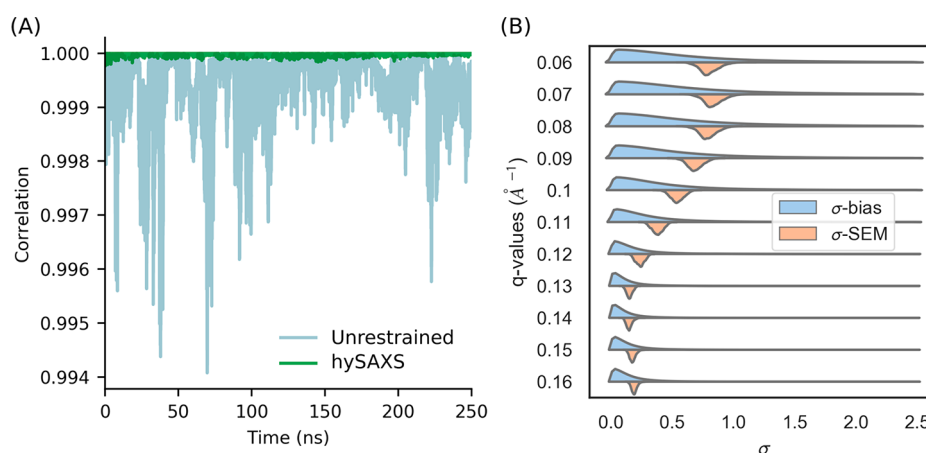
Figure 2. (A) Correlation, as a function of the simulation time, between experimental and back-calculated SAXS intensities, averaged over the replicas. The intensities considered are the ones used as restraints in the hySAXS simulation. (B) Probability density functions of the uncertainty parameters σ^{Bias} and σ^{SEM} (expressed in a.u.) for the 11 scattering angles considered.

the linear combination of numerous angles as determined by a Time-lagged Independent Component Analysis[62] (TICA) performed on the initial 100 ns MD simulation (see the Supporting Information for more details). The width of the Gaussians was determined with the dynamically adapted Gaussian approach,[63] using a time window of 4 ps to estimate CVs fluctuations and setting as minimum values for the widths 0.01, 0.05, 0.01, and 0.01 for hydContacts, polContacts, TICAcv1, and TICAcv2, respectively.

Experimental SAXS intensities for K63-Ub$_2$ are available in the SASDCG7[40] entry of the SASDB database.[64] For the hySAXS simulation, a set of 11 representative SAXS intensities at different scattering vectors, ranging between 0.06 Å^{-1} and 0.16 Å^{-1} and equally spaced, were included as restraints. The range of scattering vectors was selected based on the quality of the data, focusing on the less-noisy region of the experimental curve. We also note that the approximations of the SAXS forward model (i.e., the coarse-grain representation and the lack of hydration layer in the calculations) are more severe if wider angles also are considered, thus setting an upper limit to the possible range to be considered. These representative intensities were extracted from the experimental data, where a 21-point running average was performed to reduce the influence of experimental noise. Metainference was applied every 10 steps, using a single Gaussian noise per data point and sampling a scaling factor between experimental and calculated SAXS intensities with a flat prior between 0.5 and 1.5.

For the hySAXS simulation, each replica was evolved for 250 ns, resulting in a total simulation time of 8 μs; for the unrestrained simulation, 750 ns per replica were run, for a total of 6 μs. Convergence was assessed using the block analysis procedure, in which free-energy profiles are computed over different blocks of simulations and last, the weighted average error along the free-energy profile is computed as a function of the block length. In Figure S1 in the Supporting Information, the free-energy profiles and the block average analysis are reported, showing that both simulations converged with comparable errors. As a preliminary control, we checked the root-mean-square deviation (RMSD) of the single Ub domains. In both simulations, the Ub domains are well-folded. The comparison of RMSD distribution in the two simulations (Figure S2 in the Supporting Information) showed lower RMSD values for the hySAXS ensemble, with respect to the

unrestrained one: that could be due to the shorter simulation time per replica, as well as a protective effect of the SAXS restraints against some destabilization resulting from the use of amber03ws.

2.4. Back-Calculation of PRE. In order to back-calculate PRE values from the obtained conformational ensembles, we used the following formula:[65]

$$ \text{PRE} = K \left[4\tau + \frac{3\tau}{1 + (\omega\tau)^2} \right] \frac{1}{r^6} $$

where τ is the correlation time, $\omega/(2\pi)$ the proton Larmor frequency, K a value that is dependent on the electron g-factor, the proton gyromagnetic ratio, and the magnetic moment of the free electron. Finally, r indicates the distance between the paramagnetic center and the nuclei. In the back-calculation, these distances were approximated with the distances between the Cβ atom of N25 or K48 and all of the amide hydrogens of the proximal ubiquitin. To account for this approximation, we evaluated an error of ±3 Å on the estimation of these distances, which finally gave us an estimation of the minimum/maximum PRE values.

3. RESULTS AND DISCUSSION

To evaluate our hySAXS approach, after assessing its computational performances in comparison with conventional MD simulations and atomistic SAXS restraints, we tested its ability to improve the agreement of MD with experimental SAXS data in comparison with a state-of-the-art force-field (unrestrained simulations). As a model system, we employed K63-Ub$_2$, for which independent data are available to validate our results.

3.1. The hySAXS Approach Is Computationally Efficient. In Figure 1B, we compared the performances of (i) a conventional atomistic MD simulation (yellow); (ii) all-atom metainference simulations, where SAXS restraints with atomistic forward model were included every step (green); and (iii) all-atom hySAXS simulations, where SAXS restraints were included every step (purple) or every 10 steps (blue). The use of the hybrid approach significantly improved the performances of SAXS-driven MD simulations, compared to the ones adopting atomistic scattering evaluation. This gain can be further increased using a multiple time-step protocol (Figure
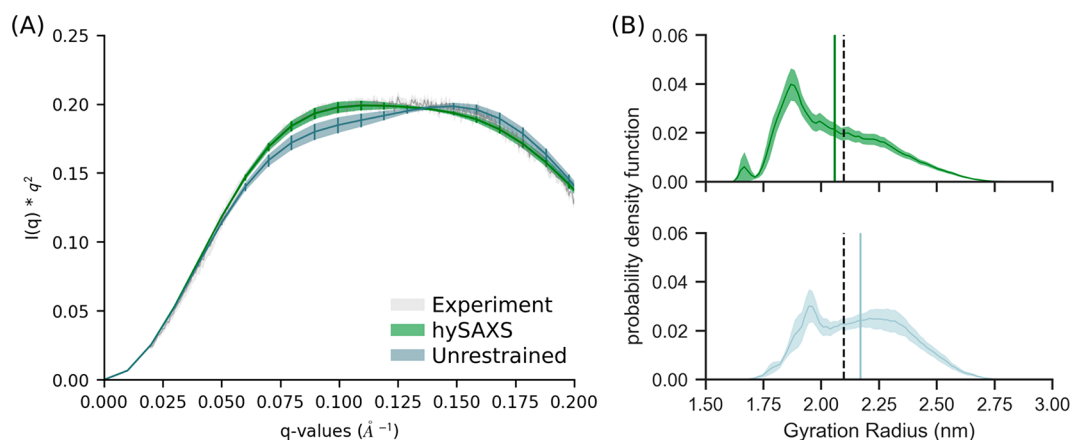
Figure 3. (A) Kratky plot comparing the experimental curve with the ones calculated (via atomistic approach) from the hySAXS and the unrestrained conformational ensembles. (B) Distribution of the gyration radius in the hySAXS (green) and in the unrestrained (light blue) conformational ensembles. Gyration radius were calculated using GROMACS. The vertical bars indicate the average back-calculated gyration radius, the shaded area indicates the standard error, computed via block-average analysis. The vertical black dashed bar indicates the SAXS-derived gyration radius.

1B, blue line), in which the restraint is applied every few time steps. This strategy is well justified in the case of SAXS data, which are characterized by slow temporal fluctuations, and allows one to approach the performances of conventional MD simulations.

3.2. Monitoring hySAXS Simulation. To evaluate, on the fly, the effectiveness of SAXS restraints, we monitored the correlation between back-calculated and experimental data, as a function of the simulation time (Figure 2A), comparing hySAXS to an unrestrained simulation. The comparison revealed a better agreement in the hySAXS simulation (Figure 2A), confirming the efficacy of the restraints. This is supported by other statistical properties, including the sum of square deviation and the slope/intercept of the linear fit (see Figure S3 in the Supporting Information).

We also monitored the intensity of experimental restraints, which depends on the square sum of the uncertainty parameters $\sigma_{r,i}^{\text{Bias}}$ and $\sigma_{r,i}^{\text{SEM}}$ (cf. see earlier section entitled Theory and Methods). To this aim, we computed the distribution across the hySAXS ensemble of both $\sigma_{r,i}^{\text{Bias}}$, which is associated with experimental and forward model inaccuracies, and $\sigma_{r,i}^{\text{SEM}}$, i.e., the standard error of the mean over the replicas. We observed a broader distribution of the sampled parameter $\sigma_{r,i}^{\text{Bias}}$, with respect to $\sigma_{r,i}^{\text{SEM}}$ (Figure 2B), with greater uncertainties associated with smaller scattering angles (where, indeed, the global conformation primarily influences SAXS profiles). The values of $\sigma_{r,i}^{\text{SEM}}$ are always within the range sampled by $\sigma_{r,i}^{\text{Bias}}$, indicating that the two sources of error comparably contribute to the restraint weight and suggesting that the number of replicas (which concurs in determining the magnitude of $\sigma_{r,i}^{\text{SEM}}$) is sufficient.

3.3. Comparison of the Resulting Conformational Ensembles. The agreement with experimental SAXS data was eventually evaluated considering the entire conformational ensembles sampled within the unrestrained or hySAXS simulations. To this aim, we needed to estimate a scaling factor λ that relates experimental and calculated data. This value could, in principle, be determined by comparing the intensities at the $q = 0$ scattering angle, but since $I(0)$ cannot be measured in SAXS experiments, we chose the λ that minimizes the χ^2 (computed over 19 q-values in the range of 0.02–0.20 Å^{-1}) between hySAXS and experimental intensities.

We observed that hySAXS provides a better match with experimental data (as confirmed by the χ^2 values: 0.44 and 3.6 for the hySAXS and unrestrained simulations, respectively), while the unrestrained ensemble strongly deviates from the experimental profile, showing a shape that is indicative of an oversampling of extended conformation (see Figure 3A, as well as Figure S4 in the Supporting Information). Importantly, our validation is conducted over a range of scattering angles (0.02–0.20 Å^{-1}) larger than the one used for restraining the simulation, and our conclusions are not dependent on the choice of the scaling factor: indeed, hySAXS simulation provides a better agreement with experiments also when choosing a λ value that minimizes the χ^2 value of unrestrained intensities (see Figure S5 in the Supporting Information).

Accordingly, we noticed a remarkable effect of SAXS restraints on the interdomain dynamics, as shown by the comparison of the probability density function of the gyration radius and the minimum interdomains distance (see Figure 3B, as well as Figure S6 in the Supporting Information). Both of the ensembles populate a wide range of gyration radius values (spanning from 1.5 nm to 3.0 nm), in agreement with the observation that K63-Ub$_2$ exists in a dynamic ensemble comprising both extended and compact states. Nevertheless, the hySAXS ensemble prefers more-compact conformations, resulting in an average gyration radius of 2.06 ± 0.03 nm, in contrast with that obtained for the unrestrained ensemble (2.17 ± 0.05 nm). The value obtained from the hySAXS ensemble better approaches the SAXS-derived gyration radius (2.10 ± 0.01 nm), with a small difference that could be explained by the contribution of the hydration layer.[67] Altogether, our results support the idea that, while the hySAXS ensemble better reproduces the correct balance between compact and open states, the unrestrained ensemble over-estimates the population of extended conformations. Importantly, here, we showed that SAXS restraints could be effectively used to contain this trend.

The propensity toward extended states for the unrestrained ensemble could be explained by the use of the amber03ws, which is a force field that was specifically designed to prevent the overstabilization of IDPs compact states by strengthening water–protein interaction terms. Our results, which are consistent with a previous report,[68] suggest that this
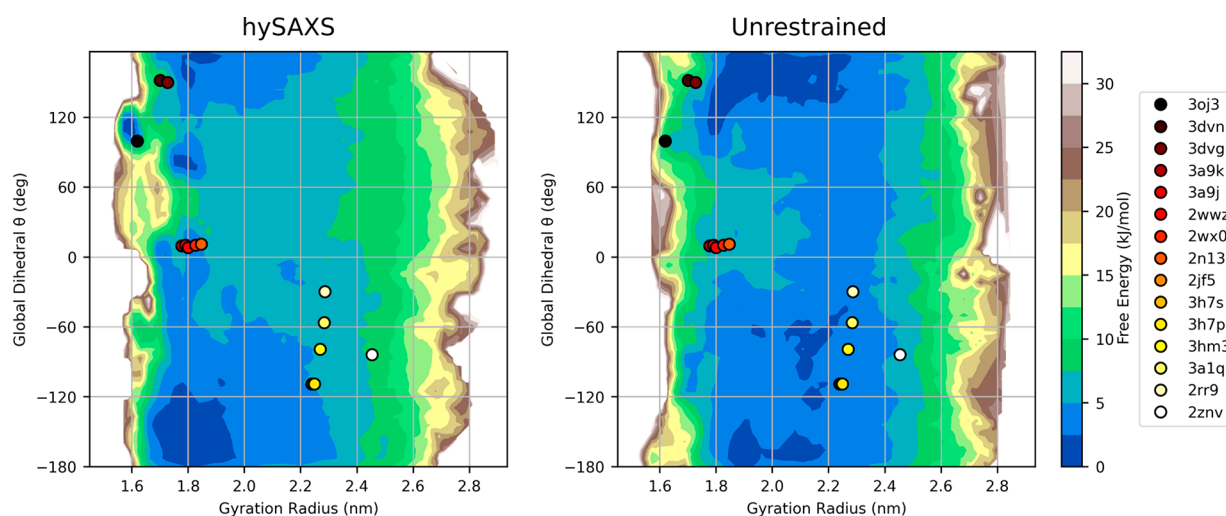
Figure 4. Two-dimensional free-energy surface for K63-Ub$_2$, derived by the hySAXS (left panel) and the unrestrained (right panel) ensembles, as a function of the Cα-gyration radius and the global dihedral angle θ (see section S1 and Figure S7 in the Supporting Information). The coordinates of the available PDB structures in this space are plotted with points (colored from red to white, from more compact to extended conformations). To make the gyration radius comparable between simulations and PDB structures, only Cα atoms of residues 1−72 of the two ubiquitin domains were considered for gyration radius calculations.
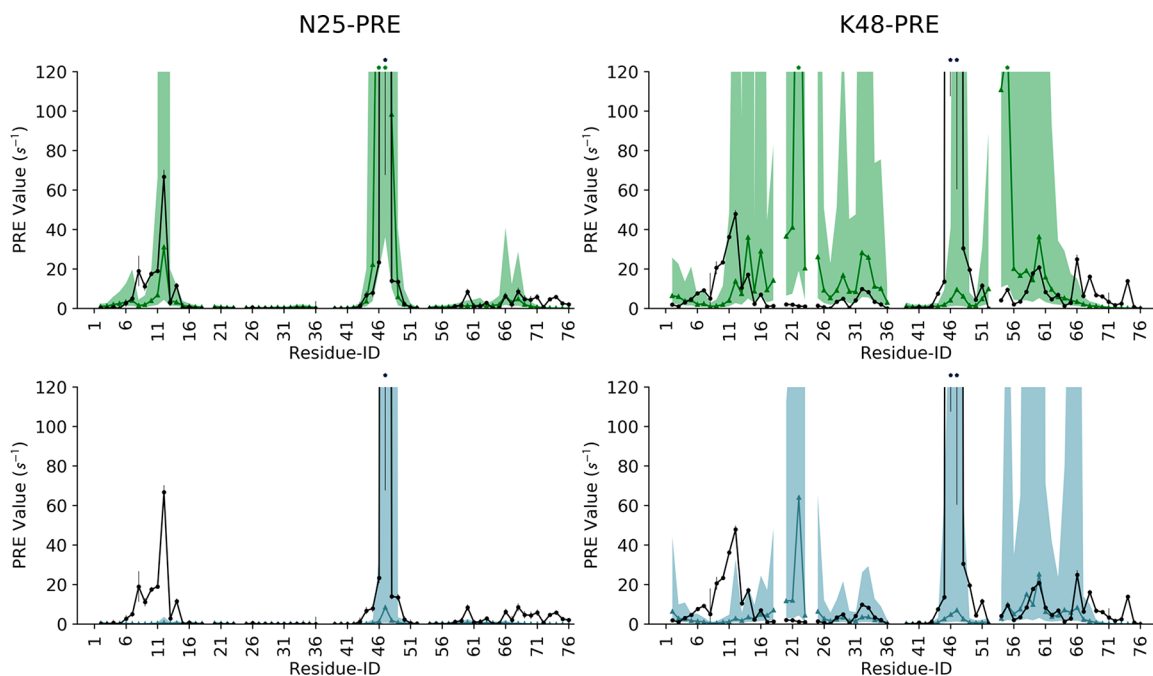

Figure 5. Comparison of experimental (black line) and back-calculated intersubunit PRE for the residues of K63-Ub$_2$ proximal ubiquitin, with the paramagnetic probe conjugated at N25C (left panels) or K48C (right panels) of the distal ubiquitin. The area between the minimum/maximum back-calculated PRE values, considering a ±3 Å error on the estimation of probe−nuclei distances, is shaded green or light blue, for hySAXS and unrestrained ensembles, respectively. The respective back-calculated PRE, without distance correction, is shown with green and light-blue lines. PRE values of >120 s^{-1} are indicated with an asterisk on the top of the graph.

modification could be too strong for folded and multidomain proteins, leading to the destabilization of compact conformations. Recent huge efforts in force-field development indicate that refinement of few force-field terms (as torsional parameters or water models), while useful in improving the description of either well-folded proteins or IDPs, could be not sufficient to provide an equally accurate description of both.[68,69]

Lastly, in Figure 4, we reconstructed two-dimensional free-energy landscapes in a space defined by the Cα-gyration radius

and a global dihedral angle θ (also used as metadynamics CVs; see Figure S7 in the Supporting Information) that describes the relative orientation of the two ubiquitin domains. Interestingly, the coordinates in this space of the available K63-Ub$_2$ PDB structures mostly fall in regions characterized by low free energy, according to both hySAXS and unrestrained simulations, indirectly supporting the reliability of the employed force field. The inspection of the free energies revealed that, in the two ensembles, the Ub domains can reorient freely when extended but prefer different Ub−Ub
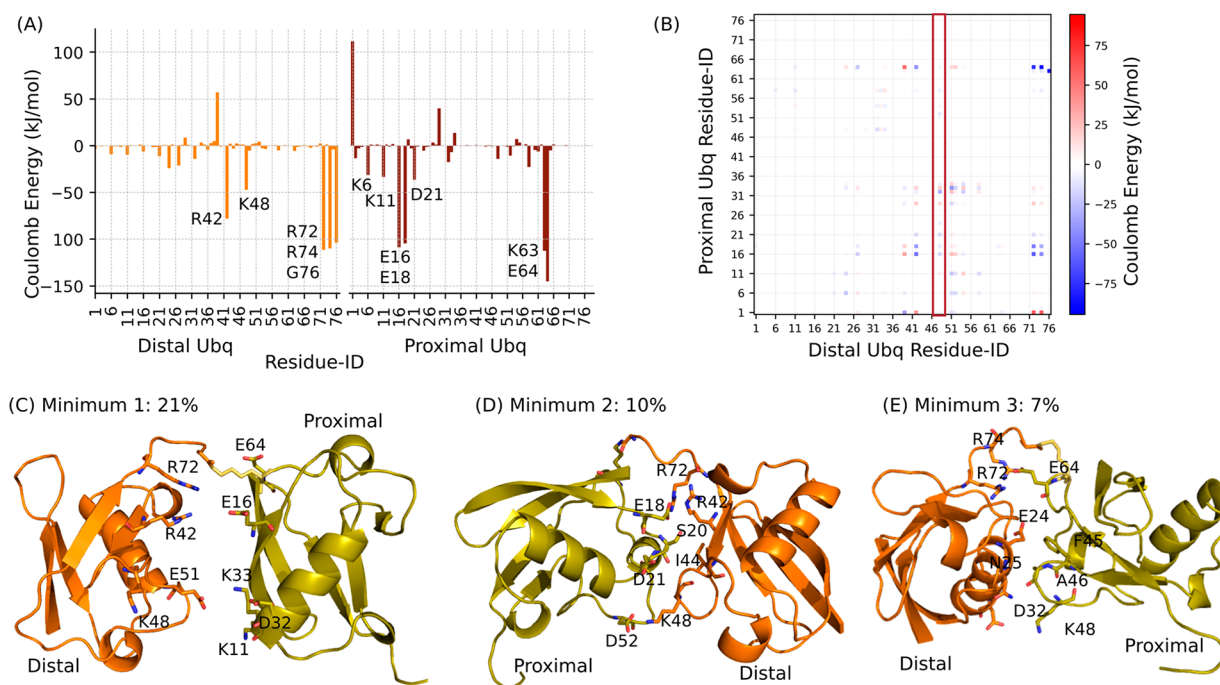
Figure 6. (A) Per-residue Coulomb energy obtained summing over the residue−residue energetic contributions for pairs of residues belonging to the two different Ub domains. Residues of distal and proximal ubiquitin are colored in orange and red, respectively; the lowest energy peaks are labeled. (B) Coulomb energy matrix reporting on the electrostatic interactions between the two domains. The column corresponding to the interactions engaged by distal K48 is highlighted in red. (C−E) Representative conformations have been extracted from the main minima of the compact state. Their population is reported and relevant residues for the interface are highlighted in sticks. Figures were created with Pymol software (The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC).

orientation in compact conformations. The contact map analysis (see Figure S8 in the Supporting Information) confirms the absence of highly stable interdomains contacts, supporting the idea that numerous different interfaces are accessible, and it shows that, in the two simulations, diverse groups of residues are preferred for interdomain interactions, where the major differences concern the residues of distal ubiquitin (the majority of contacts are engaged by distal residues 42−49, according to hySAXS and by residues 6−11, according to the unrestrained simulation). Based on this observation, we hypothesized that SAXS restraints could help in sampling more-reliable protein/protein interfaces. To test this hypothesis, we proceeded by validating our conformational ensemble against PRE data.

3.4. Validation and Analysis of the Ub/Ub Interfaces. PRE experiments from NMR are particularly suited to provide information about intersubunit distances in multidomain proteins. In these experiments, after conjugation of a specific residue with a paramagnetic probe, PRE can be measured for the other domain, where PRE values are proportional to the inverse sixth power of the distance between the paramagnetic center and the nuclei. Becuase of this functional form, PRE data are extremely sensitive to closed states even if sparsely populated.[65] Therefore, a comparison of the conformational ensemble against PRE is particularly indicated to validate the Ub/Ub interfaces of the compact states and their relative population.

Liu and co-workers previously acquired intersubunit PRE data for K63-Ub$_2$, conjugating the paramagnetic probe on residues N25 or K48 of the distal ubiquitin, after N25C/K48C mutations, and detecting many large PRE for some residues of the proximal unit.[32,70] We back-calculated the same PRE values

from our hySAXS and unrestrained conformational ensembles (see the Theory and Methods section) and compared them with those determined from experiments.

We observed that experimental N25-PRE is in good agreement with those calculated from the hySAXS ensemble (Figure 5, upper-left panel), suggesting that the compact interfaces are correctly sampled in our hySAXS run. Conversely, the unrestrained ensemble fails to reproduce N25-PRE for the proximal unit residues 8−14 (Figure 5, lower-left panel). Both the hySAXS and the unrestrained ensembles correctly identify the regions where high K48-PRE are detected (Figure 5, right panels); nevertheless, in both cases, we observe a significant overestimation of the PRE involving the residues 20−23 of the proximal unit.

Since the comparison with both N25 and K48-PRE supports the reliability of our hySAXS ensemble in sampling correct Ub/Ub interfaces, we hypothesized that the observed deviations could arise as a consequence of the introduction of the paramagnetic probe at the K48 site in PRE experiments, along with the K48C mutation. Indeed, while N25C mutation is more conservative, the replacement of a charged amino acid (K48C) could destabilize relevant interdomains contacts. To support this hypothesis, we analyzed the energetic contributions of each residue to the interface formation. We found that, according to the hySAXS ensemble, K48 of proximal ubiquitin is important in stabilizing electrostatic interactions at the interface and that a part of these contacts are indeed engaged with the negatively charged D21 residue of distal ubiquitin, belonging to the region where the major deviations were observed (see Figures 6A and 6B). Importantly, we verified that this is not the case for N25, where neither Coulomb nor Lennard-Jones interactions seem to play a major role in

stabilizing the Ub/Ub interfaces (see Figure 6A, as well as Figure S9 in the Supporting Information).

In order to have a deeper insight into the sampled Ub/Ub interfaces, we analyzed the conformational minima identified by our hySAXS run. The pool of compact conformations (defined as the ones with a Cα-gyration radius of <2.0 nm and accounting for 57% of the conformational space) were clustered, based on the backbone RMSD with a cutoff of 6 Å. This procedure identified three main clusters, with populations of 21%, 10%, and 7%, respectively. As expected, these three conformational minima contain very heterogeneous conformational states (see Figure S10 in the Supporting Information), supporting the idea that K63-Ub2 can transiently populate many different possible interfaces. Nevertheless, the inspection of both their structures and of the corresponding energy matrices (Figure S10) allowed us to characterize the interfaces and the contacts driving the interdomain recognition in greater detail (see Figures 6C−E). We observed that, in all three minima, the positive residues R42, R72, and/or R74 of distal ubiquitin engage electrostatic interactions with negatively charged residues of the proximal domain (mainly E16−E64, E18, and E64 for minima 1, 2, and 3, respectively). In addition to these interactions, further contacts characterize the different minima, again involving mainly charged residues (Figures 6C−E). While, in the most populated minimum (minimum 1), hydrophobic interactions are almost absent, these are present in the other minima: in minimum 2, contacts between distal I44 and proximal S20 are observed, whereas, in minimum 3, the interface is also stabilized by contacts between the aliphatic side chains of distal residues E24−N25 and the proximal F45−A46.

Overall, our analysis revealed the involvement of many charged residues in the Ub/Ub interface and suggests that K63-Ub$_2$ prefer electrostatic interfacial contacts, being hindered by steric constraints to interact via the common I44/I36 hydrophobic patches, which is consistent with previous reports.[71] Our results are in agreement with previous mutagenesis experiments concerning the E64 residue of the proximal unit, which plays a major role in both minima 1 and 3 interfaces. Indeed, it was reported that E64 is important for the stabilization of closed conformations, where an E64R mutation was shown to decrease the binding affinity toward ligands, known to bind the K63-Ub$_2$ closed states, via an entropically driven mechanism. Herein, our results support the conformational selection mechanism proposed by Liu and co-workers[32] for K63-Ub$_2$ ligand recognition.

4. CONCLUSIONS

In this work, we have presented a hybrid-resolution MD-based strategy, which is useful with regard to determining conformational ensembles that provide an accurate interpretation of SAXS data. The proposed approach makes the inclusion of SAXS in MD simulations feasible, in terms of computational efficiency, without losing atomistic details, and allows us to deal with highly flexible systems, aiding in the estimation of the population of the different existing conformational states.

To prove the efficacy of the method, here, it has been applied to study the conformational ensemble of the multidomain protein K63-Ub$_2$. Our results reveal that the inclusion of SAXS restraints can significantly influence the relative positioning of the different subunits and the degree of protein extension, improving the reliability of the conforma-

tional sampling, as supported by indirect validations and by quantitative comparison with independent experimental data.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.9b01181.

Description of metadynamics Collective Variables (section S1). Figures: 1. Simulation convergence (Figure S1); RMSD pdf of the single ubiquitin domains (Figure S2); statistical parameters to monitor hySAXS simulation (Figure S3); SAXS curves with residual plot (Figure S4); Kratky plot (Figure S5); interdomain minimum distance probability distribution (Figure S6); graphical representation of the global dihedral angle θ (Figure S7); contact map of the hySAXS and unrestrained ensemble (Figure S8); per-residue Lennard-Jones interface contribution (Figure S9); and energy matrices and conformational ensembles for the hySAXS compact minima (Figure S10) (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Cristina Paissoni − *Dipartimento di Bioscienze, Università degli Studi di Milano, 20133 Milano, Italy;*
Email: cristina.paissoni@unimi.it

Carlo Camilloni − *Dipartimento di Bioscienze, Università degli Studi di Milano, 20133 Milano, Italy;* Ⓞ orcid.org/0000-0002-9923-8590; Email: carlo.camilloni@unimi.it

Author

Alexander Jussupow − *Department of Chemistry and Institute of Advanced Study, Technical University of Munich, Garching 85747, Germany*

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.jctc.9b01181

Notes

The authors declare no competing financial interest.

■ ABBREVIATIONS

K63-Ub$_2$, K63-linked diubiquitin; MD, molecular dynamics; M&M, metadynamics metainference; PBMetaD, parallel bias metadynamics; CV, collective variable; SAXS, small-angle X-ray scattering; PRE, paramagnetic relaxation enhancement; NMR, nuclear magnetic resonance; FRET, Förster resonance energy transfer; TICA, time-lagged independent component analysis

■ REFERENCES

(1) Bonomi, M.; Heller, G. T.; Camilloni, C.; Vendruscolo, M. Principles of Protein Structural Ensemble Determination. *Curr. Opin. Struct. Biol.* **2017**, *42*, 106−116.

(2) Wei, G.; Xi, W.; Nussinov, R.; Ma, B. Protein Ensembles: How Does Nature Harness Thermodynamic Fluctuations for Life? The Diverse Functional Roles of Conformational Ensembles in the Cell. *Chem. Rev.* **2016**, *116*, 6516−6551.

(3) Henzler-Wildman, K.; Kern, D. Dynamic Personalities of Proteins. *Nature* **2007**, *450*, 964−972.

(4) Papaleo, E.; Saladino, G.; Lambrughi, M.; Lindorff-Larsen, K.; Gervasio, F. L.; Nussinov, R. The Role of Protein Loops and Linkers in Conformational Dynamics and Allostery. *Chem. Rev.* **2016**, *116* (11), 6391−6423.

(5) Henzler-Wildman, K.; Kern, D. Dynamic Personalities of Proteins. *Nature* **2007**, *450*, 964−972.

(6) Hub, J. S. Interpreting Solution X-Ray Scattering Data Using Molecular Simulations. *Curr. Opin. Struct. Biol.* **2018**, *49*, 18−26.

(7) Bottaro, S.; Lindorff-Larsen, K. Biophysical Experiments and Biomolecular Simulations: A Perfect Match? *Science* **2018**, *361*, 355−360.

(8) Shevchuk, R.; Hub, J. S. Bayesian Refinement of Protein Structures and Ensembles against SAXS Data Using Molecular Dynamics. *PLoS Comput. Biol.* **2017**, *13* (10), e1005800.

(9) Antonov, L. D.; Olsson, S.; Boomsma, W.; Hamelryck, T. Bayesian Inference of Protein Ensembles from SAXS Data. *Phys. Chem. Chem. Phys.* **2016**, *18* (8), 5832−5838.

(10) Bowerman, S.; Rana, A. S. J. B.; Rice, A.; Pham, G. H.; Strieter, E. R.; Wereszczynski, J. Determining Atomistic SAXS Models of Tri-Ubiquitin Chains from Bayesian Analysis of Accelerated Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2017**, *13* (6), 2418−2429.

(11) Cheng, P.; Peng, J.; Zhang, Z. SAXS-Oriented Ensemble Refinement of Flexible Biomolecules. *Biophys. J.* **2017**, *112* (7), 1295−1301.

(12) Rózycki, B.; Kim, Y. C.; Hummer, G. SAXS Ensemble Refinement of ESCRT-III CHMP3 Conformational Transitions. *Structure* **2011**, *19* (1), 109−116.

(13) Hummer, G.; Köfinger, J. Bayesian Ensemble Refinement by Replica Simulations and Reweighting. *J. Chem. Phys.* **2015**, *143* (24), 243150.

(14) Bottaro, S.; Bengtsen, T.; Lindorff-Larsen, K. Integrating Molecular Simulation and Experimental Data: A Bayesian/Maximum Entropy Reweighting Approach. *bioRxiv* **2018**, DOI: 10.1101/457952.

(15) Björling, A.; Niebling, S.; Marcellini, M.; Van Der Spoel, D.; Westenhoff, S. Deciphering Solution Scattering Data with Experimentally Guided Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2015**, *11* (2), 780−787.

(16) Chen, P. C.; Hub, J. S. Interpretation of Solution X-ray Scattering by Explicit-Solvent Molecular Dynamics. *Biophys. J.* **2015**, *108* (10), 2573−2584.

(17) Kimanius, D.; Pettersson, I.; Schluckebier, G.; Lindahl, E.; Andersson, M. SAXS-Guided Metadynamics. *J. Chem. Theory Comput.* **2015**, *11* (7), 3491−3498.

(18) Hermann, M. R.; Hub, J. S. SAXS-Restrained Ensemble Simulations of Intrinsically Disordered Proteins with Commitment to the Principle of Maximum Entropy. *J. Chem. Theory Comput.* **2019**, *15* (9), 5103−5115.

(19) Paissoni, C.; Jussupow, A.; Camilloni, C. Martini Bead Form Factors for Nucleic Acids and Their Application in the Refinement of Protein−Nucleic Acid Complexes against SAXS Data. *J. Appl. Crystallogr.* **2019**, *52* (2), 394−402.

(20) Bonomi, M.; Camilloni, C.; Cavalli, A.; Vendruscolo, M. Metainference: A Bayesian Inference Method for Heterogeneous Systems. *Sci. Adv.* **2016**, *2* (1), e1501177.

(21) Yang, S.; Park, S.; Makowski, L.; Roux, B. A Rapid Coarse Residue-Based Computational Method for X-Ray Solution Scattering Characterization of Protein Folds and Multiple Conformational States of Large Protein Complexes. *Biophys. J.* **2009**, *96* (11), 4449−4463.

(22) Marrink, S. J.; Tieleman, D. P. Perspective on the Martini Model. *Chem. Soc. Rev.* **2013**, *42* (16), 6801−6822.

(23) Bonomi, M.; Camilloni, C.; Vendruscolo, M. Metadynamic Metainference: Enhanced Sampling of the Metainference Ensemble Using Metadynamics. *Sci. Rep.* **2016**, *6* (1), 31232.

(24) Laio, A.; Parrinello, M. Escaping Free-Energy Minima. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99* (20), 12562.

(25) Löhr, T.; Jussupow, A.; Camilloni, C. Metadynamic Metainference: Convergence towards Force Field Independent Structural Ensembles of a Disordered Peptide. *J. Chem. Phys.* **2017**, *146* (16), 165102.

(26) Komander, D.; Rape, M. The Ubiquitin Code. *Annu. Rev. Biochem.* **2012**, *81* (1), 203−229.

(27) Vincendeau, M.; Hadian, K.; Messias, A. C.; Brenke, J. K.; Halander, J.; Griesbach, R.; Greczmiel, U.; Bertossi, A.; Stehle, R.; Nagel, D.; et al. Inhibition of Canonical NF-κ B Signaling by a Small Molecule Targeting NEMO-Ubiquitin Interaction. *Sci. Rep.* **2016**, *6* (1), 18934.

(28) Castañeda, C. A.; Chaturvedi, A.; Camara, C. M.; Curtis, J. E.; Krueger, S.; Fushman, D. Linkage-Specific Conformational Ensembles of Non-Canonical Polyubiquitin Chains. *Phys. Chem. Chem. Phys.* **2016**, *18* (8), 5771−5788.

(29) Castañeda, C. A.; Kashyap, T. R.; Nakasone, M. A.; Krueger, S.; Fushman, D. Unique Structural, Dynamical, and Functional Properties of K11-Linked Polyubiquitin Chains. *Structure* **2013**, *21* (7), 1168−1181.

(30) Berlin, K.; Castañeda, C. A.; Schneidman-Duhovny, D.; Sali, A.; Nava-Tudela, A.; Fushman, D. Recovering a Representative Conformational Ensemble from Underdetermined Macromolecular Structural Data. *J. Am. Chem. Soc.* **2013**, *135* (44), 16595−16609.

(31) Ye, Y.; Blaser, G.; Horrocks, M. H.; Ruedas-Rama, M. J.; Ibrahim, S.; Zhukov, A. A.; Orte, A.; Klenerman, D.; Jackson, S. E.; Komander, D. Ubiquitin Chain Conformation Regulates Recognition and Activity of Interacting Proteins. *Nature* **2012**, *492* (7428), 266−270.

(32) Liu, Z.; Gong, Z.; Jiang, W.-X.; Yang, J.; Zhu, W.-K.; Guo, D.-C.; Zhang, W.-P.; Liu, M.-L.; Tang, C. Lys63-Linked Ubiquitin Chain Adopts Multiple Conformational States for Specific Target Recognition. *eLife* **2015**, *4*, e05767.

(33) Kulathu, Y.; Akutsu, M.; Bremm, A.; Hofmann, K.; Komander, D. Two-Sided Ubiquitin Binding Explains Specificity of the TAB2 NZF Domain. *Nat. Struct. Mol. Biol.* **2009**, *16* (12), 1328−1330.

(34) Komander, D.; Reyes-Turcu, F.; Licchesi, J. D. F.; Odenwaelder, P.; Wilkinson, K. D.; Barford, D. Molecular Discrimination of Structurally Equivalent Lys 63-Linked and Linear Polyubiquitin Chains. *EMBO Rep.* **2009**, *10* (5), 466−473.

(35) Weeks, S. D.; Grasty, K. C.; Hernandez-Cuebas, L.; Loll, P. J. Crystal Structures of Lys-63-Linked Tri- and Di-Ubiquitin Reveal a Highly Extended Chain Architecture. *Proteins: Struct., Funct., Genet.* **2009**, *77* (4), 753−759.

(36) Sato, Y.; Yoshikawa, A.; Yamashita, M.; Yamagata, A.; Fukai, S. Structural Basis for Specific Recognition of Lys 63-Linked Polyubiquitin Chains by NZF Domains of TAB2 and TAB3. *EMBO J.* **2009**, *28* (24), 3903−3909.

(37) Sato, Y.; Yoshikawa, A.; Yamagata, A.; Mimura, H.; Yamashita, M.; Ookata, K.; Nureki, O.; Iwai, K.; Komada, M.; Fukai, S. Structural Basis for Specific Cleavage of Lys 63-Linked Polyubiquitin Chains. *Nature* **2008**, *455* (7211), 358−362.

(38) Sato, Y.; Yoshikawa, A.; Mimura, H.; Yamashita, M.; Yamagata, A.; Fukai, S. Structural Basis for Specific Recognition of Lys 63-Linked Polyubiquitin Chains by Tandem UIMs of RAP80. *EMBO J.* **2009**, *28* (16), 2461−2468.

(39) Newton, K.; Matsumoto, M. L.; Wertz, I. E.; Kirkpatrick, D. S.; Lill, J. R.; Tan, J.; Dugger, D.; Gordon, N.; Sidhu, S. S.; Fellouse, F. A.; et al. Ubiquitin Chain Editing Revealed by Polyubiquitin Linkage-Specific Antibodies. *Cell* **2008**, *134* (4), 668−678.

(40) Liu, Z.; Gong, Z.; Cao, Y.; Ding, Y. H.; Dong, M. Q.; Lu, Y. B.; Zhang, W. P.; Tang, C. Characterizing Protein Dynamics with Integrative Use of Bulk and Single-Molecule Techniques. *Biochemistry* **2018**, *57* (3), 305−313.

(41) Camilloni, C.; Pietrucci, F. Advanced Simulation Techniques for the Thermodynamic and Kinetic Characterization of Biological Systems. *Advances in Physics: X*, Vol. 3; Taylor & Francis: New York, 2018; pp 885−916.

(42) Bonomi, M.; Camilloni, C. Integrative Structural and Dynamical Biology with PLUMED-ISDB. *Bioinformatics* **2017**, *33* (24), 3999−4000.

(43) Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. PLUMED 2: New Feathers for an Old Bird. *Comput. Phys. Commun.* **2014**, *185* (2), 604−613.

(44) Bonomi, M.; Bussi, G.; Camilloni, C. et al. (The PLUMED Consortium). Promoting Transparency and Reproducibility in Enhanced Molecular Simulations *Nat. Methods* **2019**, *16*, 670−673

(45) Rieping, W.; Habeck, M.; Nilges, M. Inferential Structure Determination. *Science (Washington, DC, U. S.)* **2005**, *309* (5732), 303−306.

(46) Cavalli, A.; Camilloni, C.; Vendruscolo, M. Molecular Dynamics Simulations with Replica-Averaged Structural Restraints Generate Structural Ensembles According to the Maximum Entropy Principle. *J. Chem. Phys.* **2013**, *138* (9), 094112.

(47) Laio, A.; Parrinello, M. Escaping Free-Energy Minima. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99* (20), 12562−12566.

(48) Pfaendtner, J.; Bonomi, M. Efficient Sampling of High-Dimensional Free-Energy Landscapes with Parallel Bias Metadynamics. *J. Chem. Theory Comput.* **2015**, *11* (11), 5062−5067.

(49) Raiteri, P.; Laio, A.; Gervasio, F. L.; Micheletti, C.; Parrinello, M. Efficient Reconstruction of Complex Free Energy Landscapes by Multiple Walkers Metadynamics. *J. Phys. Chem. B* **2006**, *110* (8), 3533−3539.

(50) Niebling, S.; Björling, A.; Westenhoff, S. MARTINI Bead Form Factors for the Analysis of Time-Resolved X-Ray Scattering of Proteins. *J. Appl. Crystallogr.* **2014**, *47* (4), 1190−1198.

(51) Ferrarotti, M. J.; Bottaro, S.; Pérez-Villa, A.; Bussi, G. Accurate Multiple Time Step in Biased Molecular Simulations. *J. Chem. Theory Comput.* **2015**, *11* (1), 139−146.

(52) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, *1−2*, 19−25.

(53) Best, R. B.; Mittal, J. Protein Simulations with an Optimized Water Model: Cooperative Helix Formation and Temperature-Induced Unfolded State Collapse. *J. Phys. Chem. B* **2010**, *114* (46), 14916−14923.

(54) Abascal, J. L. F.; Vega, C. A General Purpose Model for the Condensed Phases of Water: TIP4P/2005. *J. Chem. Phys.* **2005**, *123* (23), 234505.

(55) Best, R. B.; Zheng, W.; Mittal, J. Balanced Protein-Water Interactions Improve Properties of Disordered Proteins and Non-Specific Protein Association. *J. Chem. Theory Comput.* **2014**, *10* (11), 5113−5124.

(56) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular Dynamics with Coupling to an External Bath. *J. Chem. Phys.* **1984**, *81* (8), 3684−3690.

(57) Bussi, G.; Donadio, D.; Parrinello, M. Canonical Sampling through Velocity Rescaling. *J. Chem. Phys.* **2007**, *126* (1), 014101.

(58) Parrinello, M.; Rahman, A. Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method. *J. Appl. Phys.* **1981**, *52* (12), 7182−7190.

(59) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. LINCS: A Linear Constraint Solver for Molecular Simulations. *J. Comput. Chem.* **1997**, *18* (12), 1463−1472.

(60) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An N·log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98* (12), 10089−10092.

(61) Barducci, A.; Bussi, G.; Parrinello, M. Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. *Phys. Rev. Lett.* **2008**, *100* (2), 020603.

(62) McCarty, J.; Parrinello, M. A Variational Conformational Dynamics Approach to the Selection of Collective Variables in Metadynamics. *J. Chem. Phys.* **2017**, *147* (20), 204109.

(63) Branduardi, D.; Bussi, G.; Parrinello, M. Metadynamics with Adaptive Gaussians. *J. Chem. Theory Comput.* **2012**, *8* (7), 2247−2254.

(64) Valentini, E.; Kikhney, A. G.; Previtali, G.; Jeffries, C. M.; Svergun, D. I. SASBDB, a Repository for Biological Small-Angle Scattering Data. *Nucleic Acids Res.* **2015**, *43* (D1), D357−D363.

(65) Clore, G. M.; Iwahara, J. Theory, Practice, and Applications of Paramagnetic Relaxation Enhancement for the Characterization of Transient Low-Population States of Biological Macromolecules and Their Complexes. *Chem. Rev.* **2009**, *109* (9), 4108−4139.

(66) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graphics* **1996**, *14* (1), 33−38.

(67) Henriques, J.; Arleth, L.; Lindorff-Larsen, K.; Skepö, M. On the Calculation of SAXS Profiles of Folded and Intrinsically Disordered Proteins from Computer Simulations. *J. Mol. Biol.* **2018**, *430* (16), 2521−2539.

(68) Robustelli, P.; Piana, S.; Shaw, D. E. Developing a Molecular Dynamics Force Field for Both Folded and Disordered Protein States. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115* (21), E4758−E4766.

(69) Piana, S.; Robustelli, P.; Tan, D.; Chen, S.; Shaw, D. E. Development of a Force Field for the Simulation of Single-Chain Proteins and Protein-Protein Complexes. *J. Chem. Theory Comput.* **2020**, DOI: 10.1021/acs.jctc.9b00251.

(70) Liu, Z.; Tang, C. Ensemble Structure Description of Lys63-Linked Diubiquitin. *Data Br.* **2016**, *7*, 81−88.

(71) Wagner, N. D.; Russell, D. H. Defining Noncovalent Ubiquitin Homodimer Interfacial Interactions through Comparisons with Covalently Linked Diubiquitin. *J. Am. Chem. Soc.* **2016**, *138* (51), 16588−16591.

The Journal of Chemical Physics

Metadynamic metainference: Convergence towards force field independent structural ensembles of a disordered peptide

Thomas Löhr, Alexander Jussupow, and 🔟 Carlo Camilloni

View Online **Export Citation** **CrossMark**

ARTICLES YOU MAY BE INTERESTED IN

Bayesian ensemble refinement by replica simulations and reweighting
The Journal of Chemical Physics **143**, 243150 (2015); https://doi.org/10.1063/1.4937786

Enhanced sampling in molecular dynamics
The Journal of Chemical Physics **151**, 070902 (2019); https://doi.org/10.1063/1.5109531

Perspective: Identification of collective variables and metastable states of protein dynamics
The Journal of Chemical Physics **149**, 150901 (2018); https://doi.org/10.1063/1.5049637
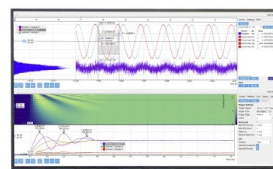
AIP Publishing

Metadynamic metainference: Convergence towards force field independent structural ensembles of a disordered peptide

Thomas Löhr, Alexander Jussupow, and Carlo Camilloni[a)]
Department of Chemistry and Institute for Advanced Study, Technische Universität München,
Lichtenbergstr. 4, 85747 Garching, Germany

Metadynamic metainference has been recently introduced as a theoretical framework to determine structural ensembles by combining and weighting their noise multiple sources of experimental data with molecular mechanics force fields and metadynamics simulations. Here we build upon these initial developments to further extend and streamline the computational approach. We also show that metadynamic metainference can actually determine a structural ensemble for a disordered peptide that is essentially independent from the employed force field. We further show that it is possible to use a very computationally efficient implicit solvent force field in the place of very expensive state-of-the-art explicit solvent ones without a significant loss in accuracy. *Published by AIP Publishing.*
[http://dx.doi.org/10.1063/1.4981211]

I. INTRODUCTION

Structural ensembles of proteins are becoming an invaluable tool to understand biological mechanisms at the molecular level.[1–5] Accurate ensembles can represent the conformational fluctuations of proteins and enable the observation of multiple substates populated by these molecules in given experimental conditions, thus providing a link between structure and function.[6,7] This is of particular importance when considering systems that are intrinsically dynamic, like multi-domain proteins with disordered linkers and, more in general, intrinsically disordered proteins and regions.[8–10]

Determining structural ensembles for such dynamic systems is, however, a challenging task that requires at the same time an accurate modelling and a thorough sampling of the system's conformational space.[1,11,12] Recently to address these challenges, we have introduced Metadynamic Metainference[13,14] (M&M). Metainference[13] is a Bayesian framework that allows integrating multiple sources of information about a system and optimally balancing them. In this approach, one modifies the *a priori* knowledge about a system (i.e., its physicochemical properties as described by molecular mechanics force fields) using data acquired from experimental measurements and balances those by sampling on-the-fly a statistical distribution of noises that can effectively take into account all the sources of errors (i.e., ensemble averaging, statistical errors, systematic errors, and experimental data modelling errors). By combining metainference with metadynamics[15,16] one can then enhance the sampling of the metainference model and explore conformational states that can be separated by significant high free-energy barriers on the time scale of standard molecular dynamics.[14]

While metainference, as well as other statistical methods to determine ensembles,[1,11] can update the prior knowledge to take the available experimental knowledge into account, the question is open about the possibility of obtaining ensembles that do not depend on the specific prior knowledge employed. From a theoretical point of view, the farther the prior is from providing a good description of a system, the more abundant and better the data must be to obtain a good representation of it.[13] In practice, one could be interested in how close the ensembles determined for a disordered system are when employing state-of-the-art molecular mechanic force fields and different sources of experimental data.[1]

In the following we have studied the disordered peptide EGAAWAASS[17] making use of M&M, two state-of-the-art force fields from different families that have given promising results in describing folded and unfolded proteins (CHARMM22*[18] and AMBER99SB with TIP4P-D[19]) and integrated three different sources of commonly available experimental data, NMR chemical shifts (CS), ^{3}J-couplings, and Residual Dipolar Couplings (RDCs). This model system has been employed recently, due to the availability of multiple accurate NMR measures,[17,20] to highlight the deficiencies of force fields in describing disordered systems.[20,21] M&M allowed us to determine structural ensembles for this peptide that are essentially indistinguishable from the point of view of a number of independent parameters. These include actual experimental observables, probability distributions of multiple global degrees of freedom, and secondary structures contents.

Prompted by this observation, we challenged M&M to determine an ensemble using CHARMM36[22] with the EEF1-SB[23,24] implicit solvent model as an extremely computationally inexpensive prior. In the case of the current peptide, we could determine an ensemble of comparable quality with those obtained from the explicit solvent force fields, but at a fraction of the computational cost. This approach could alleviate the computational cost of studying systems that require the use of large simulation boxes and huge amounts of water molecules.

a)Electronic mail: carlo.camilloni@ch.tum.de

II. THEORY AND METHODS

Metainference[13] employs Bayesian statistics to allow updating a prior distribution by considering some new additional information. In particular, metainference is derived to take into account information that is the result of ensemble averaging, i.e., averaging over a full probability distribution. In computational structural biology, the prior is usually a mechanistic force field that describes more or less accurately the interactions of the atoms. Additional information is structural equilibrium measures like those obtained by NMR spectroscopy. Equilibrium observables are always the result of time and ensemble averaging and as such should be employed to update an ensemble and not the probability of observing a single structure. This latter case is a good approximation when the single structure represents by far the most populated state of the system. Notably, metainference is equivalent to inferential structure determination if ensemble averaging is not taken into account.[25]

In M&M[14] multiple simulations, replicas, are performed in parallel for the same system, in the same conditions and using the same force field. These independent simulations are coupled with an energy defined as a function of the difference between the average over the replicas of a back-calculated quantity and a reference value for the same quantity (e.g., the difference between a chemical shift calculated for all the replicas and averaged and the experimental value for that same chemical shift). The sampling of each replica is enhanced by metadynamics[26,27] (in this case by parallel bias metadynamics[28]), which adds a history dependent bias as a function of a set of collective variables (CVs). In M&M, metainference and metadynamics are coupled by the calculation of the average over the replicas. While in standard metainference each replica contributes with the same weight to the average of back-calculated experimental observables, in M&M the weighted average provides a better estimator. This is particularly feasible in the case where the same bias is applied to all replicas in such a way that in the limit of a quasi-static bias the weight of a replica can be approximated on the fly as $w_r \propto \exp(+V^{\mathrm{MetaD}}(\mathrm{CV}(X_r))/k_{\mathrm{B}}T)$.

In the following we first reprise metainference theory, and then we introduce a simple on-the-fly estimate for the only parameter of metainference and introduce a correction to take into account the effect of the weighted average on the distribution of the forces when using metadynamics. Finally we extend metainference to work with data defined but for a multiplicative constant. All the methods described in this work are implemented and available in PLUMED;[29] a working and complete input file that allows reproducing our results and gain a better understanding of M&M is provided in the supplementary material.

A. Metainference

The energy of a metainference simulation is defined as $-k_{\mathrm{B}}T \ln(P)$, where k_{B} is the Boltzmann constant, T the temperature, and P the metainference posterior distribution, calculated over a finite number of replicas, N_r. Its general definition is

$$P(\tilde{\boldsymbol{f}}, \sigma^{\mathbf{B}}, X, \sigma^{\mathbf{SEM}}|\boldsymbol{d})$$
$$= \prod_{r=1}^{N_r} p(X_r) \prod_{i=1}^{N_d} p(d_i|\tilde{f}_i, \sigma_{r,i}^{\mathrm{B}}) p(\tilde{f}_i|X, \sigma_i^{\mathrm{SEM}}) p(\sigma_{r,i}^{\mathrm{B}}) p(\sigma_i^{\mathrm{SEM}}),$$
(1)

where $p(d_i|\tilde{f}_i, \sigma_{r,i}^{\mathrm{B}})$ is the likelihood of the experimental data d_i given $\tilde{f}_i$ and a vector of $\sigma_{r,i}^{\mathrm{B}}$. $\tilde{f}_i$ is the average of the forward model f_i used to predict the experimental observable i from a model calculated on an infinite number of replicas and $\sigma_{r,i}^{\mathrm{B}}$ is an uncertainty parameter that describes random and systematic errors in the experimental data as well as errors in the forward model. $p(\tilde{f}_i|X, \sigma_i^{\mathrm{SEM}})$ is the likelihood of observing $\tilde{f}_i$ given an estimate $\langle f_i \rangle_{N_r}$ over a finite number of replicas, where $\langle f_i(X) \rangle_{N_r} = \frac{1}{N_r} \sum_{r=1}^{N_r} f_i(X_r)$ and σ_i^{SEM} is an uncertainty parameter. $p(X_r)$, $p(\sigma_{r,i}^{\mathrm{B}})$, and $p(\sigma_i^{\mathrm{SEM}})$ are the priors on the conformations X_r (i.e., the force field), on the uncertainty σ^{B} in the experimental and back-calculated data, and on the uncertainty σ^{SEM} in the estimate of the true ensemble average, respectively.

A simple form can be obtained by choosing a Gaussian form for the likelihood $p(d_i|\tilde{f}_i, \sigma_{r,i}^{\mathrm{B}})$ given that $p(\tilde{f}_i|X, \sigma_i^{\mathrm{SEM}})$ is a Gaussian for the central limit theorem. In this case it is possible to write the metainference energy as

$$E_{\mathrm{MI}}(X, \boldsymbol{\sigma}|\boldsymbol{d}) = \sum_{r=1}^{N_r} \left\{ E_{\mathrm{ff}}(X_r) + k_{\mathrm{B}}T \sum_{i=1}^{N_d} \left[\frac{(\langle f_i(X) \rangle - d_i)^2}{2\sigma_{r,i}^2} \right. \right.$$
$$\left. \left. + \frac{1}{2} \ln 2\pi\sigma_{r,i}^2 + \frac{1}{2} \ln \frac{\sigma_{r,i}^2}{2} \right] \right\},$$
(2)

where $E_{\mathrm{ff}}(X_r)$ is the energy of the force field for the conformation X_r, $\sigma_{r,i}$ is the total uncertainty defined as $\sigma_{r,i}^2 = (\sigma_{r,i}^{\mathrm{B}})^2 + (\sigma_i^{\mathrm{SEM}})^2$, and the two logarithmic terms are the normalisation and the Jeffreys prior for $\sigma_{r,i}$, respectively.

1. Estimate of $\sigma^{\mathbf{SEM}}$

The metainference energy in the Gaussian form immediately shows the similarities between metainference and the replica-averaged simulations based on the maximum entropy[30] principle. Indeed metainference reduces to the maximum entropy replica-averaged modelling;[6,30,31] in the case that the only source of error is the ensemble averaging, σ_i^{SEM} is equivalent to the force constant employed there. In metainference as well as in replica-averaged restrained simulations, $1/(\sigma_i^{\mathrm{SEM}})^2$ is the only parameter to be set and it was shown that this must be chosen to be the largest possible force constant that can be integrated correctly by the system. Furthermore it was observed that it should scale more than linearly with the number of replicas.[31] In metainference σ^{SEM} is related to the standard error of the mean and as such, in absence of other sources of errors, the force constant actually scales $\propto N_r^2$.

In principle one should set σ_i^{SEM} for each experimental data used as restraint. A practical solution to this problem that was often employed is that of selecting one value common for all the data in a dataset. Here we introduce an alternative solution. We estimate it as $\sigma_i^{\mathrm{SEM}} = \sqrt{\max(\mathrm{Var}[f_i](t))/N_r}$, that is, the square root of the maximum over the simulation time of the variance of the forward model for the observable i divided

by the number of replicas. This guarantees the correct scaling of σ_i^{SEM} with the number of replicas, a weak time dependence, and a value proportional to the variance of the back-calculated observable and thus its dynamics. In the supplementary material, Figures S1 and S2 show how the algorithm allows us to quickly reach a stable estimate for σ^{SEM} in the first few nanoseconds of simulations.

2. Estimate of the weighted σ^{SEM}

In M&M the arithmetic average is substituted by a weighted average to consider for the effect of the bias, $\langle f_i(X)\rangle_{N_r} = \sum_{r=1}^{N_r} \frac{w_r}{N_w} f_i(X_r)$, where $N_w = \sum_{r=1}^{N_r} w_r$. In this case the forces resulting from metainference are not equally distributed among the replicas but are distributed proportionally to the weight of each single replica at each time step. In this case the standard error of the mean should take the variance of the weights into account. This is done by implementing it as[32]

$$(\sigma_i^{\text{SEM}})^2 = \frac{N_r}{(N_r - 1)N_w^2} \left[\sum_{r=1}^{N_r} (w_r f_i(X_r) - \langle w\rangle\langle f_i(X)\rangle)^2 \right.$$
$$- 2\langle f_i(X)\rangle \sum_{r=1}^{N_r} (w_r - \langle w\rangle)(w_r f_i(X_r) - \langle w\rangle\langle f_i(X)\rangle)$$
$$\left. + \langle f_i(X)\rangle^2 \sum_{r=1}^{N_r} (w_r - \langle w\rangle)^2 \right] \qquad (3)$$

and as for the unweighted case using the square root of the maximum value sampled along the simulation.

3. Generalisation for observables defined but for a scaling factor

Experimental observables can be defined modulo a multiplicative constant; this is the case with RDCs as their intensity is proportional to the fraction of aligned molecules. In these cases it is not possible to directly compare back-calculated and experimental data. One possible solution is that of considering an energy term proportional to the correlation between the experimental and the back-calculated data.[33] Alternatively one can extend the metainference formalism to take additional parameters into account, for example, a scaling factor. In this

case the metainference energy becomes

$$E_{\text{MI}}(X, \sigma, \lambda|d) = \sum_{r=1}^{N_r} \left\{ E_{\text{ff}}(X_r) + k_B T \sum_{i=1}^{N_d} \left[\frac{(\lambda\langle f_i(X)\rangle - d_i)^2}{2\sigma_{r,i}^2} \right. \right.$$
$$\left. \left. + \frac{1}{2}\ln 2\pi\sigma_{r,i}^2 + \frac{1}{2}\ln\frac{\sigma_{r,i}^2}{2} \right] \right\}, \qquad (4)$$

where λ is the scaling factor and $\sigma_{r,i}^2 = (\sigma_{r,i}^{\text{B}})^2 + \lambda^2(\sigma_i^{\text{SEM}})^2$ and the two logarithmic terms are the normalisation and the Jeffrey prior, respectively.

4. Restraint correction for high forces

The restraint intensity dependent on σ^{B} and σ^{SEM} can occasionally lead to unrealistic forces, causing instability in particular in the transient time at the beginning of the simulation when σ^{SEM} is still under estimated. To decrease the probability of this occurring, a correction factor is introduced in such a way as to temporarily decrease the applied forces. This is defined as

$$s_t = \begin{cases} s_{t-1} - \dfrac{\Delta s}{100}\ln(\dfrac{s_{t-1}}{s_{\min}}), & \text{if } n_{F_{\text{MD}}>F_{\max}} = 0, \\[2mm] s_{t-1} + \Delta s\ln(n_{F_{\text{MD}}>F_{\max}} + 1), & \text{if } n_{F_{\text{MD}}>F_{\max}} > 0, \\[2mm] s_{\max}, & \text{if } n_{F_{\text{MD}}>F_{\max}} > 0 \\ & \text{and } s_t > s_{\max}, \end{cases} \qquad (5)$$

where s_t is the correction factor at time step t to be multiplied with σ^{SEM}, Δs is the step size for the correction factor, $s_{\min}$ and $s_{\max}$ are the respective minimum and maximum possible correction values, and $n_{F_{\text{MD}}>F_{\max}}$ is the number of molecular dynamics forces above a certain threshold force $F_{\max}$. This update rule has the effect of immediately relaxing the restraint in the case of excessively high forces, followed by a slow annealing to the $s_{\min}$ value in the case of no high force events. By specifying a $s_{\min}$ value different from 1, one is able to bias the restraint intensity towards lower or higher values.

B. Simulation details

All simulations (Table I) are carried out with Gromacs 5.1.4[34] and a development version of PLUMED 2.3.[29] The peptide with the sequence EGAAWAASS is created in VMD[35] and is solvated in a rhombic dodecahedron box with side

TABLE I. All simulations performed. Force-field, water model, performances (ns/(day/replica)), and experimental data used as restraints are reported. Performances were estimated on an Intel E5-2660 2.4 GHz using one thread per replica.

Simulation	Force field	Water model	Performances	Convergence	Free energies	Back-calculations
Unrestrained	CHARMM22*	TIP3P	14.8	Fig. S5	Fig. S13	Fig. S21
	AMBER99SB	TIP4P-D	10.5	Fig. S6	Fig. S14	Fig. S22
CS, JC	CHARMM22*	TIP3P	14.6	Fig. S7	Fig. S15	Fig. S23
	AMBER99SB	TIP4P-D	10.3	Fig. S8	Fig. S16	Fig. S24
CS, JC, RDCs	CHARMM22*	TIP3P	14.4	Fig. S9	Fig. S17	Fig. S25
	AMBER99SB	TIP4P-D	10.2	Fig. S10	Fig. S18	Fig. S26
Unrestrained	CHARMM36	EEF1-SB	389.4	Fig. S11	Fig. S19	Fig. S27
CS, JC, RDCs	CHARMM36	EEF1-SB	244.9	Fig. S12	Fig. S20	Fig. S28

lengths of 4.5, 4.5, and 3.2 nm using 2118 water molecules. The system is neutralized by the addition of 3 Na⁺ and 2 Cl⁻ ions. Minimization of the system is performed with the steepest descent algorithm to a maximum force of less than 100 (kJ/(mol/nm)). Equilibration is performed over a time range of 500 ps in the NVT ensemble using the Bussi thermostat[36] and for 500 ps in the NPT ensemble using Parrinello-Rahman[37] pressure coupling while applying a position restraint on all heavy atoms. Production simulations are carried out using AMBER99SB[38] with TIP4P-D[19] water model and CHARMM22*[18,39] with TIP3P[40] water model with a time step of 2 fs at a temperature of $T = 300$ K in the NPT ensemble. Van der Waals and electrostatic interactions are modelled using the Particle-Mesh-Ewald[41,42] approach and a cutoff for the short-range interactions of 0.9 nm. Constraints are applied on all bonds with the LINCS algorithm[43] using a matrix expansion on the order of 6 and 2 iterations per step.

Metadynamics[15] is performed with the well-tempered,[44] parallel-bias,[28] and multiple-walkers[45] protocols, using a Gaussian deposition stride of 500 steps (1 ps), a bias factor of 8, and a Gaussian height of 0.3 kJ/mol for 14 replicas. The following collective variables are biased and corresponding sigma values are given in parentheses: All backbone ψ and ϕ dihedral angles ($\sigma = 0.6$) as well as the E1-S9 C^α–C^α distance ($\sigma = 0.3$ nm^{-1}), W5 χ^1, W5 χ^2 ($\sigma = 0.6$), similarities between ϕ_3 and ϕ_6 as well as ψ_3 and ψ_6 dihedral angles ($\sigma = 0.3$). Each replica is run for 100 ns for a total of 1.4 μs nominal simulation time per ensemble.

1. M&M simulation including ³J couplings and chemical shifts

Metainference calculations are performed using ensemble averages weighted according to the metadynamics bias potential. Experimental data are provided by Dames *et al.*[17] Chemical shifts were calculated using CamShift[46,47] for NH, HN, Hα, Cα, Cβ, and C′ backbone atoms while excluding the first and last residues. Hα–N, Hα–HN, W5 C–Cγ, and W5 N–Cγ ³J-coupling constants were calculated using the Karplus equation,[48]

$$^3J(\theta) = A \cos^2(\theta + \Delta\theta) + B \cos(\theta + \Delta\theta) + C, \quad (6)$$

where $^3J(\theta)$ is the coupling in Hz, A, B, and C are the Karplus parameters dependent on the type of coupling, θ is a dihedral angle, and $\Delta\theta$ is a constant shift added on to the angle. The Karplus parameters and shift $\Delta\theta$ are taken from Refs. 49 and 50. The noise is sampled independently for each datapoint through Brownian motion (flat prior) with a step size of 0.5 and hard limits at 0.001 and 25, respectively. The restraint correction for high forces was applied with s_{min} = 1.0, s_{max} = 2.0, $\Delta s = 0.001$, and F_{max} = 3500 (kJ/(mol/nm)).

2. M&M simulation including RDCs

Simulations are performed as described above with the addition of residual dipolar couplings for N–H and Cα–Hα bonds.[17] RDCs are calculated using the θ-method;[33] each coupling is calculated independently using the dipolar coupling definition

$$D_i = -\frac{\mu_0 \gamma_1 \gamma_2 \hbar}{8\pi^3} \left(\frac{3\cos^2\vartheta_i - 1}{r_i^3} \right), \quad (7)$$

where r_i is the bond length, μ_0 is the magnetic constant, γ_1 and γ_2 are the gyromagnetic ratios for the two atoms, $\hbar$ is the Planck constant, and ϑ_i is the angle between the bond and the z-axis. The coupling is then averaged and compared modulo a scaling factor λ with the experimental data. This allows us to simultaneously account for the conformational and rotational averaging measured by RDCs.[11,33,51,52] The scaling factor is sampled during the simulation using an Ornstein-Uhlenbeck process (Gaussian prior),

$$d\lambda_t = \frac{1}{2}(\mu - \lambda_t) + \Delta\lambda \frac{e}{\pi} dW_t, \quad (8)$$

where $d\lambda_t$ is the step taken, μ is the specified mean of the stationary Gaussian distribution, λ_t is the scaling value at time t, $\Delta\lambda$ is the standard deviation of the stationary Bayesian distribution, and dW_t denotes the Wiener process. The values chosen for N–H and Cα–Hα RDCs are $\mu = 8$ and $\mu = 9$, respectively, and $\Delta\lambda = 0.5$. In Figures S3 and S4 of the supplementary material the sampling of the scaling factor for N–H and Cα–Hα RDCs is shown. The sampling converges quickly after a few steps, of notice is that the average value for the scaling factor found is different depending on the force field used as a prior; this is due to the differences in the bond lengths of those bonds in the two force fields. In CHARMM22* an N–H bond is 0.0997 nm long and a Cα–Hα is 0.1080 nm long, while in AMBER99SB the same bonds are 0.1010 and 0.1090 nm long.

3. M&M using the EEF1-SB implicit solvent model

Simulations in an implicit solvent are performed using the EEF1 model originally developed by Lazaridis and Karplus[23] and subsequently optimised by Bottaro *et al.* (EEF1-SB)[24] in combination with CHARMM36.[22] EEF1-SB is a solvent-accessible surface area based model, where the free energy of solvation is computed using a pairwise interaction term for non-hydrogen atoms,

$$\Delta G_i^{solv} = \Delta G_i^{ref} - \sum_{j \neq i} f_i(r_{ij}) V_j, \quad (9)$$

where ΔG_i^{solv} is the free energy of solvation, ΔG_i^{ref} is the reference solvation free energy, V_j is the volume of atom j, and

$$f_i(r)4\pi r^2 = \frac{2}{\sqrt{\pi}} \frac{\Delta G_i^{free}}{\lambda_i} \exp\left\{ -\frac{(r - R_i)^2}{\lambda_i^2} \right\}, \quad (10)$$

where ΔG_i^{free} is the solvation free energy of the isolated group, λ_i is the correlation length equal to the width of the first solvation shell, and R_i is the van der Waals radius of atom i. The implicit solvation model is implemented in PLUMED. In our implementation, interactions are cut off after a range of $3\lambda_i$. In addition, electrostatic interactions are further screened with a position dependent dielectric constant of the form $\epsilon = 1/(\alpha r)$. Bottaro *et al.* optimised α to 15 nm^{-1} and added an energy correction for backbone dihedrals on N–C′–Cα–Cβ. All these parameters and corrections are designed to be used with the CHARMM36 force field.[22] Charged amino acids are neutralised by adjusting the partial charges, leaving a completely neutral molecule. Minimization is performed as for the explicit solvent simulations. The system is evolved by a Langevin dynamic integrator with a friction coefficient of 1 ps^{-1} at $T = 300$ K. Coulomb interactions are tabulated with a distance

dependent dielectric constant of $\epsilon = 15r$ and a cutoff at 0.9 nm, while van der Waals interactions are switched off smoothly between 0.7 and 0.9 nm. All pairwise interactions are computed using a neighbourlist with a buffer of 0.2 nm with respect to the cutoff, which is updated every 10 simulation steps. Constraints are applied on all bonds with the LINCS algorithm, as described above. Metadynamics and M&M simulations are performed as already mentioned for explicit solvent simulations using the same collective variables, parameters, and experimental data.

4. Analysis

In well-tempered metadynamics, the time-dependent bias converges to a quasi-static distribution; as a consequence, a signature of convergence can be obtained by a block comparison of the sampling after a transient time. If the simulations are converged, the histograms obtained for non-overlapping blocks of simulations should result in comparable effective free energies.

In the present case, convergence is assessed by comparing the free energies calculated from the histograms of each biased collective variable for the last two 45 ns segments of the simulation (i.e., from 10 to 55 ns and from 55 to 100 ns). The free energies represent the effective potential felt by the system as a sum of the force field, the metainference potential if present, and metadynamics. In Figures S5–S12 of the supplementary material, the comparisons of the effective free energies obtained along the 21 collective variables employed are shown, with differences that are limited to few high-energy regions and an average root-mean-square deviation (RMSD) of 0.50 kJ/mol. The converged free energies for all the collective variables are shown in Figures S13–S18 of the supplementary material.

Equilibrium distributions are then recovered by reweighting the ensembles according to the final deposited metadynamics bias.[53] The weight of each sampled conformation is given by $w_i = \exp(+V^{\text{MetaD}}(\text{CV}(X))/k_B T)/Z$, where $V^{\text{MetaD}}(\text{CV}(X))$ is the metadynamics bias calculated for conformation X at the end of the simulation and Z is the normalisation. Chemical shifts are back-calculated using CamShift;[46] residual dipolar couplings are computed using the single-value-decomposition method. ^{3}J-couplings are back-calculated using the Karplus equation. All experimental observables are calculated as weighted ensemble averages.

In addition to the conformational ensemble, the result of a metainference calculation also includes an estimate of the errors, $\sigma_{r,i}^B$, for all the experimental data added.[13] These errors incorporate in a single number an independent estimate of the experimental random and systematic errors as well as the errors in the forward-model. Indeed, while σ_i^{SEM} is an error that accounts for the use of a limited number of replicas on the fly, $\sigma_{r,i}^B$ is a useful additional source of information that results from the use of metainference.

Finally, in order to further compare the ensembles not only in terms of their agreement with experimental data but also with respect to finer properties, similarities between probability distributions are computed using the Jensen-Shannon divergence. Given two probability distributions P and Q obtained by two ensembles, their difference is

$$D_{\text{JS}}(P||Q) = \frac{1}{2}D_{\text{KL}}(P||M) + \frac{1}{2}D_{\text{KL}}(Q||M), \quad (11)$$

where $M = \frac{1}{2}(P + Q)$ and D_{KL} is the Kullback-Leibler divergence: $D_{\text{KL}} = \sum_i P(i)\ln(P(i)/Q(i))$.

III. RESULTS

In the following we present the results of eight ensembles for the EGAAWAASS[17] peptide (cf. Table I) obtained by running for each case 14 replicas for 100 ns per replica, either using only metadynamics,[15] i.e., without the addition of any experimental restraint, or by coupling the replicas using metadynamic metainference[14] and multiple experimental data. We have tested two state-of-the-art force fields in the explicit solvent, CHARMM22* in TIP3P[18] and AMBER99SB in TIP4P-D,[19] and the CHARMM36 EEF1-SB[24] implicit solvent force field recently optimised to study disordered systems. The addition of experimental data modifies the ensembles towards a result where both local and non-local properties are comparable, irrespective of the original force field employed. The implicit solvent scheme supplemented by the experimental data allows us to obtain results comparable to those obtained in the explicit solvent at a fraction of the computational cost. A PLUMED[29] input file is provided in the supplementary material to reproduce all the simulations.

A. Comparison with the experimental data

First, we assessed the quality of the force fields and the ability of metainference to successfully improve them through the weighted incorporation of experimental information. Root-mean-square deviations (RMSDs) from the experimental data, shown in Figure 1, show a clear decrease, and hence an increase in the agreement with experimental data, with the addition of more information into the system. Both force fields, CHARMM22* in blue and AMBER99SB in red, show a comparably good agreement with chemical shifts, with comparable trends in the per-residue deviations of NH, Hα, Cα, and Cβ and more marked differences in the case of C$'$ and HN chemical shifts (cf. Figures S21 and S22 of the supplementary material). Both force fields show a very good agreement with ^{3}J$_{\text{H}_\alpha\text{-N}}$, CHARMM22* describes the χ_2 angle of W5 well (i.e., ^{3}J$_{\text{C-C}\gamma}$), while both agree less well with the χ_1 angle of the same residue as well as the ^{3}J$_{\text{H}_\alpha\text{-HN}}$ and the RDCs. For the case of chemical shifts, the per-residue comparison shows comparable trends for some observables (i.e., ^{3}J$_{\text{C-C}\gamma}$, ^{3}J$_{\text{N-C}\gamma}$, and N–H RDCs) and more marked differences for others (i.e., ^{3}J$_{\text{H}_\alpha\text{-N}}$, ^{3}J$_{\text{H}_\alpha\text{-HN}}$, and C$\alpha$–H$\alpha$ RDCs). These differences suggest that the two force fields are not giving an equivalent description of the peptide and that the addition of experimental information could actually improve them.

M&M ensembles including chemical shifts (that are not expected to contribute particularly given the already good agreement) and ^{3}J-couplings have indeed a positive effect on the RMSDs of all data, including in particular
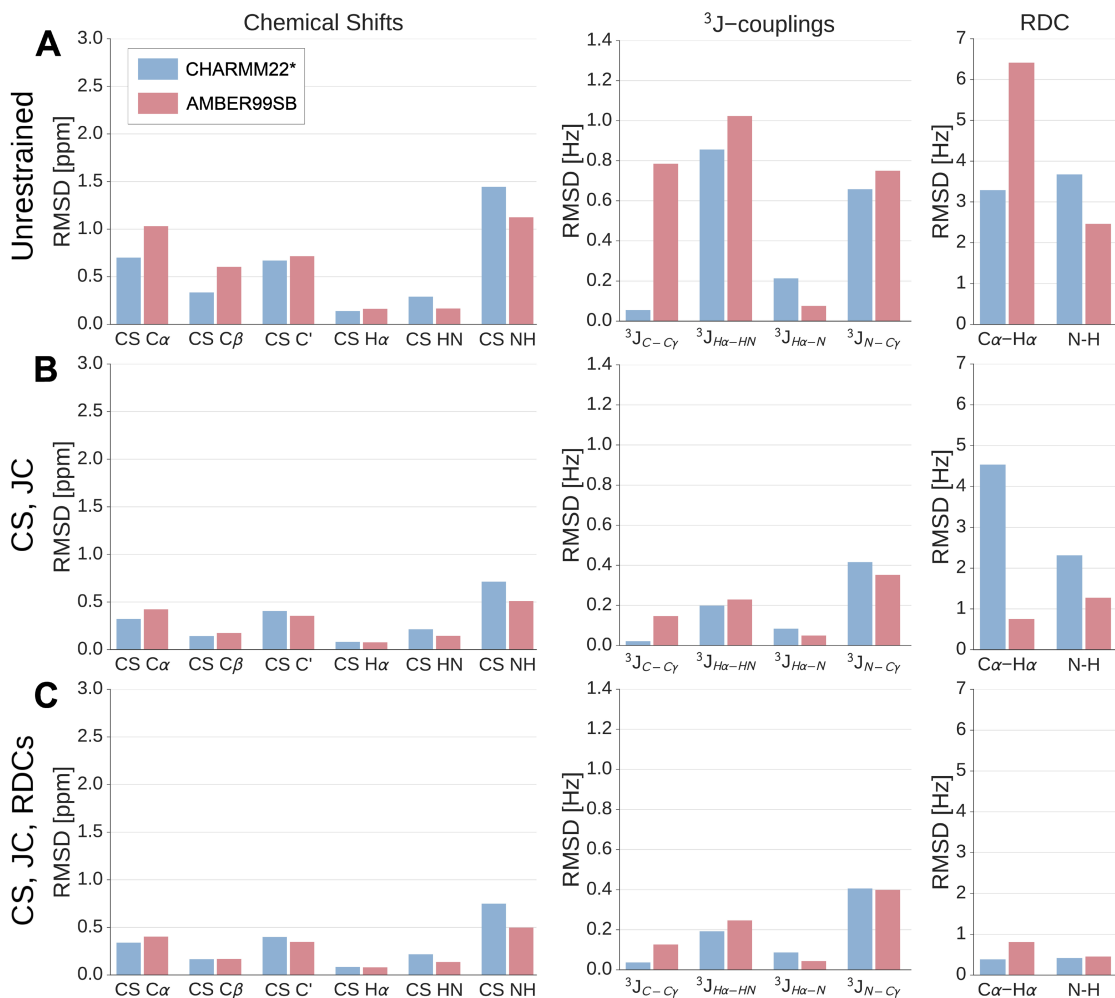
FIG. 1. Root-mean-square deviations between simulated and experimental data for all simulations performed in the explicit solvent. Transparent blue bars represent ensembles based on the CHARMM22* prior, while transparent red bars represent ensembles based on the AMBER99SB in TIP4P-D water prior. Each row indicates successive addition of data to the simulation, (a) unrestrained, (b) addition of chemical shifts and ^{3}J-couplings, (c) further addition of RDCs. Fully restrained simulations (c) show consistent improvement when compared to unrestrained simulations (a). The addition of RDC restraints has little additional impact on the quality of other experimental observables, while the addition of chemical shifts and ^{3}J-couplings has a positive impact on the quality of the RDCs. Per residue comparisons for all data and ensembles can be found in Figures S21–S26 of the supplementary material.

RDCs (cf. Fig. 1), where the improvement is more pronounced for AMBER99SB than for CHARMM22*. Interestingly, the per-residue trends are now also more comparable, with AMBER99SB showing an overall better agreement with all the available data (cf. Figures S23 and S24 of the supplementary material). Finally, the M&M ensembles also including the RDCs showed comparable, good agreements (Fig. 1) and comparable trends for all data (cf. Figures S25 and S26 of the supplementary material). This suggests that while the original force fields were providing two alternative and not completely satisfactory descriptions of the dynamics of the peptide under study, the M&M ensembles could instead provide ensembles that are indistinguishable from the point of view of the available experimental observables.

The effect of metainference on the experimental data can be also observed at a finer detail in Figure 2 where the distributions of the Cα–Hα RDCs and those for the Cα carbon chemical shifts are compared. In the case of chemical shifts, the distributions were already similar between the two unrestrained force fields, with the exception of A6. Upon restraining the chemical shifts and ^{3}J-couplings, the

distributions are translated closer to the reference experimental values but their overall shape is unchanged. The further addition of RDCs does not have any additional effect on the chemical shifts. For RDCs the unrestrained simulations show very broad distributions with average values far from the experimental data. Furthermore, the overall shape of the distribution can be very different for the two force fields as is the case for W5, A6, and A7. Upon restraining with chemical shifts and ^{3}J-couplings, the RDCs for W5, A6, and A7 showed an improvement in the agreement with the experimental data and an improved similarity between the two ensembles. The final addition of RDCs' restraints shrunk the distributions and translated them closer to the reference experimental values making the overall shape comparable between the two ensembles as visible from the quartiles. Once again the final distributions are not only similar in the average value but also in their quartiles, suggesting that the two final ensembles obtained by restraining CHARMM22* and AMBER99SB in TIP4P-D are not only in good agreement with the experimental values but also very similar to each other.
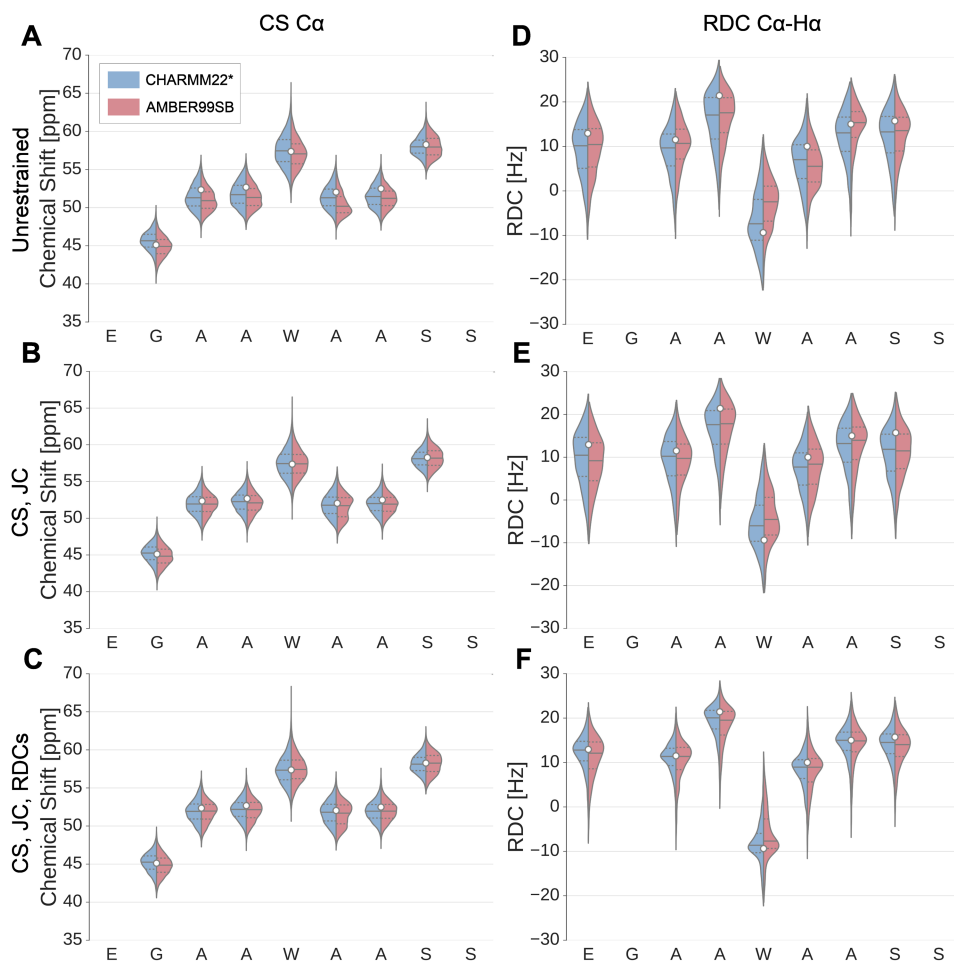
FIG. 2. Violin plots showing the probability distributions of the restrained experimental data per residue without the addition of data (top row, (a) and (d)) and with addition of chemical shifts, ^{3}J-couplings (middle row, (b) and (e)) and chemical shifts, ^{3}J-couplings and RDCs (bottom row, (c) and (f)). Transparent blue distributions represent ensembles based on the CHARMM22* prior, while transparent red distributions represent ensembles based on the AMBER99SB in TIP4P-D water prior. Means and quartiles are indicated by full and dashed lines, respectively, while the true experimental value is shown as a dot. Left panels ((a)–(c)) show the comparison of Cα carbons chemical shifts, right panels ((d)–(f)) show the comparison of Cα–Hα RDCs.

B. Convergence towards a common ensemble

While the comparison with the experimental data suggests that it is possible to use M&M to generate ensembles, starting from two alternative priors encoded in the two employed force fields—which are in remarkable good and similar agreement with the experimental data—it is still possible at least in principle that the ensembles could give different results if observed through other techniques. In order to test the hypothesis that M&M can provide at least in principle a unique ensemble, we analyzed the similarities of the ensembles with respect to other independent properties.

In Figure 3 the ensembles are compared using two alternative similarity metrics. In the left panel the probability distributions of the radius of gyration for all the pairs of ensembles are compared; their dissimilarity is measured by the Jensen-Shannon divergence (see Sec. II B 4). None of the employed experimental data is a direct measure of the radius of gyration, which makes this a good candidate for an observable that can reveal differences between the restrained ensembles. With the addition of information into the system, convergence towards a common distribution is remarkably visible (bottom right). While the unrestrained simulations show a remarkably different behaviour, with the CHARMM22* ensemble being more compact than the AMBER99SB one, both start developing a pronounced peak at about 0.85 nm with the

introduction of additional information. This is visible in the form of an increased overlap between the two distributions. The Jensen-Shannon divergence confirms the visual suggestion. Of notice is that the probability distribution for the AMBER99SB ensemble, once updated with chemical shifts and ^{3}J-couplings, seems to be already converged, in line with a good agreement of this ensemble with RDCs (cf. Fig. 1). *A posteriori* one can speculate that the experimental data employed, even if reporting about local quantities, includes indirect information about the extended state of the peptide. Since AMBER99SB with TIP4P-D prior already provides an extended ensemble, less data are needed to converge the distribution overall.

The distance matrices (Figure 3(b)) also support the notion of convergence of the ensembles towards a unique indistinguishable one. In this case, we compared the average distances among all residues calculated using the centre of mass of the residues. Again none of the experimental data reports on such high-resolution information. Here the deviation decreases in a fashion corresponding with the Jensen-Shannon divergence in the left panel. To further stress the differences between the original ensembles with respect to the similarities of the final ensemble, the distributions for the overall backbone dihedral similarities and the end-to-end distance are shown in the supplementary material (Fig. S29).
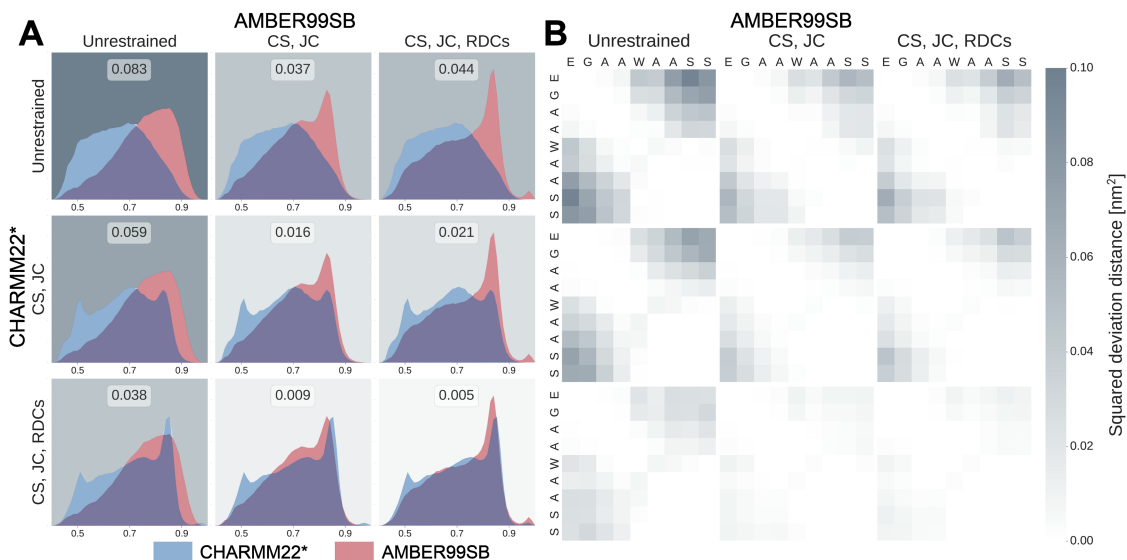
FIG. 3.   Convergence towards a unique ensemble. The two panels show two measures of ensembles-dissimilarity. (a) Ensemble convergence shown as Jensen-Shannon divergences between the probability distributions of the radius of gyration for all pairs of simulations in the explicit solvent given in nat (natural unit of information entropy). Transparent blue distributions represent ensembles based on the CHARMM22* prior, while transparent red distributions represent ensembles based on the AMBER99SB in TIP4P-D water prior. Lighter backgrounds indicate a lower divergence and thus higher similarity. The improvement is subjectively noticeable by comparing the underlying probability distributions. (b) Squared-deviation inter-residue distance matrices between each pair of simulations. Lower distance deviations correspond to a higher degree of similarity.

It is of notice that unrestrained CHARMM22* and fully restrained AMBER99SB are closer in similarity than the two unrestrained simulations and vice versa, as shown by both analyses reported in Figure 3. This again supports the notion of a funnelled picture towards a common unique ensemble that is independent from the prior knowledge and also suggests how state-of-the-art force fields seem to be converging to such a unique ensemble from different starting points.

C. Ensemble determination in implicit solvent

While the simulations discussed above show strong differences when not restrained, both use priors of similar quality with respect to experimental measures. An interesting question is therefore how a restrained simplified prior fares with respect to more conventional and more accurate priors. To this end, we performed M&M simulations using the computationally very efficient implicit solvent model EEF1-SB (cf. Table I). The combined results can be seen in Figure 4. The unrestrained ensemble is in relative good agreement with RDCs while showing a worse agreement with the other data than the explicit solvent unrestrained ensembles, suggesting that RDCs are better captured by the extended description of the peptide resulting from this prior (cf. Figure 4(b)). As expected the root-mean-square deviations showed a marked decrease with the addition of experimental data. The restrained ensemble is then, when compared with Figure 1, at least on par with the unrestrained explicit solvent simulations. The probability distribution of the radius of gyration (Figure 4(b)) shows the dramatic effect of the restraint. The prior is clearly biased towards very open states while the explicit solvent simulations show a more balanced picture (Figure 3(a)). This bias is not surprising given that EEF1-SB was explicitly optimised for disordered systems.[24] The restrained implicit solvent simulation is able to alleviate the over-extended description provided by the prior even if without reproducing quantitatively the pronounced peak at 0.8 nm present in CHARMM22*. Nonetheless, the Jensen-Shannon divergence of 0.015 between fully restrained CHARMM22* and fully restrained CHARMM36 with EEF1-SB supports their overall similarity, indeed this divergence is very similar to the one observed between the partially restrained explicit solvent simulations. Finally we compared the alternative similarity metric in the form of inter-residue distances (Figure 4(c)), which are practically indistinguishable.

Disordered systems are often characterised by transient secondary structures and elongated conformations. In order to further test the similarity of the ensembles in explicit and implicit solvents, we calculated the secondary structure populations over the ensembles using STRIDE[54] with errors estimated using the block standard error approach.[55] While all ensembles show poor helical and hairpin content (cf. Figs. S30 and S31 of the supplementary material), a convergence of the conformational space as a function of the addition of experimental information towards a common ensemble for both the explicit and implicit solvent simulations can be appreciated for the turn and coil content (cf. Figs. S32 and S33 of the supplementary material, respectively). The unrestrained AMBER99SB and the unrestrained CHARMM36 EEF1-SB ensembles show a lower turn content than the CHARMM22* unrestrained ensemble and an opposite behavior for the coil content. The successive addition of experimental data brings the AMBER99SB, CHARMM22*, and CHARMM36 EEF1-SB fully restrained ensembles to show essentially the same turn content and the same coil content.

These results further enforce the notion that M&M allows a radical reshaping of the prior ensemble to a common solution that is consistent across vastly different priors.
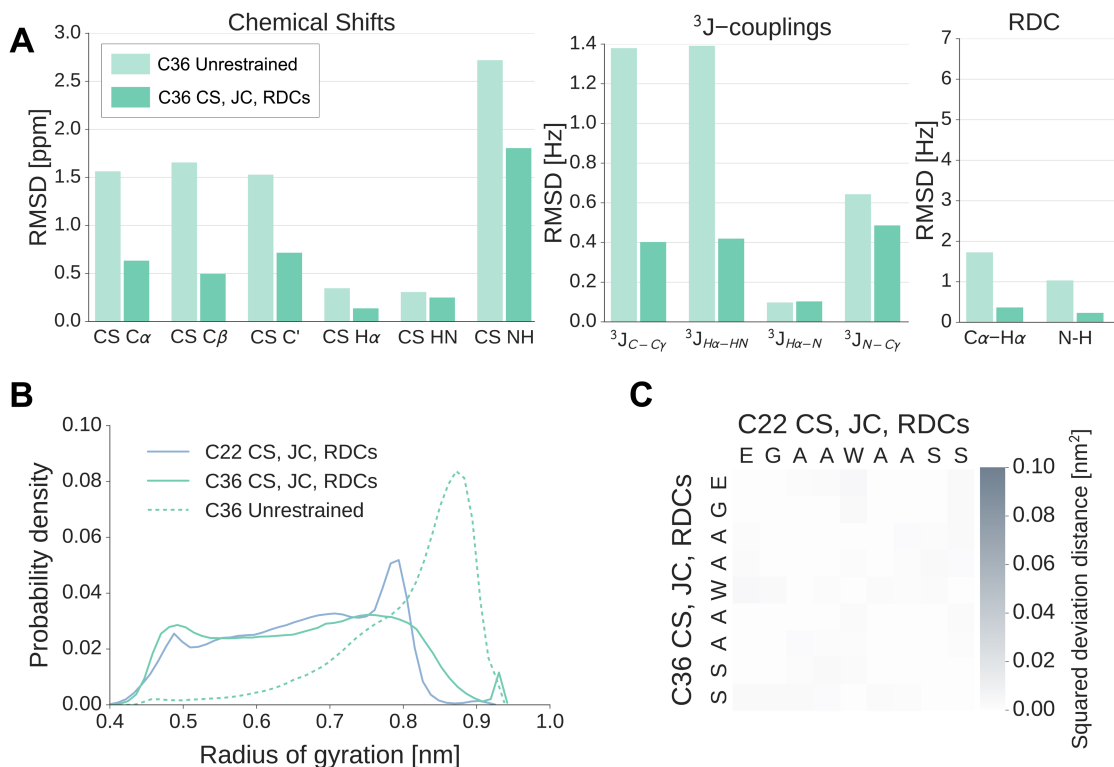
FIG. 4. Results of M&M simulations using the EEF1-SB implicit solvent model. (a) Root-mean-square deviations of the unrestrained and fully restrained implicit solvent simulations. (b) Probability distributions of the radius of gyration for unrestrained and restrained implicit solvents as well as fully restrained CHARMM22* simulations. The Jensen-Shannon divergences of the probability distributions of the radius of gyration between CHARMM36 and CHARMM22* with full restraints and CHARMM36 with restraint and unrestrained CHARMM36 are 0.015 nat and 0.203 nat (natural unit of information entropy), respectively. (c) Squared-deviation inter-residue distance matrix between fully restrained CHARMM22* and fully restrained CHARMM36 EEF1-SB simulations.

IV. CONCLUSIONS

The last few years have seen a large increase in the assessment of force fields. While there is a clear trend in the improvement of force field quality, their transferability between disordered and ordered systems and the robustness of the resulting structural ensembles for disordered systems are often questioned.[21,56–59] To circumvent force field limitations as well as limitations in the resolution of experimental techniques, hybrid methods based on the integration of experimental data and molecular dynamics simulations have seen a huge growth.[1,11] Here, we first simplified the setup of metadynamic metainference simulations to make them essentially parameter free and extended the formalism to account for experimental data that are defined as modulo a constant. Then we studied a disordered peptide to understand two concepts: first, to which extent, given enough experimental data, it is possible to obtain ensembles of structures that do not depend explicitly on the molecular mechanics force field employed; and second if it possible to obtain results of comparable quality at a fraction of the computational cost. By comparing two state-of-the-art explicit-solvent force fields and integrating them with multiple sources of experimental data, we determined two ensembles that are essentially indistinguishable from each other and different from those obtained using the force fields alone. Furthermore, results of comparable quality have been obtained using M&M, the same data, and a very inexpensive implicit solvent force field.

SUPPLEMENTARY MATERIAL

See supplementary material for convergence tests, the free energy profiles as a function of the collective variables employed in metadynamics, and more comparisons with experimental data.

ACKNOWLEDGMENTS

[1]M. Bonomi, G. T. Heller, C. Camilloni, and M. Vendruscolo, Curr. Opin. Struct. Biol. **42**, 106 (2017).

[2]G. Wei, W. Xi, R. Nussinov, and B. Ma, Chem. Rev. **116**, 6516 (2016).

[3]V. Venditti, T. K. Egner, and G. M. Clore, Chem. Rev. **116**, 6305 (2016).

[4]L. D. Cabrita, A. M. E. Cassaignau, H. M. M. Launay, C. A. Waudby, T. Wlodarski, C. Camilloni, M.-E. Karyadi, A. L. Robertson, X. Wang, A. S. Wentink, L. S. Goodsell, C. A. Woolhead, M. Vendruscolo, C. M. Dobson, and J. Christodoulou, Nat. Struct. Mol. Biol. **23**, 278 (2016).

[5] S. Milles, D. Mercadante, I. V. Aramburu, M. R. Jensen, N. Banterle, C. Koehler, S. Tyagi, J. Clarke, S. L. Shammas, M. Blackledge, F. Gräter, and E. A. Lemke, Cell **163**, 734 (2015).

[6] K. Lindorff-Larsen, R. B. Best, M. A. Depristo, C. M. Dobson, and M. Vendruscolo, Nature **433**, 128 (2005).

[7] A. Ramanathan, A. Savol, V. Burger, C. S. Chennubhotla, and P. K. Agarwal, Acc. Chem. Res. **47**, 149 (2014).

[8] H. J. Dyson and P. E. Wright, Nat. Rev. Mol. Cell Biol. **6**, 197 (2005).

[9] J. Habchi, P. Tompa, S. Longhi, and V. N. Uversky, Chem. Rev. **114**, 6561 (2014).

[10] E. Papaleo, G. Saladino, M. Lambrughi, K. Lindorff-Larsen, F. L. Gervasio, and R. Nussinov, Chem. Rev. **116**, 6391 (2016).

[11] E. Ravera, L. Sgheri, G. Parigi, and C. Luchinat, Phys. Chem. Chem. Phys. **18**, 5686 (2016).

[12] M. Schor, A. S. J. S. Mey, and C. E. MacPhee, Biophys. Rev. **8**, 429 (2016).

[13] M. Bonomi, C. Camilloni, A. Cavalli, and M. Vendruscolo, Sci. Adv. **2**, e1501177 (2016).

[14] M. Bonomi, C. Camilloni, and M. Vendruscolo, Sci. Rep. **6**, 31232 (2016).

[15] A. Laio and M. Parrinello, Proc. Natl. Acad. Sci. U. S. A. **99**, 12562 (2002).

[16] O. Valsson, P. Tiwary, and M. Parrinello, Annu. Rev. Phys. Chem. **67**, 159 (2016).

[17] S. A. Dames, R. Aregger, N. Vajpai, P. Bernado, M. Blackledge, and S. Grzesiek, J. Am. Chem. Soc. **128**, 13508 (2006).

[18] S. Piana, K. Lindorff-Larsen, and D. E. Shaw, Biophys. J. **100**, L47 (2011).

[19] S. Piana, A. G. Donchev, P. Robustelli, and D. E. Shaw, J. Phys. Chem. B **119**, 5113 (2015).

[20] H. T. A. Leung, O. Bignucolo, R. Aregger, S. A. Dames, A. Mazur, S. Bernèche, and S. Grzesiek, J. Chem. Theory Comput. **12**, 383 (2016).

[21] F. Palazzesi, M. K. Prakash, M. Bonomi, and A. Barducci, J. Chem. Theory Comput. **11**, 2 (2015).

[22] R. B. Best, X. Zhu, J. Shim, P. E. M. Lopes, J. Mittal, M. Feig, and A. D. MacKerell, J. Chem. Theory Comput. **8**, 3257 (2012).

[23] T. Lazaridis and M. Karplus, Proteins: Struct., Funct., Genet. **35**, 133 (1999).

[24] S. Bottaro, K. Lindorff-Larsen, and R. B. Best, J. Chem. Theory Comput. **9**, 5641 (2013).

[25] W. Rieping, M. Habeck, and M. Nilges, Science **309**, 303 (2005).

[26] C. Camilloni, A. Cavalli, and M. Vendruscolo, J. Chem. Theory Comput. **9**, 5610 (2013).

[27] C. Camilloni and M. Vendruscolo, J. Am. Chem. Soc. **136**, 8982 (2014).

[28] J. Pfaendtner and M. Bonomi, J. Chem. Theory Comput. **11**, 5062 (2015).

[29] G. A. Tribello, M. Bonomi, D. Branduardi, C. Camilloni, and G. Bussi, Comput. Phys. Commun. **185**, 604 (2014).

[30] A. Cavalli, C. Camilloni, and M. Vendruscolo, J. Chem. Phys. **138**, 094112 (2013).

[31] B. Roux and J. Weare, J. Chem. Phys. **138**, 084107 (2013).

[32] D. F. Gatz and L. Smith, Atmos. Environ. **29**, 1195 (1995).

[33] C. Camilloni and M. Vendruscolo, J. Phys. Chem. B **119**, 653 (2015).

[34] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl, SoftwareX **1-2**, 19 (2015).

[35] W. Humphrey, A. Dalke, and K. Schulten, J. Mol. Graphics **14**, 33 (1996).

[36] G. Bussi, D. Donadio, and M. Parrinello, J. Chem. Phys. **126**, 014101 (2007).

[37] M. Parrinello and A. Rahman, J. Appl. Phys. **52**, 7182 (1981).

[38] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling, Proteins **65**, 712 (2006).

[39] 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. 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, J. Phys. Chem. B **102**, 3586 (1998).

[40] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, J. Chem. Phys. **79**, 926 (1983).

[41] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen, J. Chem. Phys. **103**, 8577 (1995).

[42] C. L. Wennberg, T. Murtola, B. Hess, and E. Lindahl, J. Chem. Theory Comput. **9**, 3527 (2013).

[43] B. Hess, J. Chem. Theory Comput. **4**, 116 (2008).

[44] A. Barducci, G. Bussi, and M. Parrinello, Phys. Rev. Lett. **100**, 20603 (2008).

[45] P. Raiteri, A. Laio, F. L. Gervasio, C. Micheletti, and M. Parrinello, J. Phys. Chem. B **110**, 3533 (2006).

[46] K. Kohlhoff, P. Robustelli, A. Cavalli, X. Salvatella, and M. Vendruscolo, J. Am. Chem. Soc. **131**, 13894 (2009).

[47] C. Camilloni, P. Robustelli, A. De Simone, A. Cavalli, and M. Vendruscolo, J. Am. Chem. Soc. **134**, 3968 (2012).

[48] M. Karplus, J. Chem. Phys. **30**, 11 (1959).

[49] C. Pérez, F. Löhr, H. Rüterjans, and J. M. Schmidt, J. Am. Chem. Soc. **123**, 7081 (2001).

[50] B. Vögeli, J. Ying, A. Grishaev, and A. Bax, J. Am. Chem. Soc. **129**, 9377 (2007).

[51] A. N. Borkar, M. F. Bardaro, Jr., C. Camilloni, F. A. Aprile, G. Varani, and M. Vendruscolo, Proc. Natl. Acad. Sci. U. S. A. **113**, 7171 (2016).

[52] A. N. Borkar, P. Vallurupalli, C. Camilloni, L. E. Kay, and M. Vendruscolo, Phys. Chem. Chem. Phys. **19**, 2797 (2017).

[53] D. Branduardi, G. Bussi, and M. Parrinello, J. Chem. Theory Comput. **8**, 2247 (2012).

[54] D. Frishman and P. Argos, Proteins **23**, 566 (1995).

[55] H. Flyvbjerg and H. G. Petersen, J. Chem. Phys. **91**, 461 (1989).

[56] S. Rauscher, V. Gapsys, M. J. Gajda, M. Zweckstetter, B. L. de Groot, and H. Grubmüller, J. Chem. Theory Comput. **11**, 5513 (2015).

[57] F. Martín-García, E. Papaleo, P. Gomez-Puertas, W. Boomsma, and K. Lindorff-Larsen, PLoS One **10**, e0121114 (2015).

[58] F. Vitalini, A. S. J. S. Mey, F. Noé, and B. G. Keller, J. Chem. Phys. **142**, 084101 (2015).

[59] S. Piana, J. L. Klepeis, and D. E. Shaw, Curr. Opin. Struct. Biol. **24**, 98 (2014).