HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

On the Combined Power of Simulation and Experiment to Model Protein Structure, Dynamics, Function and Assembly Mechanisms

Inaugural-Dissertation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Oliver Schillinger
aus Duisburg

Düsseldorf, August 2017

aus dem Institut für Theoretische Chemie und Computerchemie
der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Referent: Juniorprof. Dr. Birgit Strodel

Korreferent: Prof. Dr. Christel Marian

Tag der mündlichen Prüfung: 13. Juli 2017

Doctoral Dissertation

On the Combined Power of Simulation and Experiment to Model Protein Structure, Dynamics, Function and Assembly Mechanisms

Oliver Schillinger
Forschungszentrum Jülich
Institute of Complex Systems 6 - Structural Biochemistry
Computational Biochemistry Group
Supervisor: Jun.-Prof. Dr. Birgit Strodel

August 29, 2017

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

JÜLICH
FORSCHUNGSZENTRUM

Education is not preparation for life.
Education is life itself.

— *John Dewey* —

Meiner Familie gewidmet

Eidesstattliche Erklärung

Ich versichere an Eides statt, dass die Dissertation von mir selbst verfasst und ohne unzulässige fremde Hilfe unter Beachtung der Grundsätze zur Sicherung der guten wissenschaftlichen Praxis an der Heinrich-Heine-Universität erstellt worden ist. Die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht. Die Dissertation wurde in der vorgelegten oder in ähnlicher Form noch bei keiner anderen Prüfungsbehörde eingereicht. Es wurden keine früheren erfolglosen Promotionsversuche unternommen.

__________________________________Mülheim an der Ruhr, August 29, 2017
Oliver Schillinger

Danksagung

Mein größter Dank gebührt meiner Frau Nicole, die mir vor allem während der vergangenen Monate den Rücken freigehalten und mich in den letzten Jahren enorm unterstützt hat. Vor allem hast du meine Launen und Krisen ausgehalten und mich immer wieder aufgebaut. Ich hoffe, ich kann mich in Zukunft revanchieren. Vielen Dank Nicole!

Entschuldigen möchte ich mich bei meinen Kindern Charlotte (3) und Franz (1), die in der letzten Zeit zu wenig von ihrem Vater hatten. In Zukunft haben wir wieder mehr Zeit miteinander. Vielen Dank Charlotte, dass du mir beim Schreiben der Doktorarbeit geholfen hast.

Ich danke meinen Eltern dafür, dass ich jederzeit meinen eigenen Weg gehen durfte und meine eigenen Entscheidungen treffen konnte. Vielen Dank auch für eure Liebe, Unterstützung und Förderung.

Besonderer Dank gilt meiner Doktormutter, Prof. Dr. Birgit Strodel, die mir während meiner Promotion alle Freiheiten gegeben hat, die ich mir wünschen konnte, mich aber gleichzeitig immer unterstützt hat und mit Rat zur Seite stand. In deiner Arbeitsgruppe habe ich mich in den letzten Jahren sehr wohl gefühlt.

Große Unterstützung habe ich auch von Prof. Dr. Christel Marian erfahren, die mir in den letzten Jahren großzügig einen Arbeitsplatz in angenehmer Atmosphäre im Institut für Theoretische Chemie und Computerchemie der Heinrich-Heine-Universität Düsseldorf zur Verfügung gestellt hat. In diesem Zusammenhang möchte ich mich auch bei meinen Mitstreiterinnen Irina Dokukina und Jelena Föller, sowie Gudrun Brauwers für die diskussionsfreudigen Kaffee- und Teepausen bedanken! Ich hoffe unsere Freundschaft reicht über das Ende unserer Promotionen hinaus.

Außerdem danke ich Andrew Dingley, der durch viele konstruktive Diskussionen den IL-6 Artikel verbessert hat, sowie allen Kooperationspartnern in meinen Projekten: Dennis Della Corte, Timo Fettweiss, Alexander Fulton, Rudolf Hartmann, Karl-Erich Jäger, Marco Kaschner, Frank Krause, Ulrich Krauss, Jakub Kubiak, Ralf Kühnemuth, Justin Lecher, Peixiang Ma, Christina Möller, Jeannine Mohrlüder, Philipp Neudecker, Christina Nutschel, Olujide Olubiyi, Vineet Panwalkar, Gunnar Schröder, Melanie Schwarten, Claus Seidel, Andreas Stadler, Matthias Stoldt und Oliver Weiergräber.

Dank geht auch an das ICS-6 am Forschungszentrum Jülich unter der Leitung von Prof. Dieter Willbold sowie an die Graduiertenschule iGRASPseed für die finanzielle Förderung.

Und vielen Dank an alle Gäste meiner Endspurt-Überraschungsparty, auch an die, die nicht kommen konnten. Ihr habt mir für die letzten Tage nochmal Kraft gegeben. Und nochmals vielen Dank an dich, Nicole, für die enorm aufwändige und geheime Organisation und Planung. Ich habe absolut nichts gemerkt und war völlig überrascht!

Kurzfassung

Klassische Moleküldynamik (MD) Simulationen und experimentelle Methoden zur Erforschung von Biomolekülen, vor allem Proteinen, können einander ergänzen und voneinander profitieren. Viele experimentelle Methoden messen die Eigenschaften von Molekülen nur indirekt, zum Beispiel als Veränderungen in elektromagnetischen Feldern oder durch die Detektion gestreuter Photonen. MD Simulationen dagegen beschreiben die Bewegungen jedes Atoms in einem molekularen System direkt. Sie sammeln jedoch nur unzureichende Daten über den gesamten Phasenraum des untersuchten Proteins. Außerdem enthalten die klassischen Kraftfelder in MD Simulationen zahlreiche Annäherungen und ungenaue Parametrisierungen zur Beschreibung der Dynamik von Biomolekülen. Ungeachtet dessen sind MD Simulationen in den letzten 40 Jahren erfolgreich in tausenden Studien von Proteinen zur Anwendung gekommen. In dieser Dissertation wird in vier unabhängigen Projekten gezeigt wie experimentelle Methoden und MD Simulationen kombiniert werden können um neue Hypothesen zu Proteinstrukturen, -dynamik und Interaktionsmechanismen zu entwickeln und zu testen. Experimentelle Beobachtungen können aus den MD Rohdaten vorhergesagt und mit experimentellen Daten verglichen werden um die Korrektheit der Simulationen zu belegen und Vorhersagen zu machen.

Abstract

Classical Molecular dynamics (MD) simulations and experimental methods used to study biomolecules, most notably proteins, can both complement and benefit each other. Many experimental techniques measure molecular properties indirectly, for instance, as changes in electromagnetic fields or through scattered photons. MD simulations, on the other hand, have direct access to the motions of each atom in a molecular system. However, they suffer from incomplete sampling of the phase space of the protein under study. Moreover, they are based on classical force fields, which involve certain approximations and parameterizations for the description of biomolecules. Nonetheless, over the past 40 years, thousands of MD simulations have been successfully applied to protein systems. In this dissertation, it is demonstrated in four independent projects, how experimental methods and MD simulations can be combined to arrive at new models of protein structures and dynamics and test hypotheses on molecular assembly mechanisms. Experimental observables are predicted from the raw MD data and compared with experimental data to assess the validity of the simulations, which in turn allow to make predictions from the MD data.

Contents

Chapter 1

Introduction

Experimental methods to study biomolecules yield only indirect information on the system of interest and hence are subject to interpretation in the light of the underlying theory. Moreover, the different experimental techniques suffer from non-natural conditions applied to the proteins, which are, however, necessary for performing the experiments and data acquisition.

X-ray crystallography measures the scattering of photons due to the electrons in a crystal grown from biomolecules. The crystals are typically obtained at low or high pH values, in the presence of high salt concentrations and organic co-solvents or heavy metals. Under optimal experimental conditions, the scattering pattern enables the calculation of very accurate electron density maps. However, the crystal structures modelled into the electron density maps represent only the thermodynamic ground state in the crystal. In solution, proteins sample many more states, especially in loops and regions of reduced order. Hence the crystal structures represent only one or the mean of many equally likely conformational states. In addition, amino acid side chains and loops at the interface of adjacent proteins in the crystal may assume conformations that exist predominantly in the crystal, but not in solution and are thus artifacts of the crystallization process.

NMR-experiments, the second most frequently used protein structure elucidation method, measures chemical shifts and relaxation rates of the magnetic moments of atomic nuclei in the protein. In order to obtain strong enough signals, NMR experiments have to be performed at high protein concentrations. From the raw data, which often shows changes to the data measured under different conditions, distances between pairs of amino acids can be deduced as well as information on local flexibilities of the protein backbone. The distances are used as constraints to generate protein structures. As several structures can be generated that match the distance constraints equally well, it is common to publish a whole ensemble of structures, hence reflecting the flexibility of the protein. NMR tends to predict protein structures well for ordered regions, yet it is unable to detect signals for very mobile protein regions. Moreover, structure generation from NMR data is a tedious method that involves molecular modelling steps. The relaxation data, the chemical shifts and data on the couplings between nuclear spins are usually used to come up with hypotheses about the dynamics and function of proteins, often with the aid of already known structures.

Small angle X-ray scattering (SAXS) experiments yield structural data of very low resolution. Several quantities like the molecular mass and properties of the overall molecular shape, such as radii of gyration, can be readily computed from the scattering data. However, hypotheses made about molecular processes underlying the changes observed in the scattering data cannot be validated without further, complementary evidence.

Fluorescence spectroscopy is another a method that allows measuring intra- and intermolecular distances and the dynamical properties of protein side chains and larger subunits. In contrast to NMR, a single experiment only yields information on the dynamical behaviour of a single residue. A further weakness of this method is the requirement of the attachment of fluorescent dyes to residues mutated to cysteine, which may affect the protein structure. Hence, models of molecular dynamics based on fluorescence spectroscopy also require further validation.

Molecular Dynamics (MD) simulations enable to study the dynamics of proteins on an atomistic level. The resulting trajectories can be viewed like videos of atomic and molecular motion. More importantly, statistical data of the various states that the molecule visits can be computed. For many experimental methods, the underlying theory permits the computation of experimental observables directly from the positions and movements of the atoms in the protein. MD simulations can therefore be validated against experimental data, assessing if they sample all relevant molecular conformations with the correct probabilities. More importantly, MD simulations give direct evidence of the structures and the molecular processes giving rise to the signals measured in experiments.

In this dissertation, four independent projects give examples of the successful complementation of experimental methods by MD simulations. Each project combines MD derived models of protein structures and dynamics with data from one or more experimental methods and demonstrates the added value that each approach yields. In the first project, a dimer model is constructed from two protein monomers of known structure and validated against experimental SAXS data (section 3.1). In the second project, several C-terminal states of the autophagy-related protein GATE-16 are characterized by MD simulations and combined with data from X-ray crystallography and NMR spectroscopy (section 3.2). A model of the assembly mechanism of the cytokine interleukin-6 with its receptors is generated based on multiple MD simulations in the third project (section 3.3), and the MD data for apo interleukin-6 validated against NMR results. A comprehensive and comparative study of the pico- to nanosecond dynamics of GABARAP, a homologue of GATE-16, is presented and a unified framework for the interpretation of data from NMR, fluorescence spectroscopy and MD simulations developed (section 3.4).

Chapter 2

Scientific Context

2.1 Experimental Methods

2.1.1 X-Ray Crystallography

X-ray crystallography is a technique that enables the determination of the atomic structure of molecular crystals at high resolution of up to ~1.0 Å. It is based on the analysis of diffraction patterns of an X-ray beam after passing through a molecular crystal [1]. According to Bragg's Law ($2d\sin(\theta) = n\lambda$, Figure 2.1), waves scattered by a uniform crystal lattice form an interference pattern with points of maximum positive interference at angles θ (d is the distance of lattice points, λ the wavelength of the incident X-ray beam and n an integer number).

It is possible, although sometimes difficult, to grow crystals from soluble biomolecules, most notably proteins. Crystals are usually grown under very unphysical conditions of extreme salt concentrations and pH values and often together with heavy metal ions. The X-ray diffraction patterns of these crystals are very complex (Figure 2.2). They are a representation of the crystal lattice in reciprocal space, which is the Fourier transform of the direct lattice. The signal intensity of the diffraction pattern is recorded as $F(\mathbf{q})$, where $\mathbf{q}$ is the reciprocal lattice vector. From the diffraction intensity, the electron density in the direct

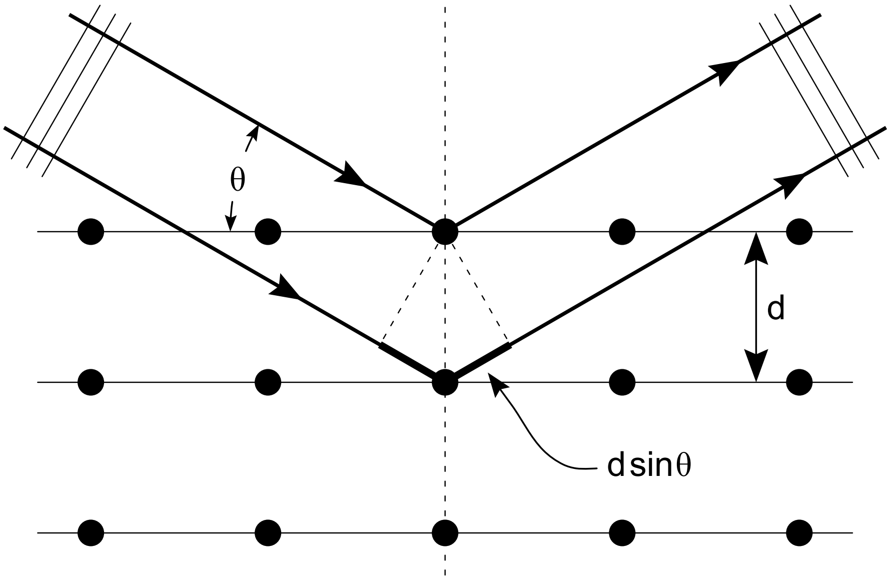


Figure 2.1: Bragg diffraction from a cubic crystal lattice. Figure reused from [2], available under a creative commons license.

crystal lattice, $f(\mathbf{r})$, can be computed by a complex Fourier transform:

$$f(\mathbf{r}) = \frac{1}{(2\pi)^3} \int F(\mathbf{q}) e^{i\mathbf{q}\cdot\mathbf{r}} \mathrm{d}\mathbf{q} \tag{2.1}$$

The problem is, that the complex value of $F(\mathbf{q})$ cannot be obtained from the diffraction pattern exactly, but rather only its magnitude, while the information on the phase $\phi(\mathbf{q})$ is lost:

$$F(\mathbf{q}) = |F(\mathbf{q})| \cdot e^{i\phi(\mathbf{q})} \tag{2.2}$$

Multiple solutions for the phase problem exist. Phases can be determined from diffraction datasets collected at different wave lengths (Multi wavelength Anomalous Dispersion, MAD) [3, 4], heavy atoms can be introduced to the crystal, either by co-crystallization or through soaking of the crystal with the heavy atoms (Multiple Isomorphus Replacement, MIP) [5], or by Molecular Replacement (MR), in which a structurally very similar homologous reference protein is used to determine the phases [1].

To judge the quality of the structural model generated from the X-ray scattering

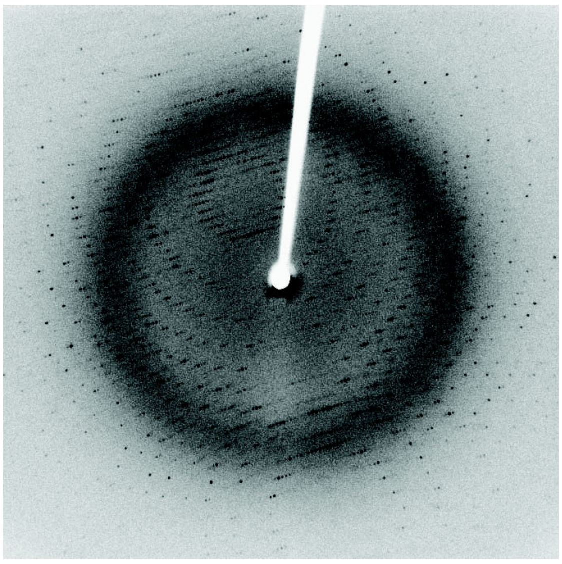


Figure 2.2: X-ray diffraction pattern of crystallized 3Clpro, a SARS protease. (2.1 Åresolution). Figure reused from [6], available under a creative commons license.

data, the agreement of the observed scattering amplitudes (F_o) with amplitudes calculated from the model (F_c) is computed as the R-factor:

$$R = \frac{\sum \|F_o(\boldsymbol{q}) - F_c(\boldsymbol{q})\|}{\sum F_o(\boldsymbol{q})}. \tag{2.3}$$

Smaller values of the R-factor denote models that are in better agreement with the experimental data.

The resolution at which a protein structure has been determined gives information about the smallest observable features in the structure, i.e., the distance beyond which two atoms appear as clearly separate structures in the electron density map. It does not equal the uncertainty in the atom position, which is generally lower. At a resolution of 5.0 Å, the backbone can already be confidently traced through the electron density and at 2.5 Å resolution the side chain conformations start to be reliable. As the resolution is a global value for the complete protein structure, some amino acids and most likely the side chains may have lower resolutions and hence care has to be taken when interpreting structures.

Apart from this, protein structures solved by X-ray crystallography suffer from additional uncertainties. Due to the crystal packing of normally soluble proteins, unnatural contacts at the protein interfaces arise. Amino acids at the surface of globular proteins, which are solvent exposed in the cellular environment, can make direct contacts with other protein molecules in the crystal. These interactions may have a distorting effect, especially on flexible parts of the protein like loop regions and termini. In addition, the heavy atoms and other agents introduced into the crystal to aid the crystallization process or data analysis, may also have locally distorting effects on the protein structure.

2.1.2 NMR – Nuclear Magnetic Resonance Spectroscopy

Nuclear Magnetic Resonance Spectroscopy (NMR) exploits the magnetic properties of atomic nuclei to obtain information about molecular structures. When an NMR active atomic nucleus is placed in a magnetic field, it absorbs and re-emits electromagnetic radiation at an isotope specific frequency, called its resonance frequency [7]. NMR active nuclei are characterized by an odd number of either protons or neutrons or both, while nuclei with an even number of both, protons and neutrons, are NMR inactive. In NMR inactive nuclei are characterized by a total nucleic spin of 0. Nuclei with an odd number of both, protons and neutrons, possess an integer total spin, while nuclei with an odd number of either protons or neutrons (and an even number of the other nucleon) possess half integer spin. This non-zero spin gives rise to an intrinsic magnetic moment and angular momentum. Isotopes that are most frequently used in NMR spectroscopy are ^{1}H and ^{13}C. While ^{1}H is the most abundant isotope of hydrogen (99.98%), ^{13}C has a natural abundance of only 1.1% and thus it needs to be introduced into molecules in artificially high concentrations during synthesis.

The intrinsic magnetic moment of an NMR active nucleus interacts with an externally applied magnetic field. Its energy due to the magnetic field is minimized by aligning with the direction of the magnetic field and it is maximized by pointing in the opposite direction. After placing a substance into the strong magnetic field of an NMR spectrometer, all active nuclei will equilibrate after some time and their magnetic moments will thus align with the external field. They can then be excited from this ground state by an electromagnetic pulse in the radio frequency range (60-1000 MHz). After excitation, the magnetic moments gradually relax back to the ground state while emitting radiation.

Two complementary descriptions of NMR spectroscopy exist. In the classical description, nuclear spins excited from the ground state rotate in the magnetic field and oscillate between parallel and anti-parallel alignments with the direction of the external field. They rotate at the isotope specific Larmor frequency ω, which is anti-proportional to the external magnetic field (B) by a proportionality constant γ, called the gyromagnetic ratio:

$$\omega = -\gamma B \tag{2.4}$$

Due to their rotation, they emit electromagnetic radiation, whose intensity decays, while the magnetic moment returns to its ground state aligned with the field lines.

In the complementary quantum mechanical description, nuclei are excited from their ground state to an excited state by photons from the external radio frequency pulse. When falling back to the ground state, they emit photons of the energy difference, which are recorded as electromagnetic radiation. The frequency of this radiation is proportional to the energy difference of the excited and ground states of the nuclei.

The electromagnetic signal measured during data acquisition of an NMR experiment is called the Free Induction Decay (FID). It is a multi-exponentially decaying signal, characterized by multiple decay time constants. Two characteristic timescales are associated with the decay from the excited to the ground state. The timescale on which nuclei realign with the external field is called "spin-lattice" or longitudinal magnetic relaxation and is characterized by the timescale T_1. In the classical view, rotating magnetic moments can also dephase, i.e., desynchronize their rotations. This process is called transverse relaxation and denoted with the timescale T_2.

2.1.2.1 NMR Order Parameter

The relaxation data obtained in NMR experiments contains information on internal motions present in the molecular system under study. Models based on physical information or ease of formulation are used for the analysis and interpretation of the relaxation data, e.g., diffusion of a bond vector in a cone or on the surface of a cone. Despite being useful, all of these models pose the risk of over interpretation of limited amounts of data. A ”model-independent” approach has been invented to interpret the relaxation data without the need to resort to a particular physical model [8].

In an isotropically reorienting protein, the orientations of all bond vectors fluctuate with respect to the external magnetic field due to fast internal motions and slower overall rotation of the protein. As all amino acids (except proline) share the same kind of backbone atoms and nitrogen and hydrogen are both NMR active nuclei, the fluctuations of the backbone amide N–H bond vectors are usually assessed in NMR experiments of proteins. The internal motions are restricted by the structure and flexibility of the protein chain at a particular residue. The motion of each bond vector can be described by two quantities: a generalized order parameter S, measuring the spatial restriction of the motion, and an effective correlation time τ_e which describes the rate of the motion. If the overall motion is isotropic, the spectral density $J(\omega)$ can be formulated as (see Reference [8]):

$$J(\omega) = \frac{2}{5}\left(\frac{S^2\tau_M}{1+(\tau_M\omega)^2} + \frac{(1-S^2)\tau}{(1+\tau\omega)^2}\right), \tag{2.5}$$

where with τ_M is the correlation time of overall rotation of the protein and τ is defined as:

$$\tau^{-1} = \tau_M^{-1} + \tau_e^{-1}. \tag{2.6}$$

The spectral density is the Fourier transform of the time-autocorrelation function of the fluctuating magnetic dipole interactions of nuclear spins. It describes a probability distribution of dynamic processes in the system with frequency ω. The relaxation times T_1 and T_2 can be expressed as functions of the spectral densities of the NMR active nuclei in the system. The values of S^2 and τ_e are obtained by fitting relaxation parameters calculated with equation 2.5 to experimental values for T_1 and T_2, where S and τ_e are the only adjustable parameters. Values of S^2 close to 1 originate from internal rigidity of the corresponding N–H bond and decreasing S^2 values towards 0 reflect increasing flexibility.

The S^2 order parameter directly relates to the internal motions of N–H bonds in a protein. The bond vector motions in a rigid reference frame attached to the protein can be cast into time autocorrelation functions defined as [9]:

$$C_I(\tau) = \langle P_2 [\hat{\boldsymbol{\mu}}(t) \cdot \hat{\boldsymbol{\mu}}(t+\tau)] \rangle , \tag{2.7}$$

where $\mu(t)$ denotes the unit length bond vector, $P_2 [x]$ is the second order Legendre Polynomial and angular brackets denote averaging over time. The autocorrelation function decays exponentially to a convergence value proportional to the S^2 order parameter.

2.1.2.2 Chemical Shift

Atomic nuclei do not all experience the same external magnetic field. Small distortions of the field due to the magnetic moments of other atoms in the local environment give rise to individual magnetic fields felt by each nucleus. The differences are proportional to the external magnetic field strength and hence stronger magnetic fields are more sensitive in the detection of individual nuclei, as the differences in the emission frequencies increase with field strength. The signal radiated during transitions from the excited to the ground state is measured as the FID. The FID can be Fourier transformed to obtain information on the frequency distribution. Each NMR active nucleus can therefore be assigned a frequency that depends on the external magnetic field and the local environment. The chemical shift of a nucleus is defined as

$$\delta = \frac{\nu_{\text{sample}} - \nu_{\text{ref}}}{\nu_{\text{ref}}}, \tag{2.8}$$

where ν_{sample} is the resonance frequency of the nucleus and ν_{ref} is the resonance frequency of a reference compound in the same magnetic field. ^{1}H and ^{13}C chemical shifts are usually referenced against tetramethylsilane (TMS) and this compound therefore has a chemical shift of zero. As the units of numerator and denominator in equation 2.8 both have units of Hz, the chemical shift is a unitless quantity. To make the usually very small quantity easier to handle and report, it is divided by 10^{-6} and thus given as *parts per million* (ppm).

The chemical shift can be used to identify chemical compounds. In protein NMR spectroscopy it serves a prominent role, as chemical shifts give hints on the secondary structure of the protein and changes in chemical shifts indicate local flexibilities.

Currently, NMR spectroscopy and X-ray crystallography are the dominating methods for protein structure elucidation, with more than 110,000 structures solved by X-ray crystallography and more than 11,000 structures solved by NMR spectroscopy (as of April 2017). Recently, Electron Microscopy (EM) is rising as a third method. In 2016, almost as many structures were solved with EM as with NMR spectroscopy (408 by EM versus 456 by NMR). EM is particularly suited to determine the structures of larger protein complexes, which are constructed from low resolution experimental data, often in combination with atomistic models of the known structures of subunits of the complex.

2.1.3 Small Angle X-Ray Scattering

Many of the challenges posed by X-ray crystallography can be circumvented by Small Angle X-ray Scattering experiments. SAXS measurements can be performed in solution, without the challenge of finding suitable crystallization conditions and hence the sample preparation is significantly simpler and requires less material. Also, the range of system sizes that can be studied by SAXS is comparably large, ranging from small molecules to large macromolecular assemblies, the latter far too big to be solved by NMR spectroscopy. However, these features come at the price of drastically reduced resolution of maximally 10.5 Å, a resolution range still high enough to distinguish molecules based on their size and shape. This information can help to validate high-resolution models of proteins [10] or to guide the modelling process of macromolecular assemblies [11].

2.1.3.1 Experimental Procedure

In SAXS experiments, a protein sample in solution is inserted into a collimated, monochromatic X-ray beam, originating ideally from a synchrotron storage ring (Figure 2.3) [12]. The angle dependent intensity $I(s)$ of the radiation scattered by the sample is detected at small angles between 0.1 and 10 degree as a function of momentum transfer $s = 4\pi \sin(\theta)/\lambda \approx \theta/\lambda$, where 2θ is the angle between the incident beam and the scattered signal and λ is the wavelength of the incident radiation. The name of s (momentum transfer) hints to its interpretation as the amount of momentum that is deflected from its original direction in the X-ray beam. In the literature, the letter q is sometimes used in place of s (Figure 2.3). In contrast to X-ray crystallography, SAXS measures an orientationally averaged, isotropic scattering intensity, which is detected as a radially symmetric signal.

SAXS is a contrast method, in which the scattering intensity of the solvent alone is subtracted from the intensity of the sample in solution. The magnitude of the remaining signal depends on the difference in electron density of solvent and solute ($\Delta\rho$) and hence denser systems give rise to larger signals ($\rho_{\text{water}} \approx 0.33$ e$^-$/Å^3; $\rho_{\text{protein}} \approx 0.44$ e$^-$/Å^3).

$I(s)$ decays quickly for small values of s. The intensity at very small angles contains information about the low resolution shape of the molecule, at medium angles about the overall fold and at larger angles (where the information content is very low and scattering curves from different molecules become very similar) about the atomic structure (Figure 2.4).

Several quantities can be directly computed from the scattering intensity. The scattering intensity at $s = 0$ is proportional to mass and volume of the scattering particles, but it needs to be extrapolated from the measured values, as it is superimposed with the original incident X-ray beam on the detector. The assembly state of macromolecules can also be determined from the estimations of the particle volume and molecular weight. The radius of gyration, R_G, is directly accessible from the logarithm of $I(s)$:

$$\ln|I(s)| = \ln|I(0)| - \frac{s^3 R_G^2}{3}. \tag{2.9}$$

The scattering intensity also depends on the electron distribution (or pair distribution function) of the scattering particles as:

$$I(s) = 4\pi \int_0^{D_{\max}} P(r) \frac{\sin(sr)}{sr} \mathrm{d}r, \tag{2.10}$$

where $D_{\max}$ is the maximum distance in a solute particle and r is the distance in the particle from an arbitrary origin.

$P(r)$ can be directly calculated from the scattering curve by means of Fourier transformation and presents information of the distances between electrons in the

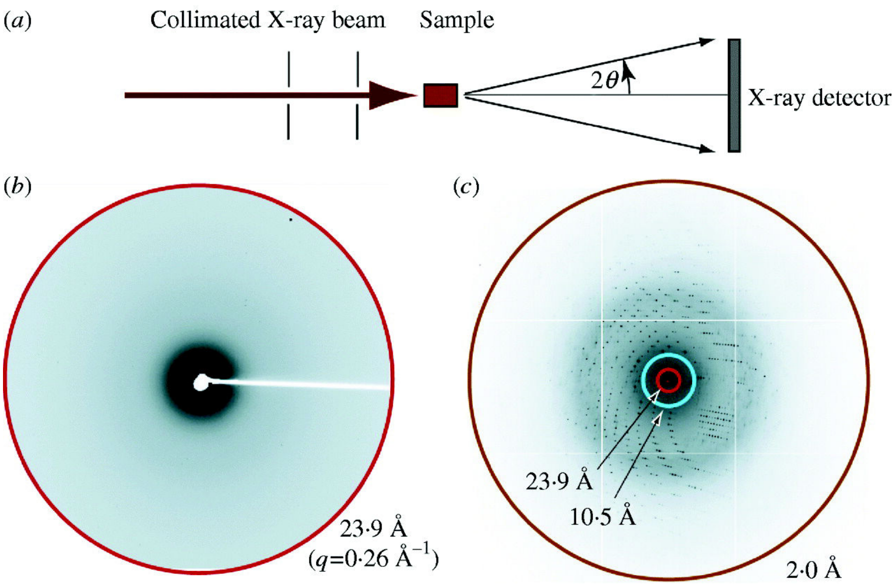


Figure 2.3: Schematic description of a SAXS experiment. (**a**) The sample is inserted into the X-ray beam and the scattering pattern is recorded. (**b**) Scattering pattern from a SAXS experiment in solution with a maximum resolution of 23.9 Å. Here the momentum transfer is indicates with q in place of s. (**c**) Scattering pattern from a protein crystal of 2.0 Å resolution. Red circles indicate radii of equivalent resolution of crystallography and SAXS experiments. The cyan circle indicates the maximum resolution achievable in SAXS (10.5 Å). Figure reused with permission from reference [13].

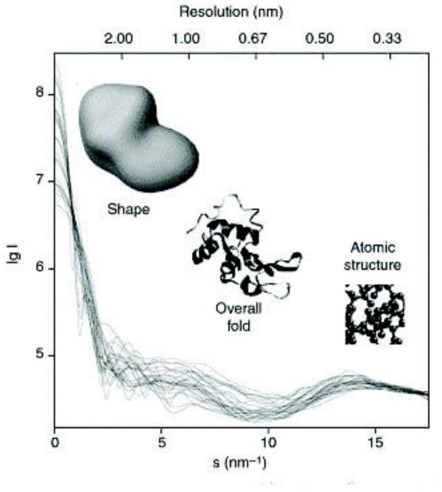


Figure 2.4: Information content at different momentum transfer ranges. The X-ray solution scattering curves were computed from atomic models of twenty-five different proteins with molecular masses between 10 and 300 kDa. Figure reused from [14], available under a creative commons license.

scattering particle. It is related to the electron density in the sample as:

$$P(r) = r^2 \left\langle \int_V \Delta\rho(\boldsymbol{r})\Delta\rho(\boldsymbol{u}+\boldsymbol{r})\mathrm{d}\boldsymbol{r} \right\rangle_\Omega , \qquad (2.11)$$

where the integration inside angular brackets gives the spatial autocorrelation function of $\Delta\rho$ over the total volume V. The angular brackets represent radial averaging of the electron density autocorrelation over the total volume, indicated by Ω to distinguish radial averaging from the autocorrelation integration. $P(r)$ can also be directly computed from computational models and presents a way of comparing experimental data to structural models. From $P(r)$, the overall shape of the molecule can be deduced, e.g., whether it is spherical, rod- or disc-shaped.

2.1.3.2 Scattering Curve Prediction

The solution scattering curves of a model structure can be readily computed from the pair distribution function $P(r)$ (Equation 2.10). To compare a model to an experimental scattering curve, several quantities were suggested. In analogy to the R-factor computed in X-ray crystallography (Equation 2.3), the discrepancy

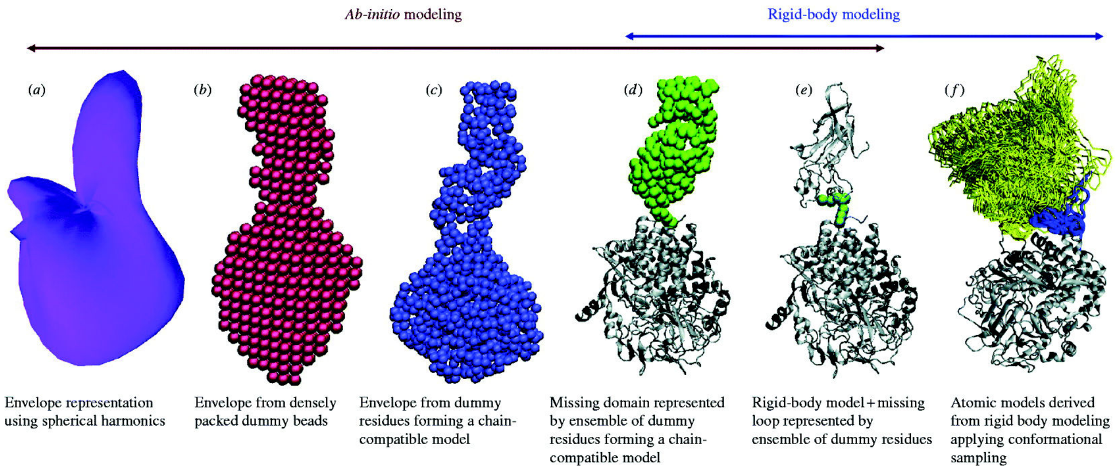

Ab-initio modeling
Rigid-body modeling
(*a*)
(*b*)
(*c*)
(*d*)
(*e*)
(*f*)
Envelope representation using spherical harmonics
Envelope from densely packed dummy beads
Envelope from dummy residues forming a chain-compatible model
Missing domain represented by ensemble of dummy residues forming a chain-compatible model
Rigid-body model+missing loop represented by ensemble of dummy residues
Atomic models derived from rigid body modeling applying conformational sampling

Figure 2.5: *In silico* models generated with different approaches from the solution scattering data of a complexed cellulase protein. Different *ab initio* modelling approaches allow the reconstruction of molecular envelopes (**a**), densely packed bead models (**b**), chain compatible models of dummy residues (**c**), chain compatible models of missing parts of the structure (**d**), rigid body fits of known atomic structures (**e**), and conformational ensembles of known atomistic structures (**f**). Figure reused with permission from reference [13].

between the calculated and experimental scattering curves can be computed:

$$R_{\mathrm{SAXS}} = \frac{\sum_i |I(s_i)_{\mathrm{exp}} - I(s_i)_{\mathrm{calc}}|}{\sum_i |I(s_i)_{\mathrm{exp}}|}, \tag{2.12}$$

where the sum runs over all N_P measurement points indexed by i. It is, however, more common to weight the scattering intensity difference by the experimental error ($\sigma(s)$) and compute the normalized discrepancy function:

$$\chi^2 = \frac{1}{N_P - 1} \sum_i \left[\frac{I(s_i)_{\mathrm{exp}} - cI(s_i)_{\mathrm{calc}}}{\sigma(s_i)} \right]^2, \tag{2.13}$$

where c is a scaling factor. χ^2, which is used in the SAXS data analysis program CRYSOL [15], gives larger weight to lower resolution data.

With the ability to evaluate model structures against the solution scattering data, attempts can be made to generate models from the experimental scattering curves (Figure 2.5).

In order to compare predicted scattering curves to experimental data, the excluded solvent volume occupied by the solute and a hydration layer of ordered solvent molecules need to be taken into account. The excluded solvent is usually modelled as continuous electron density equal to that of the real solvent. Hydration layers are either taken into account explicitly, by placing solvent molecules around the solute, or by a continuous envelope of a suitable empirical electron density.

The pair distribution function used to derive the theoretical scattering curve can be computed directly from the atom positions. However, this scales with the square of the number of atoms and becomes quickly unfeasible. Faster alternatives employ Monte Carlo integration routines or multipole expansions using spherical harmonic functions.

The degree of detail of models that can be constructed from SAXS data ranges from crude envelopes to atomistic ensembles. The envelopes are constructed from various basic shapes of continuous electron density or from spherical harmonic functions (Figure 2.5a). Instead of a continuous density envelope, a space filling bead model can be generated, approximating the envelope model (Figure 2.5b). A greater degree of detail is achieved by chain models that restrict dummy residues consisting of a single particle into conformations that trace a realistic chain, while matching the SAXS scattering curve as well as possible (Figure 2.5c). These methods are all summarized under the term *ab initio* models, as the only reference they require is the experimental data without resorting to known structural information.

In addition to *ab initio* models, it is possible to dock rigid bodies of known atomistic structure into the low resolution envelopes. This can be done for monomeric rigid models with the program SASREF and for multimeric models, with the additional complication of correctly predicting the mutual orientation, by the program BUNCH [11] (Figures 2.5d and e). Finally, multiple models can be generated that fit the experimental data comparably well, leading to a whole ensemble of structures, as with NMR spectroscopy (Figure 2.5f). In addition, the rigid docking methods can be combined, for example, to produce a multimeric model of rigid bodies, whose individual structures are known, connected by linkers of known sequence but unknown structure (Figure 2.5e).

An excellent review of Small Angle X-ray Scattering (SAXS) has been published

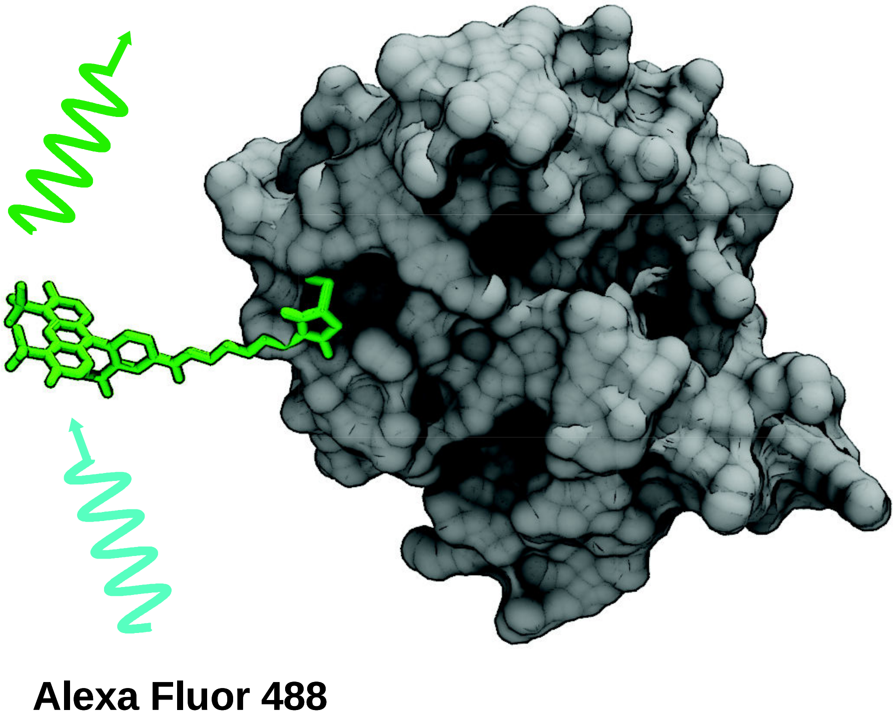


Figure 2.6: Fluorescence mechanism: A protein molecule (gray) is labeled with a Alexa Fluor 488 dye (green). The dye is excited with blue light and it can then emit green light

by Putname et al. in 2007 [13], discussing the theoretical background and data analysis methods.

2.1.4 Fluorescence

While the aforementioned methods are all ensemble methods, in the sense that they measure properties of a huge ensemble of molecules in the test tube, fluorescence is a physical phenomenon that enables single-molecule experiments. In a fluorescence spectroscopy experiment, a fluorescent dye is excited by electromagnetic radiation, often in the visible range and the intensity, the fluorescence time and the polarization of the emitted light can be detected (Figure 2.6).

To study protein structure and dynamics with fluorescence spectroscopy, the dye is attached to a macromolecule at carefully selected positions, one at a time.

The amino acids at these positions are first mutated to cysteine. Linker molecules, together with the dyes, are then covalently linked to the cysteine residues. In each experiment a different residue can be labeled. The dye positions are selected for positions that can yield information on the mechanism under investigation. Care needs to be taken to select positions, where a large linker and dye does not disrupt the proteins structure. Good choices are residues at the surface with solvent exposed side chains.

A dilute solution of the protein is then excited by a laser beam in a confocal microscope. The excitation volume and protein concentration are so small that only a single molecule is present at a time. Hence, fluorescence spectroscopy is a true single-molecule experiment, as it studies a single molecule at a time.

The fluorescence is anisotropic in the sense that the emitted light contains parallel and perpendicularly polarized components with respect to the excitation light beam. The donor anisotropy, r_{D}, is a measure for the magnitudes of the parallel (F_p) and perpendicular (F_s) fluorescence intensities emitted at time t_c after excitation at time t [16]:

$$r_{\mathrm{D}}(t+t_c) = \frac{F_p(t+t_c) - F_s(t+t_c)}{F_p(t+t_c) + 2F_s(t+t_c)} \tag{2.14}$$

It gets maximal for purely parallel fluorescence and minimal for purely perpendicular fluorescence. The time dependence of the anisotropy is very similar in its mathematical form to the N–H bond vector correlation function used to describe S^2 order parameters in NMR experiments (equation 2.7):

$$r_D(t+t_c) = r_0 C(t_0) = \langle P_2 \left[\hat{\boldsymbol{\mu}}_a(t) \cdot \hat{\boldsymbol{\mu}}_e(t+t_c) \right] \rangle \,, \tag{2.15}$$

where $\hat{\boldsymbol{\mu}}_a$ and $\hat{\boldsymbol{\mu}}_e$ are the transition dipole vectors of the fluorescent dyes for absorption and emission scaled to unit length.

Analogous to the bond vector autocorrelation function, the anisotropy can be described as a multi-exponential decay. The sources of the decay are internal motions and overall rotational diffusion, conceptually equivalent to the mechanisms behind the S^2 order parameter. In NMR, motions of the N–H bond vector are probed, while fluorescence spectroscopy yield information on the mobility of the dyes, which is hypothesized to reflect the side chain mobilities in the absence of dye and linker.

2.2 Simulation Methods and Analysis

2.2.1 Molecular Dynamics Simulations

The main method that was used in the presented research are molecular dynamics (MD) simulations based on classical molecular mechanics (MM) [17]. In molecular mechanics, molecules are modelled as systems of chemically inert atoms that interact through Newtonian mechanics, governed by Newton's second law:

$$\boldsymbol{F} = m\boldsymbol{a}, \tag{2.16}$$

stating that the acceleration $\boldsymbol{a}$ that a body undergoes is proportional to the force $\boldsymbol{F}$ acting on that body with mass m.

The interaction energy between atoms is described by preferably simple mathematical equations, mainly harmonic and trigonometric functions, as well as the well known electrostatic and Lennard-Jones potentials. A given set of potential energy equations, together with their parameters, is called a force field, as it is used to compute the interatomic forces (see also section 2.2.2).

Several programs that perform MD simulations exist. GROMACS [18] was used for all projects described in this dissertation.

2.2.1.1 Numerical Integration of Newton's Equations

Newton's equations of motion are numerically integrated to obtain a trajectory in phase space. A naive approach would be to use a simple Euler integration scheme [19] using the Taylor expansion of the particle positions truncated after the second order term:

$$\boldsymbol{r}\left(t+\Delta t\right) = \boldsymbol{r}\left(t\right) + \boldsymbol{v}\left(t\right)\Delta t + \frac{\boldsymbol{f}\left(t\right)}{2m}\Delta t^2 + \mathcal{O}(\Delta t^3), \tag{2.17}$$

where m is the particle mass and $\boldsymbol{v}$ and $\boldsymbol{f}$ represent velocities and forces of all N particles respectively:

$$\boldsymbol{f} = \begin{bmatrix} f_{1,x} \\ f_{1,y} \\ f_{1,z} \\ \vdots \\ f_{N,x} \\ f_{N,y} \\ f_{N,z} \end{bmatrix}, \qquad \boldsymbol{v} = \begin{bmatrix} v_{1,x} \\ v_{1,y} \\ v_{1,z} \\ \vdots \\ v_{N,x} \\ v_{N,y} \\ v_{N,z} \end{bmatrix}, \tag{2.18}$$

where the first index denotes the particle and the second the vector components in x, y, and z direction. For simplicity all particles are assumed to be of equal mass here, hence m becomes a scalar. The velocity update in the Euler method is computed as:

$$\boldsymbol{v}\left(t + \Delta t\right) = \boldsymbol{v}\left(t\right) + \frac{\boldsymbol{f}\left(t\right)}{2m}\Delta t + +\mathcal{O}(\Delta t^2). \tag{2.19}$$

While this algorithm is reasonably simple, it turns out to be a particularly bad choice as it is irreversible and suffers from an inherent energy shift [20]. Several different integration schemes have been introduced over the years. It is easiest to start the discussion with the Verlet algorithm [21]. This algorithm is based on two Taylor expansions of the particle position, forward and backward in time:

$$\boldsymbol{r}\left(t + \Delta t\right) = \boldsymbol{r}\left(t\right) + \boldsymbol{v}\left(t\right)\Delta t + \frac{\boldsymbol{f}\left(t\right)}{2m}\Delta t^2 + \frac{\Delta t^3}{3!}\dddot{\boldsymbol{r}}\left(t\right) + \mathcal{O}\left(\Delta t^4\right) \tag{2.20}$$

and

$$\boldsymbol{r}\left(t - \Delta t\right) = \boldsymbol{r}\left(t\right) - \boldsymbol{v}\left(t\right)\Delta t + \frac{\boldsymbol{f}\left(t\right)}{2m}\Delta t^2 - \frac{\Delta t^3}{3!}\dddot{\boldsymbol{r}}\left(t\right) + \mathcal{O}\left(\Delta t^4\right). \tag{2.21}$$

Summing equations 2.20 and 2.21 and solving for $\boldsymbol{r}\left(t + \Delta t\right)$ yields the position update relation of the Verlet integrator:

$$\boldsymbol{r}\left(t + \Delta t\right) = 2\boldsymbol{r}\left(t\right) - \boldsymbol{r}\left(t - \Delta t\right) + \frac{\boldsymbol{f}\left(t\right)}{m}\Delta t^2 + \mathcal{O}(\Delta t^4) \tag{2.22}$$

It is interesting to note, that the Verlet algorithm computes the position update from the positions at the two preceding time steps, bypassing the computation of the velocities altogether. As the velocities are essential for an estimate of the kinetic energy and hence the temperature, they might still be computed as

$$\boldsymbol{v}\left(t\right) = \frac{\boldsymbol{r}\left(t - \Delta t\right) - \boldsymbol{r}\left(t + \Delta t\right)}{2\Delta t} + \mathcal{O}(\Delta t^2), \tag{2.23}$$

but only up to an accuracy of order $\mathcal{O}(\Delta t^2)$. The Verlet algorithm also suffers from the fact that it needs two subsequent initial time steps to start the integration procedure and define the velocities.

Integration schemes exist, that are equivalent to the Verlet integration in the sense that they give rise to the same (analytical) trajectories, but yield a higher accuracy for the velocity computation and the corresponding evaluations of kinetic energy and temperature. The Leap Frog integration algorithm starts out by subtracting the Taylor expansions of equations 2.20 and 2.21, rather than summing them up, and truncating after the third order term, therefore initially losing an order of accuracy:

$$\boldsymbol{v}\left(t\right) = \frac{\boldsymbol{r}\left(t + \Delta t\right) - \boldsymbol{r}\left(t - \Delta t\right)}{2\Delta t} + \mathcal{O}(\Delta t^3) \tag{2.24}$$

Shifting the result by $+\Delta t$ and dividing the time step by 2 gives:

$$\boldsymbol{v}\left(t + \Delta t/2\right) = \frac{\boldsymbol{r}\left(t + \Delta t\right) - \boldsymbol{r}\left(t\right)}{\Delta t} + \mathcal{O}(\Delta t^3) \tag{2.25}$$

Shifting equation 2.24 instead in opposite direction results in:

$$\boldsymbol{v}\left(t - \Delta t/2\right) = \frac{\boldsymbol{r}\left(t\right) - \boldsymbol{r}\left(t - \Delta t\right)}{\Delta t} + \mathcal{O}(\Delta t^3) \tag{2.26}$$

Equation 2.25 gives a direct way of updating the positions:

$$\boldsymbol{r}\left(t + \Delta t\right) = \boldsymbol{r}\left(t\right) + \boldsymbol{v}\left(t + \Delta t/2\right)\Delta t + \mathcal{O}(\Delta t^3) \tag{2.27}$$

To get a relation for the velocity update, equations 2.25 and 2.26 need to be subtracted and combined with equation 2.22:

$$\boldsymbol{v}\,(t+\Delta t/2) = \boldsymbol{v}\,(t-\Delta t/2) + \frac{\boldsymbol{f}\,(t)}{m}\Delta t + \mathcal{O}(\Delta t^3) \tag{2.28}$$

One order of accuracy for the velocity computation has been gained at the cost of one order of accuracy for the position update compared to the Verlet integration. An additional disadvantage of this method is the fact that positions and velocities are not computed at the same but at alternating half-integer time steps. This complicates the evaluation of potential, kinetic and total energies, as these are as well not defined at the same time. It is however possible to get rid of this additional problem by casting the Verlet algorithm in a different form, known as the velocity Verlet algorithm. Its equation for the position update is the same as in the Euler integrator. To arrive at the velocity update relation of the velocity Verlet algorithm we start out by adding both sides of the relation for the Verlet position update (equation 2.22) shifted by $+\Delta t$ to the naive Euler position update relation (equation 2.17). Rearranging and collecting like terms gives:

$$\boldsymbol{r}\,(t+2\Delta t) - \boldsymbol{r}\,(t+\Delta t) = \Delta t\boldsymbol{v}\,(t) + \frac{\Delta t^2}{2m}\left[2\boldsymbol{f}\,(t+\Delta t) + \boldsymbol{f}\,(t)\right] + \mathcal{O}(\Delta t^3) \tag{2.29}$$

Insertion of the Δt-shifted Euler position update relation into the last equation and solving for $\boldsymbol{v}\,(t+\Delta t)$ results in:

$$\boldsymbol{v}\,(t+\Delta t) = \boldsymbol{v}\,(t) + \frac{\Delta t}{2m}\left[\boldsymbol{f}\,(t+\Delta t) + \boldsymbol{f}\,(t)\right] + \mathcal{O}(\Delta t^3) \tag{2.30}$$

The velocity Verlet algorithm is therefore able to compute velocities and positions at the same integer time steps, while maintaining the same order of accuracy as the Leap Frog algorithm. If the accuracy of $\mathcal{O}(\Delta t^3$ is high enough, the velocity Verlet algorithm is the best choice among the presented algorithms.

2.2.1.2 Periodic Boundary Conditions

The number of atoms included in a simulated molecular system is at most a few million, and hence much smaller than real systems in the bulk, which contain atom numbers in the order of moles ($\sim 10^{23}$ particles). The small system size introduces strong unphysical artifacts at the boundaries of the system, where it interfaces the vacuum. It is possible to restrain the positions of the molecules at the vacuum interface, preventing the system from dissociation. However, this method is still insufficient, as atoms at the boundary only interact with atoms inside the system but miss interactions with the vacuum, which leads to surface tension at the boundary interface. A solution to this problem, which mimics the bulky conditions of real systems closest, is the introduction of periodic boundary conditions. Periodic boundary conditions allow particles to leave the box at one side and to reenter at the opposite side. More importantly, interactions act across the periodic boundaries and hence each particle experiences the simulation box as if it was located at its center.

2.2.1.3 Temperature and Pressure Control

According to the ergodic hypothesis in thermodynamics, the distribution of states visited by a system during an infinite time span equals the distribution of an infinite number of copies of the same system in equilibrium at one instant of time. In other words, averages of system properties over time will equal averages over states in phase space. These distributions of states at thermodynamic equilibrium are summarized in the concept of ensembles [22]. Each of a number of possible ensembles (i.e. distributions of states) is associated with system properties that are constant for all states of that ensemble. The thermodynamic ensemble that is produced by a molecular dynamics simulation is controlled by the macroscopic properties that are kept constant during the simulation The ensemble directly affects the behaviour of the system and therefore the measured quantities. A naive approach to molecular simulation would keep the number of particles (N), the system volume (V) and the total energy of the system (E) constant, as this does not require any control of extensive quantities, that are hard to influence. This will yield distributions from the microcanonical, or NVE ensemble, which does not correspond to any situation that is easy to create in a laboratory. Two other ensembles result in distributions that are much more directly comparable to real world experiment. The canonical, or NVT ensemble, requires a proper control of temperature (T) and reflects experimental conditions of constant volume and constant temperature. Another ensemble, the NPT ensemble, needs an additional mechanism to control the pressure (P). NPT is the desired ensemble for most biomolecular simulations, as it corresponds directly to *in vivo* situations, where the pressure and temperature are controlled by the environment and in general thought to remain constant on chemically relevant time scales.

Several algorithms exist for temperature control during MD simulations. The Berendsen algorithm [23] rescales the velocities in order to match the differential equation

$$\frac{dT}{dt} = \frac{T_0 - T}{\tau_\mathrm{T}}, \tag{2.31}$$

that couples the system exponentially with a time constant τ_T to a heat bath of temperature T_0. The Berendsen thermostat suppresses fluctuations of kinetic energy and therefore does not generate a proper thermodynamic ensemble. It might still be used for equilibration purposes.

The Nosé-Hoover thermostat [24, 25] extends the system Hamiltonian by coupling to a thermal reservoir with a friction term ξ. The reservoir has its own momentum p_ξ, a mass-like parameter Q, and it follows its own dynamic equations of motion:

$$\frac{dp_\xi}{dt} = (T - T_0), \tag{2.32}$$

coupling the reservoir momentum to a reference temperature T_0. The new equation of motion for the particles is:

$$\frac{d^2\boldsymbol{r}_i}{dt^2} = \frac{\boldsymbol{F}_i}{m_i} - \frac{p_\xi}{Q}\frac{d\boldsymbol{r}_i}{dt} \tag{2.33}$$

The Nosé-Hoover thermostat can be extended to couple the heat reservoir to another bath, which in turn couples to a further bath and so forth to generate a chain of baths. This Nosé-Hoover chain is guaranteed to be ergodic in the limit of infinite chain length. The system temperature under control of a Nosé-Hoover chain will oscillate around a reference value and it is therefore not suited for equilibration if the system starts out at a totally different temperature.

A third temperature control algorithm is the velocity rescaling thermostat [26], which, as the name suggests, rescales the velocities similar to the Berendsen thermostat with the improvement of an additional stochastic term that ensures the generation of a proper thermodynamic ensemble distribution.

Pressure control happens along the same lines of the algorithms outlined for temperature control. A Berendsen analog for pressure coupling exists and the Parinello-Rahman [27] and Martyna-Tuckerman-Tobias-Klein (MTTK) [28] algorithms are similar in spirit to the Nosé-Hoover thermostat, coupling the pressure to a pressure bath. One additional complication of pressure control is the definition and computation of the pressure in a bulk system. It is defined by the virial equation:

$$PV = Nk_{\mathrm{B}}T + \frac{1}{D}\left\langle \sum_{i=1}^{N} \boldsymbol{r}_i \boldsymbol{F}_i \right\rangle , \tag{2.34}$$

where D is the number of dimensions. The virial equation relates the pressure to the average over the forces acting on all particles. The pressure computed in this way is a highly fluctuating quantity and only its time average can be meaningfully interpreted.

2.2.2 Force Fields

An exact calculation of the behaviour of a molecular system involves solving the Schrödinger equation for all electrons and nuclei. As the computaional complexity for this problem grows extremely fast with increasing system size, the direct solution of all but rather small molecular systems is infeasible. Parameterized, analytical models have been developed for a simplified description of interatomic interactions in and between molecules that enable simulations of large molecular systems. The mathematical form of these equations is inspired by knowledge of the physics and chemistry that govern molecular interactions. The parameters are chosen to reproduce energy functions of quantum mechanical descriptions of model systems. As the large number of parameters, the choice of the model systems and the quantum chemical methods used for reference computations are all degrees of freedom in the paremterization, several different sets of parameters exist. A set of equations with its corresponding parameters is called a force field. This section first discusses the mathematical form of relatively simple, yet fast to compute, non-polarizable force fields and then gives reasons for the choice of a particular implementation.

2.2.2.1 Lennard-Jones Potential

A simple but often reasonable model of atomic structure treats atoms as nuclei surrounded by a diffuse electron cloud. Some elements possess a permanent dipole moment due to the shape and occupation of their electronic orbitals. Three kind of interactions between electron clouds give rise to the van der Waals interaction, which is the attractive contribution to the Lennard-Jones potential. First, two permanent dipoles interact and will attract or repel each other depending on their mutual alignment. The second contribution to the van der Waals interaction comes from dipoles induced in the electron cloud by permanent dipoles of neighbouring atoms. This interaction is attractive, as the induced dipole will have the same polarization as the inducing dipole due to repelling electron clouds. Finally, the electron clouds of two unpolarized atoms repel each other, therefore inducing dipoles that in turn give rise to a dipole–dipole interaction. The sum of these three effects is the source of the van der Waals interaction and gives rise to an overall attractive force that decays with the sixth square root of the interatomic distance.

Two atoms that approach each other will be affected by the Pauli exclusion principle that prohibits electrons from having the same state. This prevents two atoms to occupy the same space and will result in a repelling force at close distances. This and the dipole interactions of the electron clouds together are cast in the Lennard Jones potential that contains a term for short range repulsion and one for long range attraction (Figure 2.7):

$$V_{LJ}(r_{ij}) = 4\epsilon \left[\left(\frac{\sigma}{r_{ij}} \right)^{12} - \left(\frac{\sigma}{r_{ij}} \right)^{6} \right], \tag{2.35}$$

where ϵ is the depth of the potential well, which has its minimum at $r_m = 2^{1/6}\sigma$, and r_{ij} is the distance between atoms i and j. For each pair of atom types a force field needs to specify two parameters to define their equilibrium distance and the depth of the interaction well. Many force fields list parameters per atom

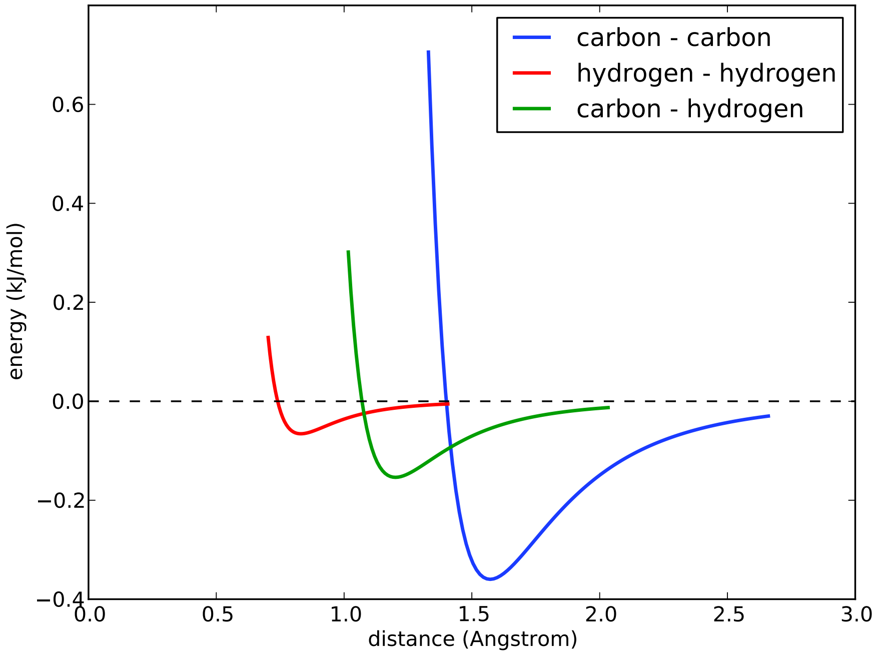


Figure 2.7: Lennard Jones potentials for carbon–carbon, hydrogen–hydrogen and carbon–hydrogen interactions.

type and give a relation for the computation of parameters for the Lennard-Jones interaction of two different atoms. The Lennard-Jones interaction converges fast to zero for large interatomic distances, and with it the resulting force goes to zero. This fast convergence makes a truncation at a suitable cutoff distance possible. However, for an accurate computation of the potential energy, the abrupt cutoff distance needs to be accounted for. This can be done by switch or shift functions that set the Lennard-Jones interaction smoothly to zero instead of a hard cutoff.

2.2.2.2 Coulomb potential

The computation of the Coulomb interaction is particularly tricky in systems with periodic boundary conditions (section 2.2.1.2), as the electrostatic energy and force decay slowly with increasing distance and suffers from ill convergence:

$$V_C(r_{ij}) = \frac{1}{4\pi\epsilon_0}\frac{q_i q_j}{r_{ij}}, \tag{2.36}$$

where q_i is the charge on atom i and ϵ_0 is the dielectric constant of vacuum. This problem is treated with sophisticated methods for long range electrostatic interactions such as the Particle-Mesh-Ewald (PME) method [29].The partial charges on each atom that are needed for the evaluation of the Coulomb potential are part of the force field parameter set and are usually derived from quantum chemical calculations.

2.2.2.3 Bonded Forces

Apart from the non-bonded inter-atomic forces there also exist interaction potentials due to bonds between the atoms of a molecule. Bond stretching (equation 2.37), bond angle bending (equation 2.38) and improper dihedral torsion angles (equation 2.40) can all be modelled by harmonic potentials, but more sophisticated mathematical forms exist as well. Proper dihedral torsion angle potentials are modelled by a cyclic cosine potential (equation 2.39), that is usually maximal for eclipsed and minimal for gauche conformations. Bond stretching is defined by an equilibrium distance between two atom types, r^0_{ij}, and an interaction constant k^b_{ij}. Bond angles are parameterized by an equilibrium angle between three atom types and a constant (θ^0_{ijk}, k^a_{ijk}). Dihedral and improper dihedral angles depend on interaction constants and equilibrium angles between planes spanned by three atoms ijk and jkl (ϕ^0_{ijkl}, ξ^0_{ijkl}, k^d_{ijkl} and k^{id}_{ijkl}).

$$\begin{aligned}
V_b(r_{ij}) &= \frac{1}{2}k^b_{ij}(r_{ij} - r^0_{ij})^2 && (2.37)\\
V_a(\theta_{ijk}) &= \frac{1}{2}k^a_{ijk}(\theta_{ijk} - \theta^0_{ijk})^2 && (2.38)\\
V_d(\phi_{ijkl}) &= k^d_{ijkl}(1 + \cos(n\phi_{ijkl} - \phi^0_{ijkl})) && (2.39)\\
V_{id}(\xi_{ijkl}) &= \frac{1}{2}k^{id}_{ijkl}(\xi_{ijkl} - \xi^0_{ijkl})^2 && (2.40)
\end{aligned}$$

The parameter n in equation 2.39 determines the periodicity of the dihedral angle potential, i.e., the number of minima and maxima. Dihedral angles with $n = 1$ have a single minimum, usually in either *cis* or *trans* conformation. Angles with larger n have more minima and maxima. More complicated potentials can be constructed from a combination of multiple dihedral angle potential with different periodicities and positions of the minima and are used to model nucleic acids.

A great variety of force fields exists and a careful and well informed choice about the set of equations and parameters that can model the system under investigation best has to be made. Recent benchmarks of protein force fields give insights into the strengths and weaknesses of the available options [30], indicating that one of the most up to date corrections to the widely used Amber force field (`amber99sb-ildn-nmr`) [31] is a very good choice for simulations of proteins in explicit solvent. For the cited benchmark a statistical evaluation of the performance of various force fields compared to real experiments was performed. A number of di- and tripeptides served as model systems for chemical shift measurements in NMR experiments. The deviations of the measurements from chemical shifts, computed on the basis of molecular dynamics trajectories performed with the investigated force fields, are reported. In general, older force field parameterization sets performed worse than newer corrections for all force fields studied. The chosen correction to the `amber99` force field performed best in simulations of ubiquitin, the only full length protein studied in the benchmark. All studies presented in this work make either use of the `amber99sb-ildn-nmr` force field or its predecessor `amber99sb-ildn`, which performs nearly as well in combination with the `TIP3P` water model.

2.2.3 Hamiltonian Replica Exchange MD

In order to sample the dynamics of a protein during an MD simulation with sufficient statistical accuracy, the sampling time needs to be longer than the timescale on which the dynamical process takes place. This is problematic for processes happening on the microsecond timescale or longer, as sampling times are limited to a few microseconds for systems consisting of 10,000 to 100,000 atoms on modern computing hardware (as of 2017). In order to speed up sampling of the energy landscape, the replica exchange method was invented [32]. The central idea is to replicate the system multiple times and run equivalent MD simulations of each replica in parallel at narrowly spaced increasing temperatures. The hotter systems will overcome energy barriers, separating stable states on the energy landscape, more quickly. At regular intervals, coordinate exchanges are performed. Each exchange is accepted with a probability of

$$p = \min\left(1, e^{(E_i - E_j)\cdot\left(\frac{1}{k_\mathrm{B}T_i} - \frac{1}{k_\mathrm{B}T_j}\right)}\right), \tag{2.41}$$

where i and j denote two separate replicas, E_i is the potential energy of replica i, T_i denotes the corresponding temperature and k_B is the Boltzmann constant. In order to enable similar exchange rates between each pair of neighbouring replicas, temperatures are drawn from a geometric progression in which the ratio of subsequent temperatures is constant. The most efficient exchange ratio is approximately 20% for atomistic systems [33]. Iterative algorithms have been devised to find the temperature distribution that ensures an exchange ratio close to the optimum. The replica simulated at the target temperature samples the desired phase space distribution, while the replicas at higher temperatures enable to overcome energy barriers more quickly. Due to the coordinate exchanges, the ground state replica can overcome these barriers as well, as it frequently receives coordinates sampled at higher temperatures, while it is not required to sample the high energy transition states inaccessible at the ground state temperature.

While temperature replica exchange speeds up the sampling process at the target temperature, running many parallel replications of the simulated system is expensive. To limit the number of replicas, more efficient replication schemes have been invented. One widely used scheme is the Hamiltonian Replica Exchange MD (HREMD) or Replica Exchange with Solute Tempering (REST) method [34]. Instead of scaling the temperature of the complete system, the system is partitioned into separate regions: a hot region whose sampling is to be increased, for example a protein, and a cold region which is of lesser interest, for example the solvent. Instead of scaling the temperature directly, the energy function, also called Hamiltonian, of the hot region is modified using a scaling factor $\lambda \leq 1$.

- Interactions within the hot region are kept at an effective temperature T/λ.
- Interactions between the hot and cold region are kept at an effective temperature of $T/\sqrt{\lambda}$
- Interactions within the cold region are unaffected

In order to achieve this, the interactions in the hot region and at the interface of both regions need to be scaled:

- Charges affecting the Coulomb potential (equation 2.36) in the hot region by a factor $\sqrt{\lambda}$,
- the Lennard-Jones ϵ parameter (equation 2.35) in the hot region by a factor λ,
- dihedral angles within the hot region by λ and those crossing the region boundary by $\sqrt{\lambda}$.

The acceptance probability function for exchanges between replicas is changed to

$$p = \min\left(1, e^{\frac{E_i(\boldsymbol{r}_i)-E_i(\boldsymbol{r}_j)+E_j(\boldsymbol{r}_j)-E_j(\boldsymbol{r}_i)}{k_BT}}\right), \tag{2.42}$$

where $E_i(\boldsymbol{r}_j)$ is the potential energy of the coordinates of replica j under the Hamiltonian of replica i and all other quantities are as before. This ensures efficient sampling of the hot region in the replicas with $\lambda < 1$, while the cold solvent does not have a large negative impact on the exchange acceptance probability. This is different from temperature REMD, where in addition to the protein, the solvent has to be heated, requiring many replicas in order to obtain sufficient energy overlap between replicas, which is necessary for acceptable exchange rates. HREMD is, for example, available in the MD program GROMACS through the PLUMED plugin [35, 36].

2.2.4 Analysis of MD Simulations

While inspiring ideas about the structural dynamics of proteins can be gained through visual inspection of the molecular dynamics trajectories, statistical analysis methods together with prediction of experimentally available quantities is required for a sound scientific evaluation.

2.2.4.1 RMSD

One way to quantify the similarity between conformations of an equivalent set of atoms is the Root Mean Square Deviation (RMSD) of atomic positions. It is defined as

$$\mathrm{RMSD} = \frac{1}{N}\sqrt{\sum_{i=1}^{N} \delta_i^2}, \tag{2.43}$$

where i enumerates the N atoms of the selection and δ_i is the distance between the atom positions in both conformations. Usually, a rigid superposition that minimizes the RMSD of the two structures is performed prior to the RMSD computation. This is achieved by a translation of the centers of mass of the molecules to the same position and a subsequent optimal rotation that minimizes the RMSD [37, 38].

The structural similarity of protein structures is often assessed by computing the RMSD over their C_α atom positions or over all backbone atoms. The smaller their RMSD, the more similar two structures are. However, the RMSD is only a good measure for similarity for closely similar structures, as structures with larger RMSDs to a target structure cannot be discriminated against each other.

2.2.4.2 RMSF

Through the computation of the Root Mean Square Fluctuation (RMSF) of atom positions, the local flexibility of protein structures can be measured. The RMSF is defined as:

$$\mathrm{RMSF} = \sqrt{\left\langle \left\| \boldsymbol{r}_{i,t} - \boldsymbol{r}_{\mathrm{ref}} \right\|^2 \right\rangle}, \tag{2.44}$$

where $\boldsymbol{r}_{i,t}$ is the position vector of atom i at time t, $\boldsymbol{r}_{\mathrm{ref}}$ is a reference position and angular brackets denote averaging over all frames in the MD trajectory. Common choices for the reference position are an experimentally determined reference structure, starting structures or average structures of the simulation. If the average structure is used as the reference position, the RMSF can be interpreted as the standard deviation of an atom position around its mean position. RMSFs can be computed for all backbone atoms or all heavy atoms and subsequently averaged on a per residue basis to describe the flexibility of the corresponding set of atoms per residue.

2.2.4.3 Dihedral Angle Principal Component Analysis

The RMSD fails as a measure of protein structure similarity (and similarly RMSF as a measure for flexibility), when the differences in the structures become too large, for example, if parts of the protein start to unfold. The protein backbone dihedral angles ϕ and ψ define the backbone protein structure as a set of internal coordinates without the need of a structural alignment. Moreover, using ϕ and ψ instead of Cartesian coordinates allows to better disentangle motions in highly flexible or even unordered parts of the protein from small fluctuations in the rigid regions.

Principal Component Analysis (PCA) is a way to transform a set of observations with linearly correlated variables into a set of coordinates that are linearly uncorrelated, called principal components. This is achieved by diagonalization of the covariance matrix of all observations. The resulting eigenmodes of the covariance matrix represent the principal components and the variance of the data along each principal component is equal to the corresponding eigenvalue. The spectrum of eigenvalues decays quickly for highly correlated datasets, with only a few eigenvalues of significant magnitude. The data can be projected into a space spanned by only a subset of the eigenmodes to ease the interpretation of the input data by concentrating on the significant fluctuations only, while reducing noise in the data along the ignored eigenmodes. By selecting only eigenmodes with eigenvalues larger than a defined threshold, all fluctuations of the input data along the ignored eigenmodes are removed. The eigenmodes themselves give information on correlations and anticorrelations of different variables in the input data.

It is possible to perform a PCA on the dihedral angles of an MD trajectory. As the dihedral angles are circular coordinates with a discontinuity between 0° and 360°, they need to be transformed into a Cartesian space first. This is achieved by introducing new coordinates:

$$v_i = \cos(\phi_i) \tag{2.45}$$

$$w_i = \sin(\phi_i) \tag{2.46}$$

$$x_i = \cos(\psi_i) \tag{2.47}$$

$$y_i = \sin(\psi_i) \tag{2.48}$$

These coordinates are now of Cartesian nature. Distances between points in this space can be interpreted as Cartesian distances between points on a unit circle. These distances approximate the distances of circular coordinates along the arc of the unit circle well, especially for small angles. The PCA is then performed in a space spanned by $4(N_{\text{res}} - 1)$ coordinates, where N_{res} is the number of residues. The first and last residues do not possess *phi* and *psi* angles respectively.

In an MD study of decaalanine, dihedral angle PCA (dPCA) has proven to be "a one-to-one representation of the original angle distribution and that its principal components can readily be characterized by the corresponding conformational changes of the peptide" [39].

2.2.4.4 NMR/Fluorescence Order Parameter Determination

Molecular dynamics trajectories lend themselves to a direct computation of S^2 order parameters (section 2.1.2.1), as the evolution of the N–H bond vectors is known directly from the MD trajectories. The S^2 order parameter is equal to the convergence value of the internal bond vector correlation functions (equation 2.7 in section 2.1.2.1). This convergence value can be estimated in multiple ways. The MD trajectory can be fitted to a reference structure, to remove overall rotation of the molecule. The internal correlation functions can then be computed from the fitted trajectory and the S^2 order parameter estimated as the convergence value, either by taking the final value or as the average over a number of points at the end of the correlation function to reduce the influence of noise.

However, there is an alternative route to this procedure. The order parameter can be computed as an ensemble average over the trajectory as [40]:

$$S^2 = \frac{1}{2}\left(3\sum_{i=1}^{3}\sum_{j=1}^{3}\langle\mu_i\mu_j\rangle^2 - 1\right), \tag{2.49}$$

where μ_1, μ_2 and μ_3 are the x, y and z components of $\hat{\boldsymbol{\mu}}$. This method avoids the computation of bond vector correlation functions and the accompanied problems if the correlation functions of flexible residues do not decay to constant values.

It is also possible to fit one or more (Υ) exponential decays to the internal correlation functions and obtain multiple S_i^2 values, together with their correlation times τ_i:

$$C_I(t) = \sum_{i=1}^{\Upsilon} A_i e^{-\frac{t}{\tau_i}} + A_0, \tag{2.50}$$

where the decay processes are associated with their individual correlation times, τ_i, and amplitudes of the decay process, A_i. The slowest correlation time is defined to be τ_1, and correlation times are sorted in decreasing order. Each of these decay processes then reflects a different dynamical process of the bond vector, giving rise to the decorrelation. Each decay process can therefore be associated with an S_i^2 order parameter, which is the sum of amplitudes of all slower processes: $S_i^2 = \sum_0^{j=i-1} A_j$. The S_i^2 order parameter of the slowest internal process ($i = 1$) is the S^2 order parameter commonly discussed in NMR literature.

For very rigid bonds, the correlation functions decay quickly (within a few picoseconds) to a constant value and the obtained S^2 values have little variance when computed for many uncorrelated subtrajectories of the same simulation. However, flexible bonds tend not to converge quickly and hence the error of the derived S^2 values is large. The simplest possibility to circumvent this problem of unconverged correlation functions is to average the correlation functions from multiple equally long and uncorrelated subtrajectories and hence obtain a single well converged correlation function, from which the S^2 values can be readily computed.

It is also possible to take the exponential fitting procedure one step further and fit the global bond vector correlation functions directly with exponential decays, including the bond vector decorrelation due to overall rotation of the molecule. The global correlation function is assumed as a product of the internal correlation function and the global decorrelation:

$$C(t) = C_I(t) \cdot C_{\text{global}}(t) = C_I(t) \cdot e^{-\frac{t}{\tau_{\text{global}}}} \tag{2.51}$$

The noise in the global correlation functions is too large to allow for meaningful fits of a single correlation function. Only by averaging many correlation functions of a single bond vector over subtrajectories, sufficiently smooth global correlation functions can be obtained.

The optimal number of decay processes that can be fit to a correlation function depends on the underlying number of decay processes with sufficiently well separated decay times. Hence the risk of overfitting the correlation functions with too many exponential decays exists. The Akaike Information Criterion (AIC) can be employed to find the optimal number of decay processes [41]:

$$\mathrm{AIC} = 2K + n \cdot \ln(\mathrm{RSS}), \tag{2.52}$$

where K is the number of parameters used for fitting, n is the number of data points and RSS is the Residual Sum of Squares of the fit. AIC penalizes more complex models and has been shown to result in the optimal choice of model: as simple as possible while maintaining a good quality of fit.

The resulting order parameters need to be scaled to be comparable to experimental data by a factor ζ:

$$\zeta = \left(\frac{1.02}{1.04}\right)^{6} \approx 0.89 \tag{2.53}$$

The reason is, that for the computation of S^2 order parameters during the analysis of NMR data, S^2 is interpreted as a ratio of experimental dipolar coupling strengths to those that would be observed in a totally rigid reference molecule [42]. An internally rigid molecule ignores quantum-mechanical zero point vibrational motions and the effective bond length chosen for data analysis affects the scaling of the S^2 order parameters. For historical reasons, a bond length of 1.02 Å is most often used, while a bond length of 1.04 Å would yield S^2 order parameters much closer to unity. As classical MD simulations ignore quantum mechanical effects and treat N–H bond vectors as rigid, yielding S^2 order parameters near unity independently of the bond length, it is helpful for the comparison of MD and NMR derived order parameters to scale the experimental S^2 values by $1/\zeta$ in order to obtain close to unity values for very rigid N–H bonds as well. However, it is more common to scale the simulation derived S^2 values by ζ, achieving an equivalent effect for comparison. One should keep in mind that it is relative differences in S^2 that matter, and not absolute differences between experiment and simulation.

Side chain mobilities probed in fluorescence spectroscopy can be readily computed from MD data with the same procedure described above for N–H bond vector S^2 order parameters measured in NMR experiments. As most side chains consist of multiple bonds, order parameters for all bonds should be computed and then averaged. Bonds able to rotate without affecting the overall orientation of the side chain should not be considered, such as bonds in the Tyrosine and Phenylalanine rings not colinear with the C_α-C_β bonds during rotations of the ring around the axis of that bond. A good example are the C_β-C_γ bonds of Tyrosine.

The computation of order parameters from MD trajectories has been automated in the software **MOP$\boldsymbol{S^2}$** and is made available to the public by the author [43].

2.2.4.5 Chemical Shift Prediction

The chemical shift of an atomic nucleus is defined as its resonant frequency in a magnetic field relative to a reference frequency (more details on chemical shifts are provided in section 2.1.2.2). Many methods to predict chemical shifts have been developed over the last decades [44]. Bioinformatics based methods predict chemical shifts from protein sequence data alone, utilizing the fact that similar sequences adopt similar protein structures, which in turn exhibit very similar chemical shifts. By comparing a target protein sequence to a large database of sequences with known chemical shifts, it is often possible to generate reliable shift predictions from the known structures. In contrast to sequence based methods, shift prediction methods have been developed that base their prediction on three-dimensional structural data. These methods fall into different categories, based on the level of theory employed for the predictions. Quantum Mechanical (QM) prediction methods employ Hartree-Fock (HF) or Density-Functional Theory (DFT) calculations to find estimates for the chemical shifts in a molecule. These methods are very costly, as they require huge amounts of computing power, especially for larger molecules. Semi-empirical prediction methods are a faster alternative to QM based predictions, at the cost of a lower performance. However, for large scale protein chemical shift predictions, empirical prediction methods are needed, that do not resort to quantum mechanics and are therefore a lot faster to compute. The most successful of these methods employ modern machine learning techniques, either solely or in combination with other empirical methods. Two of the most successful methods (according to references [44, 45, 46, 47, 48]) are described in more detail.

SPARTA+ The chemical shift prediction with SPARTA+ [45] is based on an artificial neural network. The network is trained with a carefully selected set of 580 protein structures, for which high-resolution X-ray structures and complete experimental backbone and C_β chemical shifts are available. The network consists only of a single level (one hidden layer plus input and output layers). Input to the neural network consists of backbone and side chain torsion angles of a residue and its two immediate neighbours, information on H-bonds, the presence of ring currents and electrostatic interactions and a prediction of the local backbone flexibility. In total this amounts to up to 113 parameters per residue, depending on the identity of the amino acid. The prediction of a 100-residue protein takes approximately 1 to 2 seconds.

SHIFTX2 SHIFTX2 divides the chemical shift prediction into two subpredictions, which are finally combined for the end result. The first prediction (SHIFTX+) is based on the protein structure, while the second (SHIFTY+) is based on the amino acid sequence. SHIFTY+ aligns the amino acid sequence of the target protein with a large database and predicts the chemical shift from shifts of known structures with highly similar sequence and hence presumably similar chemical shifts. SHIFTX+ uses a machine learning approach for chemical shift prediction that combines multiple statistical machine learning methods and trains these on a set of 63 parameters. These parameters include the amino acid type and the type of its neighbours, the secondary structure, the hydrogen bonding status, the solvent accessible surface area, backbone and side chain torsion angles and the presence of ring currents and electrostatic fields. Finally, the predictions of SHIFTX+ and SHIFTY+ are combined. However, the sequence based prediction of SHIFTY+ is only considered if the sequence identity with any sequence from the database was at least 40%. If the difference in shift predictions from SHIFTX+ and SHIFTY+ is sufficiently small, the sequence based prediction of SHIFTY+ is used exclusively. Otherwise, a linear combination of the two predictions is used as the final prediction, with a bias towards the structure based prediction of SHIFTX+ that grows with the difference in the two methods.

The motivation behind this procedure is, that for highly conserved proteins, the chemical shifts of the known homologue are very reliable predictions, but their accuracy decreases quickly with dissimilarity in the sequences and the predictive value of the structure based prediction increases.

According to reference [46], SHIFTX2 outperforms SPARTA+ slightly, but both methods are far more accurate in terms of chemical shift RMSDs and correlation coefficients relative to the experimental shifts than all other methods used in the benchmark. As the comparison was done by the developers of SHIFTX2, the possibility of the selection of a favorable test set for SHIFTX2 exists and hence the inferiority of SPARTA+ is questionable. Hence, both methods represent very good choices for accurate and state of the art chemical shift predictors.

2.3 Molecules

2.3.1 CitAP – BsLA Fusion Protein

2.3.1.1 *Bacillus Subtilis* Lipase LipA

Bacillus Subtilis is a gram-positive, aerobic bacterium found in soil and water that is subject to active research since 1992 [49]. *Bacillus Subtilis* is of major industrial interest because of its highly efficient protein secretion system [50], enabling the production of proteases and amylases in bulk quantities [51]. Of special interest are its excreted lipases, as these catalyze the hydrolysis and synthesis of long chain triacylglycerols [52]. These lipases possess a wide substrate specificity. They find applications in transesterification reactions, the synthesis of esters, the resolution of racemic mixtures, and are used as additives in laundry detergents [53].

The reported experiments where performed using lipase LipA from *Bacillus Subtilis* (BSLA). Compared to lipases from other organisms, BSLA is relatively small, consisting of 181 amino acids and weighing 19.36 kDA. BSLA belongs to the α/β hydrolase fold enzymes. Members of this protein family are constituted by a twisted β-strand core flanked by α-helices on both sides [54]. Because of its small size and because other known α/β hydrolase fold enzymes extend on this core structure, it can be considered as a minimal α/β hydrolase. BSLA hydrolyses sn-1 and sn-3 glycerol esters with long fatty acid chains, while its highest activity is exhibited on medium sized fatty acids [52]. The optimum environment is reported to be found at pH 10 [52].

Most known lipases possess an α-helical segment that covers the active site. The encounter of lipid aggregates triggers the opening of this lid, exposing the active site and enabling catalytic activity. This behaviour is known as interfacial activation [54]. BSLA lacks this lid, resulting in an active site that is solvent exposed and thus actively hydrolysing fatty acids even in the absence of lipid aggregates.

The active site is located at the bottom of a shallow cleft formed by four turns (Figure 2.8). Its structure is in accordance with the secondary structure topology of the canonical α/β hydrolase fold, in which the active site is formed

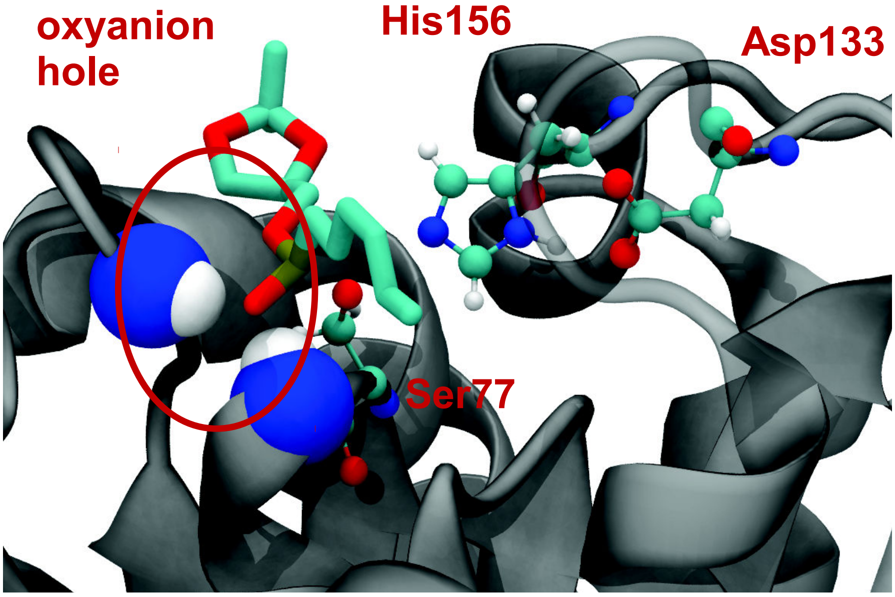


Figure 2.8: Active site of *Bacillus Subtilis* lipase LipA with 1,2-O-isopropylidene-sn-glycerol (IPG) phosphonate-inhibitor bound. The catalytic triad residues are colored, as well as the backbone nitrogen atoms forming the oxyanion hole. Coordinates taken from PDB accession code 1R50.

by a nucleophile, an acid and a histidine, together called the catalytic triad [52]. In BSLA this catalytic triad is constituted of serine 77 (nucleophile), aspartate 133 (acid) and histidine 156. Serine 77 is located at the so called nucleophilic elbow [55], a sharp turn at the bottom of the binding pocket connecting a central β-sheet to a peripheral α-helix. The ester hydrolysis reaction, catalyzed by BSLA is depicted in Figure 2.9. The ligand binds covalently with its phosphoryl group to Serine 77. A tetrahedral intermediate structure is formed. The phosphoryl oxygen of the substrate is hydrogen bound to the NH atoms of methionine 78 and isoleucine 12, forming the oxyanion hole, a pocket stabilizing the negatively charged transition state common to most lipases [52]. The N^ϵ atom of His156 is located at hydrogen bonding distance from the Ser77 O^γ atom and the substrate ester oxygen.

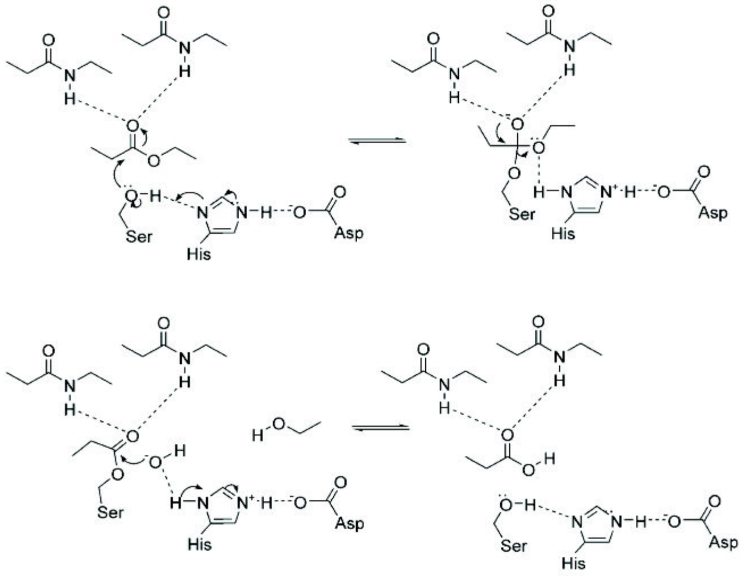


Figure 2.9: Ester hydrolysis reaction catalyzed by BSLA. Figure reused with permission from reference [56].

2.3.1.2 Periplasmic Domain of Sensor Histidine Kinase CitA

CitA is the sensor histidine kinase of a two component regulatory (TCR) system of *Klebsiella Pneumoniae* [57]. TCR systems generally consist of a sensor histidine kinase and a response regulator (CitB in the present case). Sensor histidine kinases respond to environmental conditions by autophosphorylation of a conserved histidine residue. The phosphoryl group is subsequently transferred to a conserved aspartate residue of the response regulator. The phosphorylated response regulator in turn triggers a change in gene expression or cell behaviour [57]. The CitA/CitB regulatory system controls the expression of citrate fermentation genes. Citrate metabolism requires careful regulation, as an uncontrolled transcription of citrate fermentation genes in the absence of citrate would break the citric acid cycle with severe consequences for the organism [57].

Like most known sensor histidine kinases, CitA is a membrane protein, the domain composition of which is shown in Figure 2.10. Its extracellular N-terminal sensor domain is flanked by two transmembrane helices, one of which links to the intracellular autokinase domain. Like virtually all known sensor histidine kinases, CitA appears to be dimeric [58]. In many cases only the structures of the

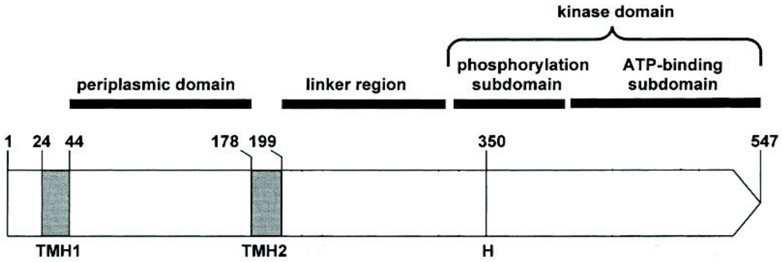


Figure 2.10: CitA domain organization. The transmembrane helices are indicated with TMH1 and TMH2. Only the structure of the periplasmic domain (CitAP) is known and was used in the reported research. Histidine 350 gets phosphorylated after signal transduction as described in the text. Figure reused with permission from [57].

extracellular domains of sensor kinases are known, due to the inherent difficulties of obtaining structural information about membrane proteins. This is also true for CitA, where only structures of the periplasmic domain (CitAP) have been resolved. CitAP consists of 135 amino acid residues, weighing a total of 14.68 kDA.

The overall fold of the periplasmic domain of sensor histidine kinases classifies them as members of the Per-Arnt-Sim domain superfamily. This superfamily comprises many well studied sensor histidine kinases [59].

The mechanism of action of CitA [57] is initiated by citrate binding, triggering a conformational change of the periplasmic domain that is mediated by a transmembrane helix (TMH) to the cytoplasmic kinase domain. It has been hypothesized that this mediation happens by a piston-like mechanism (Figure 2.11) in which transmembrane helix 2 is pulled to the periplasmic side of the membrane. The resulting conformational change in the kinase domain allows the interaction of the ATP-binding subdomain with the cytoplasmic phosphorylation subdomain. After autophosphorylation the subdomains dissociate. The phosphorylation subdomain can now bind to the receiver domain of the response regulator CitB. A final phosphotransfer in the phosphorylation domain from histidine to an aspartate completes the CitA cycle and a new response can be triggered.

The binding pocket of CitAP includes a β-sheet core that forms the bottom and two loops, the minor loop (residues 96 to 106) and the major loop (residues 63 to 92, Figure 2.12). These loops form tight lids covering the bound ligand, forming

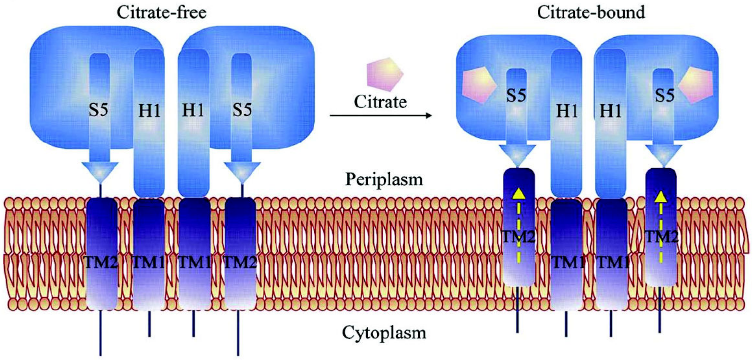


Figure 2.11: Hypothetical mode of action of CitA. The citrate-triggered conformational change of the periplasmic domain is mediated to the cytoplasmic kinase domain by a piston-like mechanism. Figure reused with permission from [60]

a closed pocket. While the conformation of the termini remains uncertain due to crystal packing artifacts, the core structure of the bound CitAP has been solved with 1.6 Å resolution and has been experimentally found to be well folded in solution (by NMR spectroscopy, [60]). The major loop is unresolved in the unbound CitAP structures, indicating that it is highly mobile without the stabilizing effect of the ligand. The core structure of the protein is highly similar for the bound and unbound structures with similar backbone positions and side chain orientations of the binding residues not located in the minor or major loop [60].

A sodium ion attached to the first helix of the periplasmic domain (H1 in Figure 2.11) and the β-sheet core has been crystallized in the citrate bound structures. This ion seems to be of physiological relevance, as sodium is required for the induction of the CitA/CitB target genes [61] and is backed by the fact that no sodium ion has been found in the citrate free crystal structure.

It has been reported that citrate binding to CitAP is highly specific and that virtually all CitAP molecules are in the bound state under physiological citrate concentrations [60]. Citrate interacts mainly with Thr58, Arg66, His69, Glu80, Ser101, Leu102, Arg107, Lys109 and Ser124 by hydrogen bonding (Figure 2.13). A further hydrogen bond is mediated by a water molecule to Gly103.

The conformational change in the periplasmic domain of sensor histidine kinase CitA is illustrated in Figure 2.12. The main conformational change is a flattening of the β-sheet that forms the minor loop. Upon citrate binding this β-sheet is

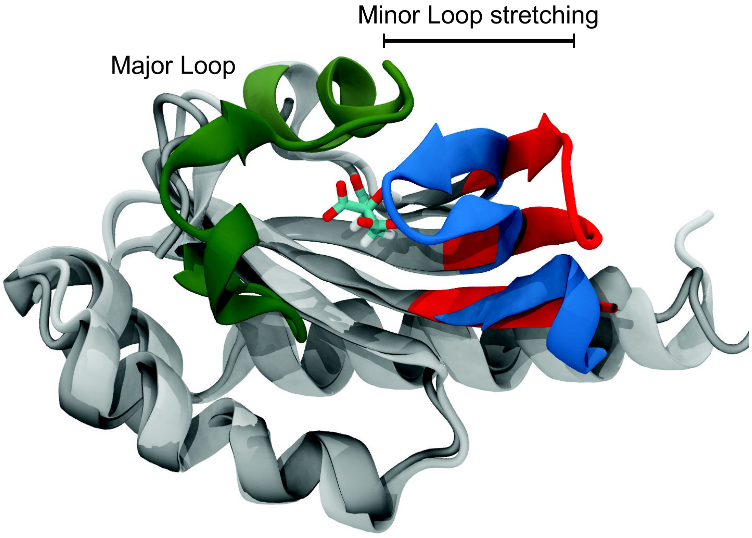


Figure 2.12: The conformational change in the periplasmic domain of sensor histidine kinase CitA. The minor loop changes from the bound conformation (blue) to the open conformation (red). Citrate bound and free coordinates have been taken from PDB structures 2J80 and 2R50, respectively. Coordinates for the major loop (green) are only available in the bound structure as it was indicated to be unstructured in the ligand free form, at least under crystallization conditions.

pulled in, due to interaction with the ligand. This disrupts some of the interactions with neighbouring β-sheets and subsequently pulls the C-terminus away from the membrane. The unstructured major loop in the citrate free crystal structure indicates that the citrate binding pocket is less stable in absence of citrate.

2.3.1.3 Fusion Complex

The two proteins, CitAP and BSLA, have been fused in the laboratory, motivated by the desire to regulate the lipolytic activity of BSLA via citrate binding to CitAP. A controlled upregulation by a cheap and non-toxic molecule such as citrate could be utilized to switch on the desired catalysis at specified times. It was assumed that the conformational change of CitAP triggered by citrate binding could have a regulatory effect on lipase activity.

The sequence of the fused construct, as used in the experiments performed in the group of Karl-Erich Jaeger (Heinrich-Heine-University Düsseldorf, Germany),

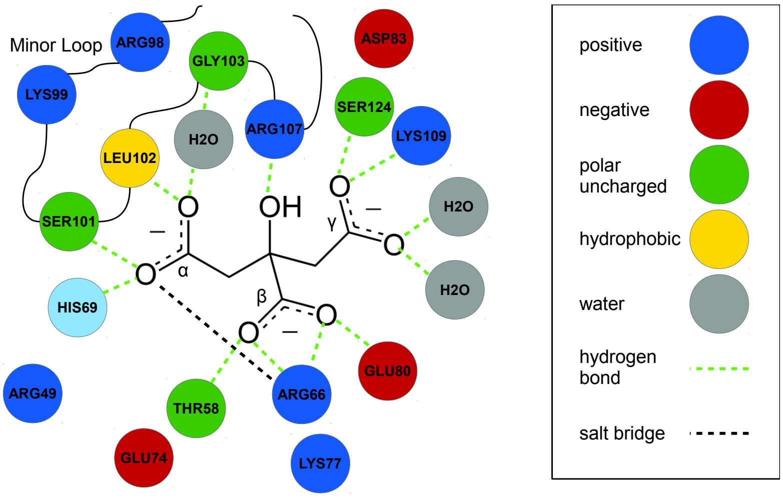


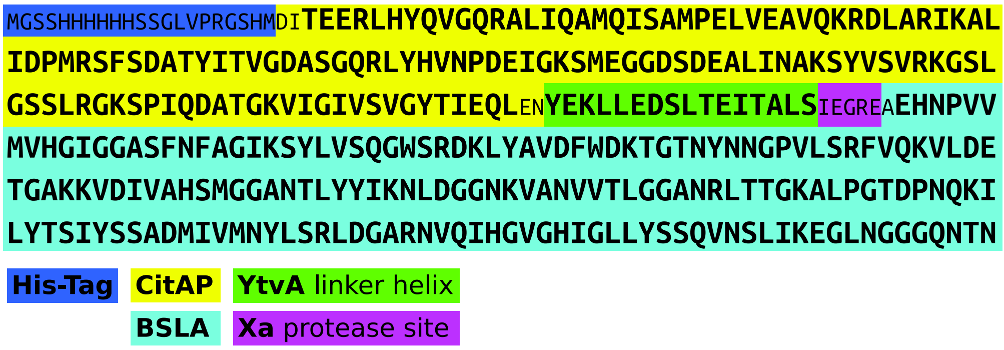
Figure 2.13: Interactions of CitAP with its ligand citrate as determined by visual inspection of a 1 ns molecular dynamics trajectory of citrate-bound CitAP. The minor loop is the domain with the largest RMSD in the transition from the citrate bound to the citrate free conformation.

is shown in Figure 2.14. CitAP was connected to BSLA by means of a linker helix that was taken from YtvA, the blue light sensor of *Bacillus Subtilis* (PDB accession code: 2PR5). A polyhistidine-tag (His-tag) was attached to the N-terminus of CitAP to allow easy purification and a protease site was included in the complexed protein as it was found experimentally that this increased the effect of citrate on BSLA activity. The construct consists of 336 amino acid residues and weighs 36.21 kDA.

Although it was attempted to up-regulate the lipase activity by citrate binding, experiments indicated the opposite effect (section 3.1). It was further demonstrated that increasing citrate concentrations had no effect on the activity of isolated BSLA. Lipolysis was significantly reduced at higher citrate concentrations in the CitAP–BSLA complex to about 40% of the initial activity of the complex.

2.3.2 GABARAP and GATE-16

Throughout a eukaryotic cell's life cycle, many of its components become dysfunctional or unnecessary [62, 63]. To prevent damage due to toxic protein assemblies and to reuse building blocks such as lipids and amino acids, these components are



```
MGSSHHHHHHSSGLVPRGSHMDITEERLHYQVGQRALIQAMQISAMPELVEAVQKRDLARIKAL
IDPMRSFSDATYITVGDASGQRLYHVNPDEIGKSMEGGDSDEALINAKSYVSVRKGSL
GSSLRGKSPIQDATGKVIGIVSVGYTIEQLENYEKLLEDSLTEITALSIEGREAEHNPVV
MVHGIGGASFNFAGIKSYLVSQGWSRDKLYAVDFWDKTGTNYNNGPVLSRFVQKVLDE
TGAKKVDIVAHSMGGANTLYYIKNLDGGNKVANVVTLGGANRLTTGKALPGTDPNQKI
LYTSIYSSADMIVMNYLSRLDGARNVQIHGVGHIGLLYSSQVNSLIKEGLNGGGQNTN
```

Figure 2.14: Amino acid sequence of the CitAP–BSLA complex.

disposed and recycled in a controlled manner through a process called autophagy [64]. Autophagy involves the creation of an autophagosome: a spherical vesicle formed by a double layer membrane that engulfs the cellular waste [65]. Eventually, the autophagosome fuses with spherical, membrane bound vesicles called lysosomes [66]. In the lysosomes slightly basic interior, biopolymers are degraded into their components by hydrolytic enzymes.

The biogenesis of the autophagosome consists of multiple stages [67]. First, a cup-shaped membrane forms at the endoplasmic reticulum, called the phagophore. In a subsequent step, the phagophore enwarps a volume of the cytoplasm and forms the autophagosome vesicle. Finally, the autophagosome fuses with the lysosome, in which the breakdown of the cargo takes place.

Many proteins are involved in the creation and constitution of autophagosomes and lysosomes and in the signalling process that targets proteins for degradation. The proteins involved in autophagosome formation are collectively denoted as autophagy-related proteins (ATG). In yeast, the autophagic process is dependent on the protein ATG8, of which several orthologs exist in higher eukaryotes. These orthologs are collectively referred to as the LC3/GABARAP family, which in turn is constituted by two subfamilies: LC3 and GABARAP. The LC3 subfamily consists of four distinct proteins (LC3A/B/C, where LC3A exists in two different splicing variants), while the GABARAP subfamily consists of GABARAP, GABARAPL1 and GABARAPL2 (GABARAP like proteins 1 and 2), where GABARAPL2 is also known by its alternative name GATE-16 (Golgi-associated ATPase Enhancer of 16 kDa [68]). The acronym GABARAP, referring to γ-aminobutyric acid receptor type A ($\mathrm{GABA_A}$) receptor associated protein, stems from the function it was first associated with during its discovery [69]. The subfamily classification is based on their degree of homology, characterized by intrafamily sequence identities larger than 80% and interfamily sequence identities of approximately 60% [70]. The structure of all LC3/GABARAP proteins is highly conserved. They consist of two N-terminal α-helices, followed by a ubiquitin like core [71]. A key feature responsible for their ability to selectively associate with specific binding partners is the difference in electrostatic properties of these N-terminal helices ($\alpha_1 - \alpha_2$). Helix 1 of LC3 is basic, while it is acidic in GATE-16 and GABARAP. Helix 2, on the other hand, has an acidic, neutral and basic surface in LC3, GATE-16 and GABARAP, respectively. This gives rise to their specific electrostatic patterns: LC3 (basic-acidic), GATE-16 (acidic-neutral), GABARAP (acidic-basic). Even though the LC3/GABARAP family members are structurally very similar, they exhibit non-redundant, unique functions during autophagy.

In order for LC3/GABARAP proteins to fulfill their functions, post-translational modifications are necessary. The proteins exist in an unmodified pro-form, which is then cleaved at the C-terminus by members of the ATG4 protease family to expose a conserved glycine residue, yielding isoform I. Subsequently, isoform II is produced by other ATG proteins covalently conjugate phosphatidylethanolamine (PE), a class of phospholipids found in biological membranes, to the conserved glycine [72]. The conjugated PE is used to anchor the proteins to membranes at which they fulfill their functions.

The different known functions of LC3/GABARAP proteins are summarized in Figure 2.15. GABARAP can act as scaffolding proteins, recruiting other binding partners and forming complexes with them, enabling the nucleation of the autophagosomal membrane (Figure 2.15A) [73]. Whether the initial membrane nucleation happens due to GABARAP and the membrane is formed *de novo* or if the membrane is derived from other organelles, inheriting a preformed lipid assembly, and held in place by the protein scaffold is still unknown. Selective engulfment of organelles is possible with the aid of GABARAPL1 and LC3B (Figure 2.15B). Elongation of the phagophore membrane is facilitated predominantly by LC3B, mainly due to ionic interactions with the lipid heads (Figure 2.15C) [67]. Closure of the phagophore to produce the autophagosome is mediated by LC3B, but more importantly by GABARAP (Figure 2.15D). It has been hypothesized, that GABARAP exists in an open conformation, in which the two N-terminal helices detach from the ubiquitin core of the protein and penetrate the adjacent membrane by hydrophobic interactions of the unfolded helix 1, enabling the phagophore closure [74]. In autophagosome fusion with the lysosome, several members of the LC3/GABARAP family play important roles in the interactions with other proteins relevant for the fusion process. However, the exact interaction partners and processes at work are yet unclear (Figure 2.15E).

In addition to the importance for autophagy, LC3/GABARAP proteins perform functions unrelated to autophagy. GABARAP, as its name suggests, was originally found to facilitate intracellular transport due to interactions with the GABA_A receptor. The effect of this interactions might be the incorporation of this receptor into the transport vesicle membrane. GATE-16 enhances the transport across membranes of the Golgi apparatus (Figure 2.15F) [75]. Other known involvements are the degradation or proteins in the endoplasmic reticulum, phagocytosis, mobility and growth, as well as functions as a tumor suppressor.

More information can be found in a detailed review of LC3/GABARAP family proteins, published by Schaaf et al. in 2016 [76], providing an ideal starting point for any reader wishing to familiarize with the current state of LC3/GABARAP research.

2.3.3 Interleukin-6

Interleukin-6 (IL-6) is a pleiotropic cytokine, i.e., a signaling molecule affecting multiple cell types. Its main function is to control the immune response, but many other functions outside of the immune system are known as well, such as metabolic control and the generation of pain. While predominantly acting in a pro-inflammatory way, stimulating the acute phase response to injury [77], IL-6 has also been associated with anti-inflammatory functions, helping cells to recover from inflammation [78].

IL-6 levels are elevated during inflammation and hence it constitutes an interesting target for medical treatments. It could be therapeutically exploited to treat

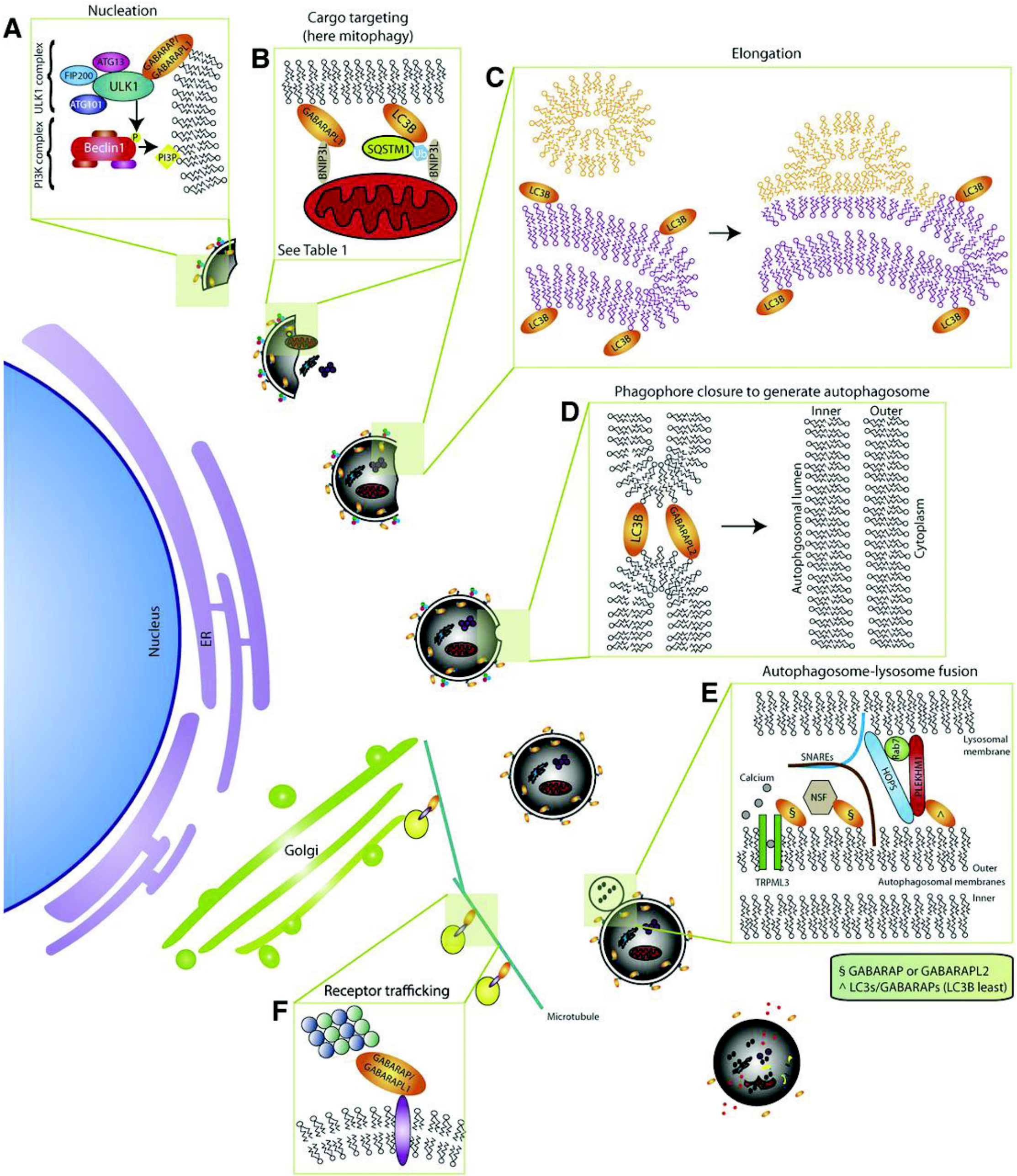


Figure 2.15: Graphical depiction of the different functions of LC3/GABARAP family proteins. Figure reused with permission from reference [76]

chronic inflammation diseases.

IL-6 signalling acts through the gp130 receptor, which it shares with several other interleukins [79, 80]. Gp130 is a transmembrane glycoprotein, expressed ubiquitously in all cell types. It consists of eight domains: six extracellular, one transmembrane and one cytosolic domain. The first domain (D1) is immunoglobulin-like and the following two domains are the main cytokine binding domains (D2 and D3). Preformed but inactive dimers of gp130 exist on the cellular membrane [81].

IL-6 cannot bind gp130 directly. It first needs to form a heterodimer with its α receptor (IL-6Rα) to form the IL-6/IL-6Rα complex. There exists evidence that IL-6Rα exists as a preformed homodimer as well, similar to gp130 [82].

IL-6Rα exists in a membrane bound form (mbIL-6Rα) and in a soluble form (sIL-6Rα), which is produced from the membrane bound form by proteolytic cleavage or through an alternatively spliced IL-6Rα messenger RNA, which does not contain the transmembrane domain [83]. IL-6 alone can only affect cells that express both mbIL-6Rα and gp130 receptors, which is called classic-signaling [84]. However, IL-6Rα is only expressed in a few cell types, such as macrophages and neutrophils, and hence the classic-signaling of IL-6 is responsible for the pro-inflammatory behaviour of IL-6. In addition, IL-6 can bind to sIL6-R and either initiate a response in cells not expressing mbIL-6Rα or amplify signalling on cells that do contain mbIL-6Rα. This process is called trans-signaling. One of its functions is to attract lymphocytes to the inflamed area [85]. The pro-inflammatory functions of IL-6 are likely due to its classical-signaling behaviour, while its anti-inflammatory behaviour is mediated by trans-signaling [86].

IL-6 consists of a four-helix bundle, characteristic for IL-6 type cytokines [87, 88]. The four helices (A, B, C, D) are arranged in an up-up-down-down fashion with short loops connecting helices B and C (BC-loop) and longer loops connecting helix A to B (AB-loop) and C to D (CD-loop), as shown in Figure 2.16.

A crystal structure of a hexameric complex, consisting of two IL-6/IL-6Rα/gp130 trimers, has been solved by X-ray crystallography [89]. However, an alternative model of the receptor association has been proposed, stating that only one IL-6/IL-6Rα complex is needed to bind to and activate the gp130 dimer. [90]. This alternative model suggests that at low IL-6/IL-6Rα concentrations,

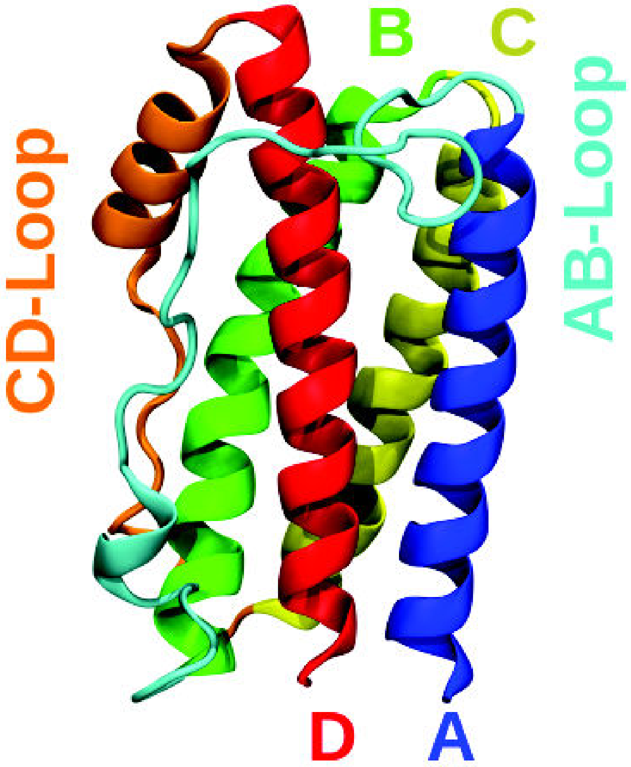


Figure 2.16: The four-helix bundle of IL-6, showing helices A to D in blue, green, yellow and red, respectively, as well as the AB-loop in cyan and the CD-loop in orange.

the complex with a single IL-6/IL-6Rα activates the signal transduction pathway, while at higher concentrations the hexamer forms, in which gp130 is inactive. The hexameric form would in this model present a self-inhibition mechanism to damp the response to IL-6 signaling.

IL-6/IL-6Rα, in its soluble and membrane bound forms, initiates signal transduction of gp130 to the cytosol through associated kinases [79]. IL-6 possesses three distinct receptor binding sites (Figure 2.17), binding domains D2 and D3 of IL-6Rα at site I, domains D2 and D3 of gp130 at site IIa and the immunoglobulin-like domain D1 of the second gp130 in the complex at site IIIa. In the hexameric complex, there are additional interactions between both D3 domains of IL-6Rα and gp130 (site IIb), as well as the D2 domain of IL-6Rα and the D1 domain of gp130 (site IIIb). Site III is characterized by a larger hydrophobic binding interface to which Trp158 makes the largest contribution. Trp158 is conserved among nine species of the IL-6 family. In addition, the AB-loop makes several contacts with gp130's D1 domain. It is highly mobile in the apo state of IL-6, as is evident from 32 structures of an ensemble originating from NMR spectroscopy [87].

A viral variant of IL-6 (vIL-6) exists in human herpes virus 8, which signals through the gp130 receptor without the need for IL-6Rα [91]. Human and viral IL-6 share a sequence identity of 25% and a sequence similarity of 66%. The X-ray crystal structure of the vIL-6/gp130 tetrameric complex shows a very similar binding pose with two gp130 molecules compared to human IL-6 [92]. The binding interfaces IIa and IIIa of viral IL-6 are at equivalent positions to human IL-6, however, the amino acids constituting these interfaces are not well conserved. To identify the crucial binding interfaces that enable the IL-6Rα independent binding of vIL-6, a chimeric variant was constructed in which three subsites of vIL-6 binding site III were introduced into human IL-6 (Figure 2.18) [93]. Interestingly, human IL-6 was only able to bind gp130 in the absence of IL-6Rα if subsites a, b and c were introduced in the chimeric variant at the same time, even though subsite c does not make contact with the D1 domain of gp130 in any of the receptor bound crystal structures. Subsite c is located at the BC-loop and does not directly contact the D1 domain of gp130. Its contribution to binding site III is therefore likely through interactions with the other two subsites, a and b, presumably through salt bridges.

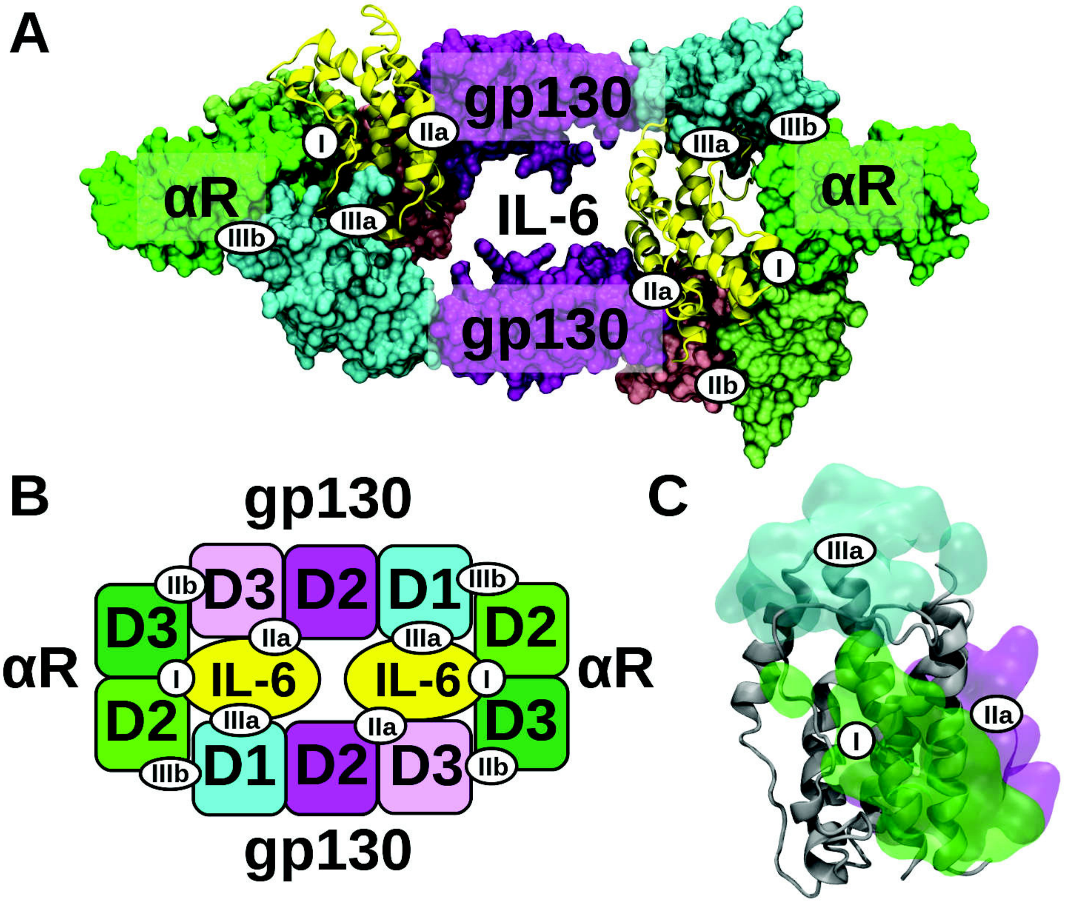


Figure 2.17: **A:** The hexameric human IL-6 assembly (PDB accession code 1P9M). IL-6 is shown in yellow, the α receptor in shades of green, and the gp130 receptor in cyan, purple and rose to indicate the individual subdomains. The three binding sites of IL-6 are indicated with circles. **B:** Schematic representation of the hexameric assembly corresponding to panel A. **C:** The binding epitopes of IL-6 are shown with transparent surface representation with the same color coding as in panels A and B.

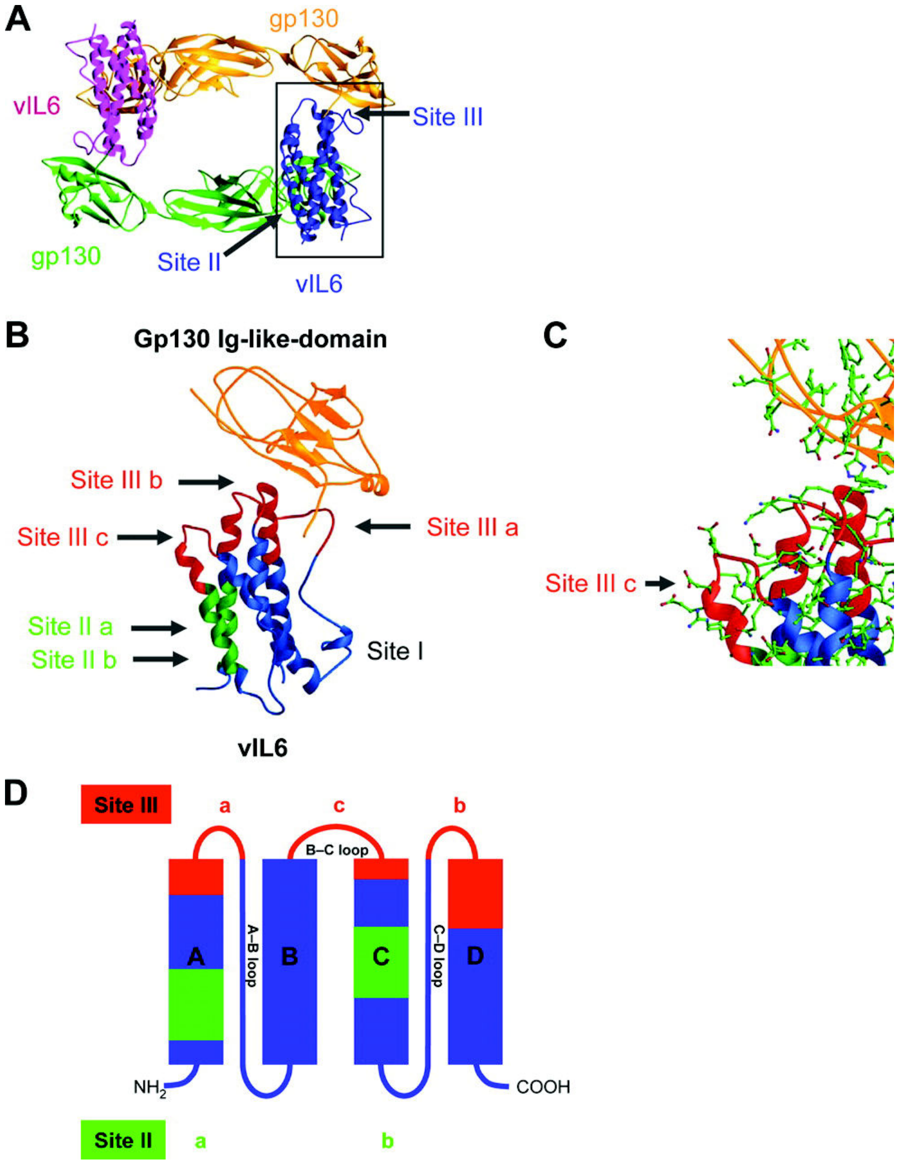


Figure 2.18: Binding of vIL-6 to gp130. (**A**) Ribbon model of the recently solved vIL-6/gp130 X-ray structure. (**B**) Interaction of gp130 (only immunoglobulin-like domain D1) and vIL-6. Sites I, II and III are indicated. (**C**) The BC loop of site III (subsite c) is not in direct contact to gp130-D1. (**D**) Schematic drawing of the common four-helix bundle cytokine fold. The different parts of site II and site III are color coded as green (site II) or red (site III). Figure and caption reused with permission from reference [93].

A detailed review of IL-6, discussing the receptor complex assembly, the different signaling mechanisms and its biological activity has been published by Scheller et al. in 2011 [94]. A more recent review published by Hunter et al. in 2015 focuses predominantly on the biological mechanisms and its experimental evidence [95].

Chapter 3

Applications

3.1 A combination of mutational and computational scanning guides the design of an artificial ligand-binding controlled lipase

3.1.1 Summary

Bacillus Subtilis Lipase A (BsLA, section 2.3.1.1) is set under ligand-dependent regulatory control through the introduction of a regulatory domain (CitAP, section 2.3.1.2) from a different organism. Structures of both individual molecules are known. The proteins were fused together with a helical linker to produce a fusion protein of unknown structure in the hope that CitAP is able to regulate the activity of the lipase. It was found that citrate, the natural ligand of the regulatory domain, is able to inhibit lipase activity. Furthermore, the fusion protein was shown to exist primarily as a homodimer at physiologically relevant concentrations.

The research questions addressed in this publication are: What is the structure of the fusion protein and how is BsLA activity controlled by CitAP?

The most important experimental data to elucidate the structure was obtained by small angle X-ray scattering (section 2.1.3). With the aid of molecular modelling, using MD simulations and SAXS curve reconstruction, we were able to generate a structural model of the fusion protein based on the sub-domains of known structure that is in agreement with the experimental SAXS data. However, it was not possible to fully explain the underlying mechanism of allosteric BsLA regulation based on this structural model.

3.1.2 Contribution

I built the fusion protein models from low resolution SAXS data by rigid body fitting of known X-ray structures. The linker domains were reconstructed from C_α traces, which were obtained by a fit of the SAXS data with single-bead protein residues (section 2.1.3). Four different models of fusion-protein homodimers were generated. Each model was subject to 100 ns MD simulation, after which SAXS scattering curves were computed for the MD trajectories. Finally, the best models were selected, based on χ^2 values, which estimate the agreement between simulated and experimental data.

I wrote manuscript sections on MD simulation methods and model building, the results section on computation modelling and SAXS envelope reconstruction and produced Figures 5a, 5b, 6a and 6b (in total approximately 15 % of the complete manuscript).

3.1.3 Reprint

This section contains a complete reprint of the publication [96]. The supporting information to this article is located in section 6.1.

The publication is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/

www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

Received: 30 September 2016
Accepted: 11 January 2017
Published: 20 February 2017

A combination of mutational and computational scanning guides the design of an artificial ligand-binding controlled lipase

Marco Kaschner[1], Oliver Schillinger[2], Timo Fettweiss[1], Christina Nutschel[1], Frank Krause[3], Alexander Fulton[1], Birgit Strodel[2], Andreas Stadler[4], Karl-Erich Jaeger[1,5] & Ulrich Krauss[1]

Allostery, i.e. the control of enzyme activity by a small molecule at a location distant from the enzyme's active site, represents a mechanism essential for sustaining life. The rational design of allostery is a non-trivial task but can be achieved by fusion of a sensory domain, which responds to environmental stimuli with a change in its structure. Hereby, the site of domain fusion is difficult to predict. We here explore the possibility to rationally engineer allostery into the naturally not allosterically regulated *Bacillus subtilis* lipase A, by fusion of the citrate-binding sensor-domain of the CitA sensory-kinase of *Klebsiella pneumoniae*. The site of domain fusion was rationally determined based on whole-protein site-saturation mutagenesis data, complemented by computational evolutionary-coupling analyses. Functional assays, combined with biochemical and biophysical studies suggest a mechanism for control, similar but distinct to the one of the parent CitA protein, with citrate acting as an indirect modulator of Triton-X100 inhibition of the fusion protein. Our study demonstrates that the introduction of ligand-dependent regulatory control by domain fusion is surprisingly facile, suggesting that the catalytic mechanism of some enzymes may be evolutionary optimized in a way that it can easily be perturbed by small conformational changes.

Allosteric regulation represents a general mechanism which is used throughout all kingdoms of life to achieve control of protein activity. In terms of their evolution it appears reasonable to assume that allosteric proteins evolved from non-allosteric ones. Hereby, the evolution of multidomain (sensory) proteins is of particular interest for engineering purposes mimicking natural evolution, as they potentially arose through establishing domain interactions between independently functioning, ancestral proteins[1–3]. Thus, a key to understanding allostery in multidomain sensory proteins is to understand how those proteins gain, lose and rearrange domains. In theory, new functionalities can emerge by at least two mechanisms: i) the interchange of sensor and effector domains between different sensory receptors in a process called domain shuffling[4] and ii) the recruitment of a sensor domain to an existing non-allosteric protein module[3]. One of the most widespread and versatile sensor domain families, e.g. present in sensory histidine kinases (SHKs)[5] and other multidomain sensory receptors are Per-Arnt-Sim (PAS) domains[6]. Signal perception by PAS domains is usually determined by covalently or non-covalently bound small molecule ligands[7]. Structurally, PAS domains possess a mixed α/β-fold, where usually five anti-parallel β-strands together with a variable set of α-helices form a tight pocket in which the respective ligand is bound[7]. Known ligands include heme, flavins (flavin mononucleotide, FMN and flavin adenine dinucleotide, FAD), 4-hydroxycinnamic acid (4-HCA), divalent metal cations, C3-C4 carboxylic acids (malonate, malate, succinate), C6 carboxylic acids (citrate)[7]. The environmental stimuli that a given PAS domain can recognize, are equally diverse, ranging from chemical signals such as metabolite concentration (e.g. carboxylic acids)[8–10], oxygen

[1]Institut für Molekulare Enzymtechnologie, Heinrich-Heine Universität Düsseldorf, Forschungszentrum Jülich GmbH, D-52425 Jülich, Germany. [2]Institute of Complex Systems ICS-6: Structural Biochemistry, Forschungszentrum Jülich GmbH, D-52425 Jülich, Germany. [3]Nanolytics, Gesellschaft für Kolloidanalytik GmbH, Am Mühlenberg 11, 14476 Potsdam, Germany. [4]Jülich Centre for Neutron Science JCNS and Institute for Complex Systems ICS, Forschungszentrum Jülich GmbH, D-52425 Jülich, Germany. [5]Institute of Bio- and Geosciences IBG-1: Biotechnology, Forschungszentrum Jülich GmbH, D-52425 Jülich, Germany. Correspondence and requests for materials should be addressed to U.K. (email: u.krauss@fz-juelich.de)

(heme)[11,12], redox potential (FAD)[13,14] to physical signals such as light (FAD, FMN and 4-HCA)[15–17]. Based on this diversity and the modular nature of PAS-domain containing sensory receptors, efforts have been made recently to engineer allosteric behaviour into naturally non-allosteric proteins by fusion of PAS sensory domains[18–20]. Although successful in several cases[18–25], the rational engineering of allostery into an existing non-allosteric protein still represents a challenging endeavour. Several strategies have been brought forward all relying on the above described two evolutionary mechanisms, i.e. domain swapping to reprogram allosteric control altering the sensory input of the system[24,25], insertion[18,23] or terminal fusion[19,20] of a sensory domain. Often, the screening of several fusion constructs[20,22–24] and/or circular permutation and several rounds of directed evolution[26] were necessary to obtain an efficient switch. Thus, the general question arises, which of the already explored strategies represents the best for a given target protein, and, more importantly, is it possible to rationally predict the best strategy (N-terminal fusion, C-terminal fusion or insertion) based on i.e. bioinformatics analyses or already available mutational data for a given target protein? To this end, several bioinformatic methods have been developed that infer the evolutionary (statistical) coupling between residue pairs in a given protein family sequence alignment[27,28]. It is reasoned, that this co-evolutionary information captures the statistical signature of functional constraints arising from conserved communication between positions and thus enable the identification of chains of residues facilitating the flow of information necessary for allosteric communication[18,27–29].

In the present contribution, we explore the possibility of rationally engineering allosteric control into the naturally not allosterically regulated lipase A of the Gram-positive bacterium *Bacillus subtilis* (BsLA)[30] by fusion of the citrate-binding CitAP PAS domain of the periplasmic CitA citrate-sensor of *Klebsiella pneumonia*[9], hereby mimicking evolutionary processes that could lead to the emergence of new multidomain proteins. The site of domain fusion (N-terminal, C-terminal or insertion) was rationally determined based on a whole-protein site-saturation mutagenesis dataset of BsLA, backed by computational evolutionary coupling analysis. Functional assays, complemented by a set of biochemical and biophysical studies, suggest a mechanism for control of the artificial citrate-binding lipase, similar but distinct to the one suggested to be realized in the parent sensor-domain containing CitA SHK. Our study demonstrates that the generation of ligand-binding dependent control of an enzyme by sensory domain fusion can easily be achieved in a simple "plug and play" manner.

Results

Computational predictions and site-saturation scanning mutagenesis data identify a network of functionally and evolutionary coupled residues at the N-terminus of BsLA. BsLA is a monomeric α/β-hydrolase that hydrolyses glycerol-esters with medium chain length (C8) as well as *sn-1* and *sn-3* glycerol esters with long fatty acid chains to the corresponding alcohols[30]. It is one of the smallest known lipases, that, in contrast to other lipases, lacks a lid-domain structure and hence does not show interfacial activation[30]. No allosteric effects have so far been described for BsLA. In order to infer chains of evolutionary coupled residues and hence to identify the best site for sensor domain fusion, we computationally inferred the evolutionary coupling between residues in BsLA by using the EVcoupling webserver (www.evfold.org)[31,32]. In order to obtain reliable evolutionary constraints (EC) values, we constructed a large hydrolase core alignment with the BsLA sequence as query for alignment generation using the tools available as part of the EVcoupling webserver. In an unrestrained run, an alignment containing 149.524 sequences was generated (E-value cutoff 10E-3) which was subsequently used to infer EC scores for every residue in the conserved BsLA core. The resulting EC values were mapped onto the BsLA X-ray structure (Fig. 1a, see also Supplementary Figure 1). Evolutionary coupled residues are color-coded from grey (low EC values) to red (high EC values). A network of evolutionary coupled residues appears to be centred around the anti-parallel β-scaffold of BsLA, with the highest values obtained for residues on β3, β5, β6. To experimentally validate those findings we used a set of data obtained by complete site-saturation mutagenesis of BsLA[33,34] and parsed this data for residues whose substitution led to severe loss of function. From this data, the number of inactive variants per residue was determined (Supplementary Figure 2) and the respective values were mapped on the X-ray structure of BsLA (Fig. 1b). Interestingly, very similar to the data obtained from evolutionary-coupling analyses, most "mutationally-sensitive" residues, i.e. those where mutations led in many cases to loss of enzyme activity, are found within the β-scaffold of BsLA, namely on strands β3, β5, β6. In particular, the first N-terminal 11 amino acids including the β3 strand (residues 6 to 9) appear especially sensitive to mutation. Importantly, a similar network of functionally important residues seems to be absent at the C-terminus or within loop regions of BsLA.

Design of the fusion protein. Based on the above described analyses, a potential allosteric communication pathway was predicted extending from the BsLA N-terminus *via* the first β-strand to the enzyme active site (Fig. 1). Thus, in order to gain control over BsLA function we fused the citrate-binding PAS domain CitAP of the CitA SHK of *Klebsiella pneumoniae*[9] N-terminally to BsLA as a putative "effector" module. Hereby, the CitAP PAS domain (residues 44 to 178 of full-length CitA) and full-length BsLA were linked *via* the Jα linker (residues 126 to 147) of the *B. subtilis* YtvA photoreceptor[35], resulting in a tripartite fusion protein (Fig. 2a). In wild-type CitA, a transmembrane helix (TM2) connects the periplasmic CitAP PAS sensor domain and the cytosolic histidine kinase (HK) effector domain (Fig. 2a). We decided to replace this TM2 helix (residues 179 to 199) of wild-type CitA by the YtvA Jα linker, to allow for soluble expression in *E. coli*. As suggested for full-length CitA, we reasoned, that in the here designed, potentially ligand-binding controlled lipase, the conformational change induced by ligand binding in the CitAP PAS domain could be transmitted *via* the Jα linker to affect BsLA activity.

Lipase activity of CitAP-BsLA depends on citrate. The gene-fusion coding for CitAP-BsLA was expressed in *E. coli* as a hexa-histidine (His6)-tagged fusion protein and purified to homogeneity by immobilized metal affinity chromatography and preparative size exclusion chromatography. A specific activity of 509 ± 5 U/mg was determined for purified CitAP-BsLA, while purified wild-type BsLA showed an activity of 181 ± 3 U/mg with

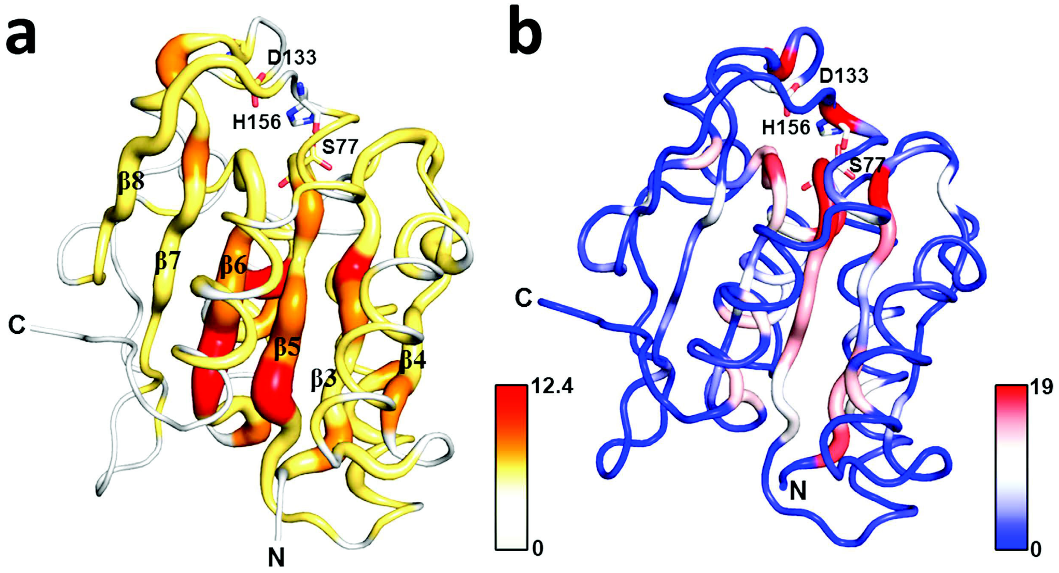


Figure 1. Comparison of evolutionary-coupling analyses **(a)** and site-saturation scanning mutagenesis data **(b)** mapped onto the X-ray structure of BsLA. Evolutionary coupled residues were inferred from a multiple sequence alignment using the EVcoupling webserver (www.evfold.org). The obtained evolutionary constraints (EC) values were mapped onto the X-ray structure of BsLA (PDB Entry: 1I6W)[30]. The magnitude of the obtained EC scores is color-coded (low values in yellow; high values in red). Additionally, EC values are encoded by sausage thickness representing the magnitude of the EC score. For orientation, the central β-scaffold (β3- β8) of BsLA is labelled according to topological order[30]. The number of inactive BsLA variants per residue was obtained from a complete site-saturation mutagenesis dataset (**b**) and mapped onto the BsLA X-ray structure. The number of inactive variants is encoded by color (blue: low values; red: high values) and sausage thickness. The N- and C-termini of BsLA are indicated. The residues of the catalytic triad, Ser77, Asp133 and His156 are shown as sticks with oxygen in red, carbon in grey and nitrogen atoms in blue. The color-bars next to the respective figure represent the plotted scale of EC values and the number of inactive BsLA variants per mutated site.

p-nitrophenylbutyrate as a model substrate. This suggests that fusion of CitAP to BsLA had no negative influence on the lipolytic activity of BsLA. On the contrary, the specific activities of CitAP-BsLA exceeded those of the isolated wild type BsLA. This observation might be related to the fact that fusion of CitAP to BsLA results in an increased solubility of the protein. While BsLA starts to aggregate at pH 10 at concentrations higher than 1 mg/ml, CitAP-BsLA can easily be concentrated to 5–10 mg/ml (data not shown). The effect is even more pronounced at neutral pH values, i.e. under assay conditions. This might result in higher stability of the fusion protein under assay conditions and thus could account for the increased apparent specific activity.

To address citrate sensitivity of CitAP-BsLA, we performed lipase assays in the presence of different concentrations of sodium citrate. Figure 2b shows the dose response curve recorded for the citrate-dependence of CitAP-BsLA lipase activity, displaying a clear sigmoidal response, characteristic for specific binding interactions and ligand-dependent functional regulation (Fig. 2b; red line). In contrast, isolated wild-type BsLA, without attached sensor domain, did not show any response toward citrate in the tested concentration range (Fig. 2b; blue line). From the fit of experimental data, an apparent K_D of 32 ± 8 μM and a Hill coefficient (n_H) of 0.94 ± 0.11 can be derived. During setup of the lipase assay for CitAP-BsLA, we realized, that the detergent Triton-X100 (TX100), which is added to the assay to solubilize the hardly water-soluble lipase substrate, apparently influences the magnitude of the functional citrate dependent response of CitAP-BsLA. We therefore performed an experiment where we kept the citrate concentration constant but varied the TX100 concentration in the assay (Fig. 2c). Please note, that the maximally employed TX100 concentration (160 μM) is well below the critical micelle concentration (CMC) of the detergent (0.22 mM)[36]. In this way, we are able to derive dose response curves for the TX100-dependent response of CitAP-BsLA at three different sodium citrate concentrations (Fig. 2c). The dose response curves display sigmoidal character, indicative of specific binding of TX100 to the protein. At different citrate concentrations, different apparent K_D and n_H values are obtained. At a concentration of 1 mM citrate, an apparent K_D of 38 ± 1 μM can be derived, with a Hill coefficient of 3.64 ± 0.42. In the absence of citrate the K_D is increased to 67 ± 1 μM (n_H = 7.53 ± 0.42), revealing an increased inhibitory potential for TX100 in the presence of citrate. In order to further analyze the role of the detergent TX100 on the citrate dependent activity response of CitAP-BsLA and wild type BsLA, we determined the functional response, i.e. the lipolytic activity in the presence and absence of 1 mM citrate, at different TX100 concentrations (Supplementary Figure 10). While the measurement conducted using wild type BsLA does not show a clear TX100 dependency and a relatively large associated error, the measurement for CitAP-BsLA reveals a maximal activity response at approx. 50 μM TX100. In the absence of the detergent no functional response of CitAP-BsLA is observed. In light of those findings, the

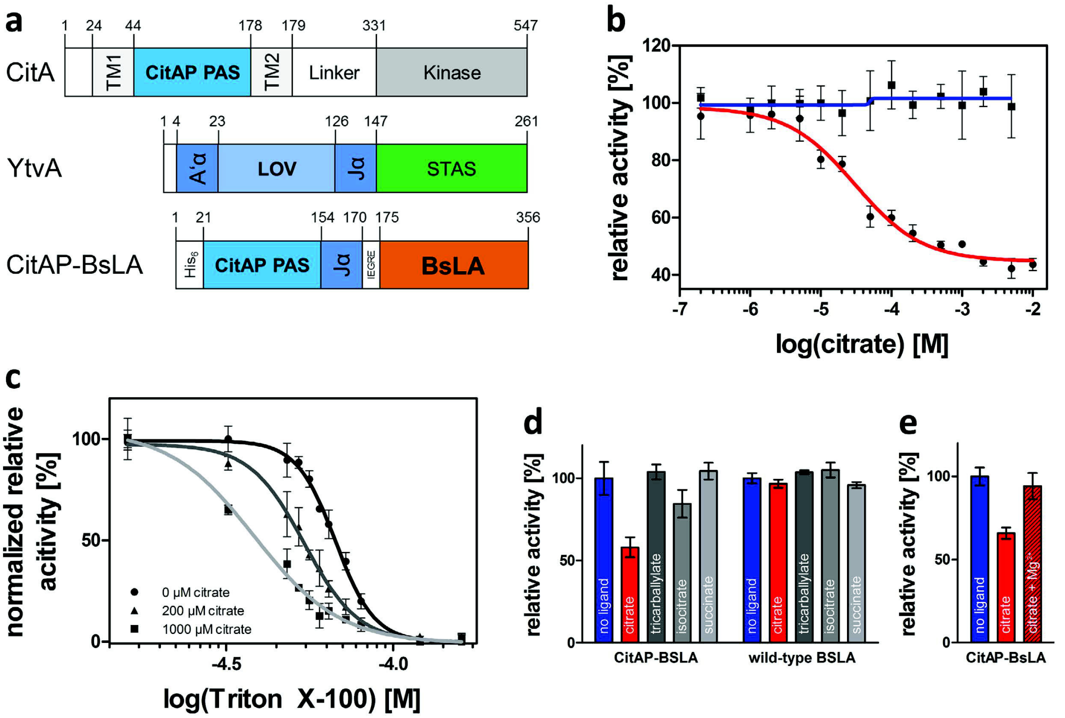


Figure 2. (**a**) Schematic representation of the multidomain architecture of the sensory histidine kinase CitA of *Klebsiella pneumoniae*, the blue-light photoreceptor YtvA of *Bacillus subtilis* and the here constructed artificial ligand-binding controlled lipase CitAP-BsLA. The numbers above the boxes denote amino acid numbers at domain boundaries of the respective full-length proteins. Abbreviations: TM1, TM2: transmembrane helices, CitAP PAS: periplasmic citrate-binding sensory domain of CitA, A'α: N-terminal N-cap α-helix of YtvA, LOV: blue-light sensing light oxygen voltage domain of YtvA, Jα: α-helical linker connecting LOV and STAS domains of YtvA, STAS: sulfate-transporter anti-sigma factor antagonist domain, His$_6$: Hexa-histidine tag, IEGRE: protease Factor Xa cleavage site, BsLA: *B. subtilis* Lipase A. Domain boundaries of CitA according to Kaspar *et al.*[37]. (**b**) The lipolytic activity of CitAP-BsLA (red-line, black circles) and wild type BsLA (blue-line, black squares) were determined in the presence of increasing concentrations of sodium citrate. The experimental data was fitted using a four parameter logistic dose-response model (red line). (**c**) The activity change in the presence of increasing concentrations of Triton X-100 was determined at 0 (black line, circles), 200 μM (dark grey line, triangles) and 1000 μM sodium citrate (light grey line, squares). (**d**) Sodium citrate, as well as the respective citrate analogues, were added to the assay in a final concentration of 1 mM and the lipolytic activity of CitAP-BsLA and BsLA was determined relative to the activity without ligand (**e**) The presence of Mg^{2+} ions, which are known to scavenge citrate[37], abolishes the functional response of CitAP-BsLA. 10 mM $MgCl_2$ was added to the assay containing CitAP-BsLA and 1 mM sodium citrate. Lipolytic activity was measured using *p*-nitrophenylbutyrate as substrate. Error bars depict the standard deviation of the mean derived from three independent measurements.

observed citrate-dependent reduction of CitAP-BsLA lipase activity has to be interpreted as a citrate-dependent modulation of TX100 inhibition of CitAP-BsLA.

CitAP-BsLA fusion and isolated CitA display similar ligand-binding characteristics. The specifity of CitAP-BsLA was further probed by using different citrate analogues. The isolated CitAP sensor domain was reported to be highly specific for citrate[37]. We therefore used isocitrate, succinate and tricarballylate as potential ligands and analysed the functional response of CitAP-BsLA. As expected for a highly specific citrate-sensor, CitAP-BsLA did not respond to any of the tested analogues (Fig. 2d). Similarly, the isolated BsLA protein did not show any change in activity due to presence of citrate analogs (Fig. 2d). Moreover, it was reported that Mg^{2+} ions can form a stable complex with citrate[37]. Therefore, the addition of $MgCl_2$ to the assay solution is expected to inhibit the citrate-dependent functional response of CitAP-BsLA by interfering with citrate-binding. As expected, addition of 10 mM $MgCl_2$ to the assay solution containing 1 mM citrate completely abolished the functional response of CitAP-BsLA (Fig. 2e). Please note that, all experiments using citrate analogues and $MgCl_2$ were performed in the presence of TX100, which indicates that the detergent does not influence the ligand-binding properties of the CitAP domain in CitAP-BsLA, i.e compared to the isolated CitAP domain.

Global citrate-induced structural changes in CitAP-BsLA probed by fluorescence spectroscopy. In order to assess global structural changes in CitAP-BsLA induced by citrate binding we initially monitored the fluorescence of the aromatic amino acid residues of CitAP-BsLA and wild-type BsLA. Excitation of tryptophan (Trp) residues of CitAP-BsLA and wild-type BsLA at 295 nm did not reveal any spectral changes due to the

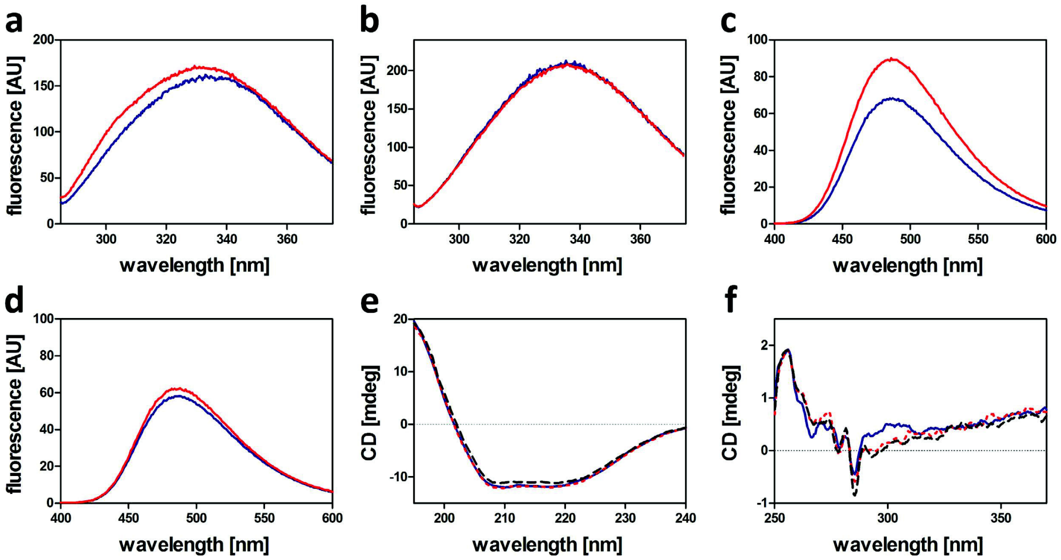


Figure 3. Fluorescence of aromatic amino acids of CitAP-BsLA (**a**) and wild-type BsLA (**b**). Protein samples, diluted to 3 µM in 10 mM glycin buffer pH 10 supplemented with 10 mM NaCl, were excited at 278 nm. (**c**) Fluorescence emission spectra of 4,4′-dianilino-1,1′-binaphthyl-5,5′-disulfonic acid (bis-ANS) of samples containing CitAP-BsLA or wild-type BsLA (**d**). 6 µM of bis-ANS was added to protein samples (3 µM) in 10 mM glycin buffer pH 10 supplemented with 10 mM NaCl. Bis-ANS was excited at 385 nm. All fluorescence emission spectra were recorded in the presence (red line) and absence (blue line) of 1 mM sodium citrate. Far-UV (**e**) and near-UV (**f**) circular dichroism (CD) spectra of CitAP-BsLA in 10 mM glycine buffer pH 10 supplemented with 10 mM NaCl (blue solid line), after addition of 1 mM sodium citrate (red dashed line) and in the presence of 1 mM sodium citrate and 50 µM Triton X-100 (black dashed line). All spectra were recorded at 25 °C.

presence of citrate (Supplementary Figure 3). In contrast, excitation at 278 nm, thus exciting both tyrosine (Tyr) and Trp residues, resulted in distinctly different emission spectra for samples with and without 1 mM citrate. In the presence of 1 mM citrate, an increased emission (Fig. 3a) with a maximum at around 303 nm is observed for CitAP-BsLA (maximum derived from the resulting difference spectrum, (data not shown). This small, but reproducible, increase in fluorescence emission could be attributed to an increased emission from Tyr residues and thus to reduced Förster-Resonance-Energy Transfer (FRET) between Tyr and Trp residues in the protein. In contrast, no citrate-dependent change in Tyr-Trp FRET was observed for the isolated BsLA protein (Fig. 3b). To further probe global structural changes in CitAP-BsLA we employed the fluorescent dye 4,4′-dianilino-1,1′ -binaphthyl-5,5′-disulfonic acid (bis-ANS)[38,39], which binds to hydrophobic surface patches of proteins[40]. Upon dye binding, an increased fluorescence emission as well as a blue-shift of the emission maximum, compared to the free dye, can be observed. Bis-ANS emission was markedly increased for CitAP-BsLA samples containing 1 mM citrate (Fig. 3c), suggesting that upon citrate binding additional hydrophobic surface patches become exposed.

In contrast, only a negligible citrate-dependent change in bis-ANS fluorescence was observed for a sample of the isolated wild-type BsLA protein (Fig. 3d).

Far- and near-UV circular dichroism (CD) spectroscopy hint at citrate-induced conformational changes. Far-UV CD spectroscopy was used to analyse CitAP-BsLA for potential secondary structural changes associated with citrate-binding. Additionally, due to the observed effect of TX100 on BsLA activity a CitAP-BsLA sample containing the detergent was included (Fig. 3e). Far-UV CD spectroscopy suggested that CitAP-BsLA is well folded in solution but does not reveal any significant secondary structural changes due to presence of sodium citrate or TX100. This notion is further corroborated by deconvolution of the corresponding CD spectra (Supplementary Table 3) and a comparison to the theoretical secondary structure composition of the fusion protein derived from the X-ray structures of the components (Supplementary Table 4). This further suggests that TX100 does not influence the proper folding of CitAP-BsLA. In contrast, near-UV CD spectra revealed citrate-dependent tertiary structural changes independent of the presence of TX100 (Fig. 3f). In the presence of citrate, we observed a decrease in ellipticity at around 265 nm and increased values at 285 nm as well as in the region between 290 nm and 310 nm. While the changes seen at around 285 nm may be attributed to a rearrangement of Tyr side chains which are distributed throughout the whole fusion protein (15 residues), the most pronounced citrate-dependent spectral changes are observed in the 290 nm–310 nm region corresponding to the absorption band of Trp residues. Since Trp residues are only found within the BsLA domain (W31 and W42 of BsLA) of the fusion protein, those spectral changes must be interpreted as a tertiary structural change in the BsLA part of the construct.

	relative oligomer distribution			
	AUC[$,§]		SAXS[&]	
	−	+	−	+
	[%]	[%]	[%]	[%]
Monomer	73	85	73	86
Dimer	23	12	27	14
>Dimer	4	3	n.d	n.d

Table 1. Comparison of the relative oligomer distribution of CitAP-*Bs*LA samples with (+) and without (−) 1 mM sodium citrate derived from analytical ultracentrifugation (AUC) and small-angle X-ray scattering (SAXS) data obtained for CitAP-*Bs*LA at low concentration (0.5 mg/ml). [$]For AUC experiments, the relative oligomer distribution was estimated using Bayesian statistics assuming the presence of discrete species of known molecular mass. [&]From SAXS experiments the average molecular mass was determined, which corresponds directly to the average molecular mass of a population between monomer and dimer with known molecular mass. The molecular mass was determined from the concentration normalized forward scattering and the excluded volume multiplied by the protein density, and the average molecular mass is given. Theoretical molecular mass: monomer: 38.5 kDa, dimer: 77 kDa; [§]Values represent the mean of two independent sedimentation velocity runs, with an experimental error below 5%.

	sedimentation coefficent (AUC)[§]		Guinier Radius R_g (SAXS)[&]		frictional ratio (AUC)[§]		D_{max} (SAXS)[&]	
	−	+	−	+	−	+	−	+
	[S]	[S]	[nm]	[nm]	[f/f0]	[f/f0]	[nm]	[nm]
Monomer[$]	3.53	3.61	2.94	2.81	1.15	1.12	10.1	9.7
Dimer[$]	4.49	5.15	3.37	3.44	1.37	1.25	11.1	11.6
>Dimer	6.65	8.36	n.d	n.d	n.d.	n.d.	n.d	n.d

Table 2. Comparison of analytical ultracentrifugation (AUC) and small-angle X-ray scattering (SAXS) data for CitAP-*Bs*LA samples with (+) and without (−) 1 mM sodium citrate. [$]Molecular mass: monomer: 38.5 kDa, dimer: 77 kDa; n.d. not detected; [&]*Rg* and *Dmax* were determined from SAXS measurements at 0.5 and 5 mg/mL. At low concentration, the determined parameters primarily inform about the structural properties of the monomer, which is the predominwant population at that concentration. At high concentration the dimer is the prevalent species and the structural parameters inform primarily about the properties of the dimer. Dmax and Rg were determined from the distance distribution in real space. [§]Values represent the mean of two independent sedimentation velocity runs, with an experimental error below 5%.

Citrate-dependent quaternary structural changes studied by small angle X-ray scattering (SAXS) and analytical ultra-centrifugation (AUC). Many, though not all, bacterial SHKs are functionally active as dimers. In those cases, signal relay was suggested to occur *via* a rotation/piston/torque-like movements[41–43] initiated in the sensor domains which are transduced through rigid coiled-coils in case of soluble SHKs, or transmembrane (TM) helices in case of membrane bound SHKs[44–46]. Given that CitAP is reported to be a dimer[9,46], while BsLA appears to be monomeric, the question arises whether the fusion protein CitAP-BsLA is a monomer or dimer in solution. We therefore initially used AUC to determine the oligomerization state of CitAP-BsLA in solution with or without 1 mM citrate for samples of low concentration (0.5 mg/ml) and subsequently employed SAXS to cover a broader concentration range (0.5–5 mg/ml) to address the possibility of concentration-dependent oligomerization and obtain a low-resolution structural model of the fusion protein. AUC and SAXS data for CitAP-BsLA are summarized in Supplementary Figure 4 and Supplementary Figure 5, respectively. At low concentrations, both AUC and SAXS reveal the presence of monomeric and dimeric species of CitAP-BsLA (Table 1). Moreover, both methods provide an identical estimation of the relative monomer:dimer distribution with the monomer (SAXS: 73%, AUC: 73%) representing the predominant species in the absence of citrate. At low concentrations, this equilibrium is slightly influenced by the presence of citrate, shifting the equilibrium further toward the monomer (SAXS: 86%, AUC: 85%) (Table 1). Additionally, a small but significant citrate-dependent increase in the sedimentation coefficient of the CitAP-BsLA monomer/dimer is observed in AUC experiments (Table 2). Moreover, a citrate-dependent change in the AUC determined frictional ratio f/f$_0$ can also be seen in the radius of gyration (R_g) and the maximal elongation of the molecule (D_{max}), derived from SAXS experiments (Table 2). Thereby, the frictional ratio reflects both the shape and hydration of the protein molecule and can be considered as an approximate measure of the molecules' globularity. Here, a smaller frictional ratio (f/f0) is observed in the presence of citrate indicative of a more globular conformation and/or lower hydration. This observation is corroborated by the SAXS data, where smaller R_g and D_{max} values are found in the presence of citrate for the monomer. Please note that we cannot rule out that the change in R_g and D_{max} observed by SAXS is caused by the altered monomer:dimer ratio between citate-free and citate-bound protein samples (*vide supra*).

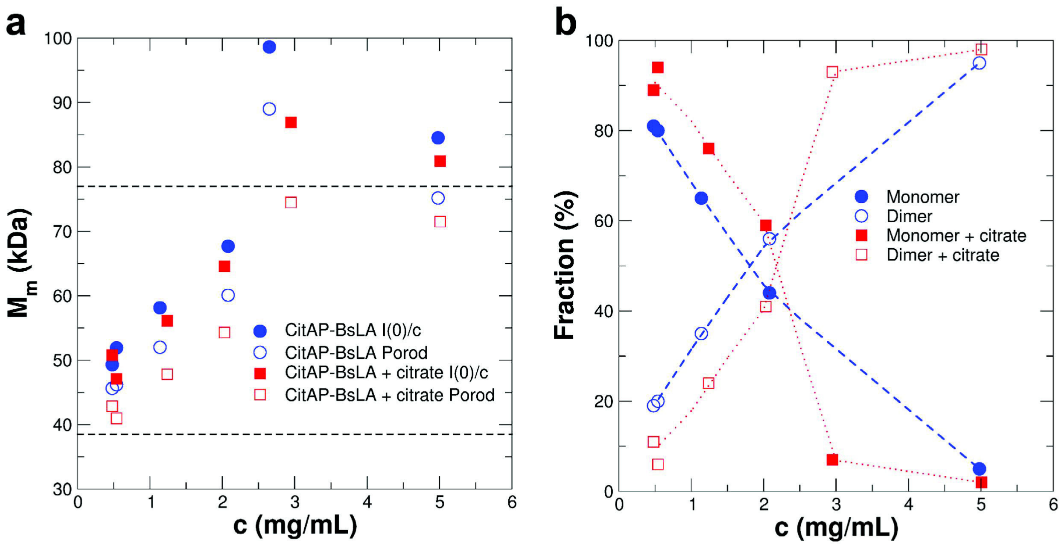


Figure 4. **(a)** Average molecular mass of the scattering particle determined from the concentration normalized forward scattering I(0)/c or from the Porod volume. **(b)** Monomer and dimer fraction of CitAP-BsLA as a function of the protein concentration. Values were determined using the M_m determined from the Porod volume. The respective values were obtained by analysing SAXS data recorded for CitAP-BsLA in the absence (blue) and presence (red) of 1 mM sodium citrate.

Dimerization of CitAP-BsLA depends on protein concentration. SAXS measurements provide direct information about the oligomerization state of a protein. Here, the average molecular mass of the scattering particle was calculated (Fig. 4a) from the forward scattering $I(0)/c$ normalized by the protein concentration c, which is directly proportional to the molecular mass M_m of the scattering particle, and by the Porod volume multiplied with the appropriate protein density[47]. By comparison of scattering curves for CitAP-BsLA at different protein concentrations, in both the absence and the presence of 1 mM citrate, a concentration dependent monomer:dimer equilibrium was observed (Fig. 4b). At all protein concentrations, this equilibrium was shifted by the presence of citrate, resulting in a reduction of the dimer content (Fig. 4b). At concentrations of about 5 mg/ml more than 90% of CitAP-BsLA was present as a dimer. Additionally, from Fig. 4b, a dissociation constant of approximately 1.8 to 2.2 mg/ml (24–29 μM) for the dimer can be estimated, indicating that dimer association is rather weak. Probably, this is a direct consequence of fusing monomeric BsLA to dimeric CitAP thus altering the dimer forming capacity of CitAP by presenting non-evolved protein-protein interaction via the BsLA part of the fusion protein. Given the rather high dissociation constant of the dimer, it seems reasonable to assume that under assay conditions (at 1 μM protein concentration)) CitAP-BsLA is present as a monomer. This implies that the citrate-induced structural changes in monomeric CitAP-BsLA are sufficient to induce the observed functional response.

Computational modelling and SAXS envelope reconstructions. In order to gain more insight into the structural arrangements of CitAP and BsLA in the monomer as well as the assembly of the CitAP-BsLA dimer, we reconstructed low-resolution bead models from SAXS data, further investigated the resulting models using molecular dynamics (MD) simulations, and compared the final MD-derived models to the experimental data obtained in SAXS experiments of CitAP-BsLA with and without citrate. Four different starting models of the dimeric CitAP-BsLA complex were obtained using different strategies. Details about model generation can be found in the Materials and Methods section and the Supplementary Materials. The models differed with regard to the manner of generation and the conformation of the CitAP-PAS domain, being either in the citrate-bound (closed) (models: $M_{low\text{-}cit}$, $M_{high\text{-}cit}$) or citrate-free (open) state (models $M_{low\text{-}free}$, $M_{high\text{-}free}$) (Supplementary Table 2). In order to improve the initial models, a 100 ns MD simulation was performed for each dimeric assembly (Supplementary Figure 7). To evaluate the quality of the resulting models sampled during the MD runs, a theoretical scattering curve was calculated for every 200 ps snapshot of each trajectory and fitted against the experimental data (with and without citrate) (Supplementary Figure 8). Hereby, only the MD simulation of $M_{low\text{-}free}$ yielded acceptable χ values, as a measure of the goodness of the fit between the experimental data and the theoretical model. Thus, only the data of the $M_{low\text{-}free}$ simulation is summarized in Fig. 5. The respective data for all models is given in the Supplementary Materials (Supplementary Figures 7 and 8). During the MD simulation, $M_{low\text{-}free}$ (Fig. 5a) and most of the other models (Supplementary Figure 7) underwent significant structural rearrangements. Figure 5b depicts the evolution of χ during the $M_{low\text{-}free}$ MD simulation. The corresponding data for the alternative models is shown in Supplementary Figure 8. Hereby, the MD-derived models were compared against the experimental SAXS data measured at protein concentrations at which CitAP-BsLA is predominately dimeric (5 mg/ml), both in the presence (blue) and absence of citrate (red) (Fig. 5b and Supplementary Figure 8). For both cases, the data shows the same overall trend and appears to be scaled by a constant factor, indicating a better experimental accuracy of the scattering curves obtained in absence of citrate and hence a larger χ value. The

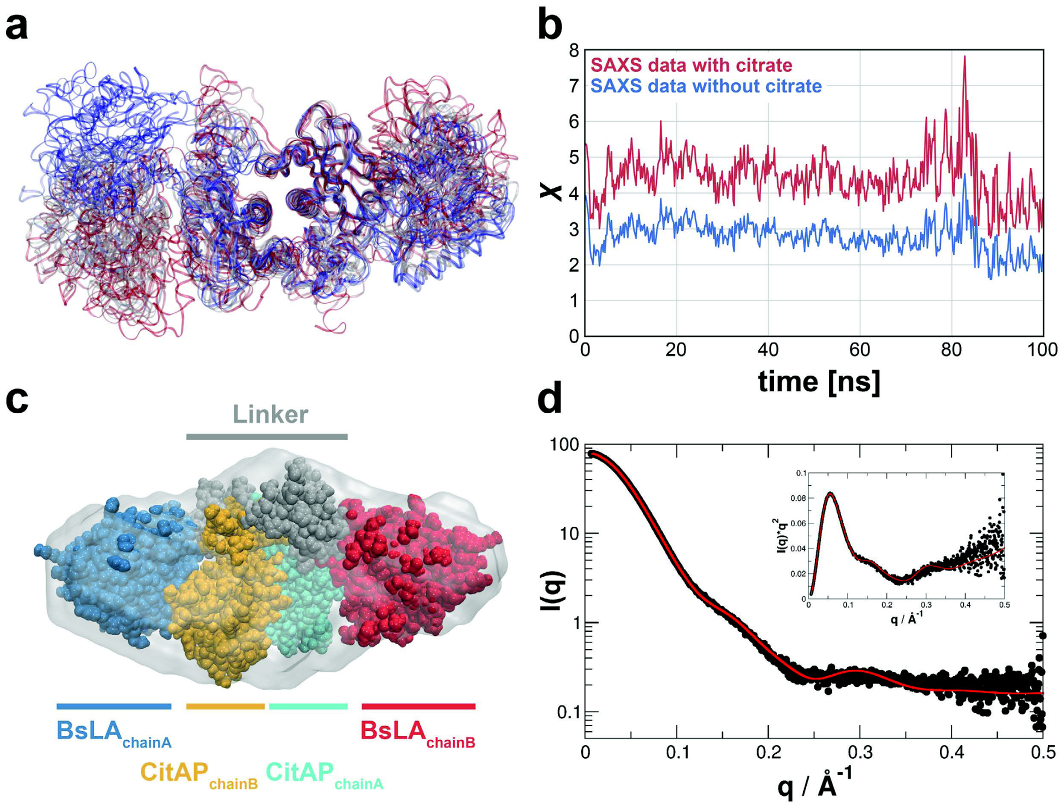


Figure 5. (**a**) Structural changes during the molecular dynamics (MD) simulations of the dimeric $M_{low\text{-}free}$ CitAP-BsLA model. The proteins are shown as ribbons and the colors represent structures at different times, changing from red at t = 0 ns to blue at t = 100 ns. (**b**) Time evolution of χ during the MD simulations of the $M_{low\text{-}free}$ model. Deviation between the model and experimental data measured at 5 mg/ml (blue) and without citrate (red) are shown. (**c**) Final model of the CitAP-BsLA dimer complex superimposed onto the SAXS derived low-resolution envelope obtained from SAXS data at high protein concentration in the presence of citrate. (**d**) SAXS scattering curve recorded for CitAP-BsLA at a protein concentration of 5 mg/ml in the presence of citrate (black dots) and the CRYSOL-fitted theoretical scattering curve of the final CitAP-BsLA dimer model (red line).

similar trends in both datasets are likely due to highly similar structures of the fusion protein dimers with and without citrate. After a structural rearrangement at around 80 ns, the χ value reaches a minimum of 1.50 when compared to the experimental scattering data at in the presence of citrate. Thus, this model appears closest to the physical structure of the fusion protein. This model was further optimized by constructing symmetric dimers by superimposing chain A onto the Cα atoms of chain B and *vice versa*. This yielded a structure with a χ value of 1.3 when chain A is superimposed onto chain B. A subsequent energy minimization of this model further improved χ to 1.16. The resulting structure represents the best model in terms of χ and is thus taken as the final model (Fig. 5c). Models from the last frame of each MD simulation are shown in Supplementary Figure 9. For comparison, the MD-derived final model was fitted to a low-resolution SAXS envelope obtained from *ab initio* bead-modelling (Fig. 5c). For the final model, the maximal elongation (D_{max}) and the radius of gyration (R_g) were calculated and compared to the corresponding experimental values. Both values (D_{max} = 12.2 nm; R_g = 3.23 nm) are in good agreement with the corresponding experimentally derived values (D_{max} = 11.6 nm; R_g = 3.44 nm; see Table 2). As depicted in Fig. 5d, the corresponding theoretical scattering curve agrees nicely with the experimental one.

Discussion

The computational prediction of allosteric communication pathways in signalling proteins represents an important line of investigation in both basic science and applied pharmaceutical research either enabling or facilitating the design of inhibitors for a given pharmaceutical target. Likewise, the rational design of allosteric communication, so far successful in a few cases only in the recent past, is still challenging due to the lack of an atomic level understanding of the underlying signal-relay principles. Utilizing the small, not-allosterically regulated, lipase A from *Bacillus subtilis* (BsLA) as model protein, we show that sequence-based methods which capture the evolutionary coupling (see ref. 28 and references therein) between residues in a protein family can yield valuable information about the functional importance and hence potential modes of information flow within proteins (Fig. 1a). So far, those bioinformatic predictions have only been in rare cases experimentally validated by alanine-scanning[48] or site-saturation mutagenesis[49]. The here presented site-saturation mutagenesis data for

BsLA (Fig. 1b) shows that both computational predictions and the experiment essentially yield similar results. Both evolutionary-coupling analyses and site-saturation scanning mutagenesis identified a stretch of residues at the N-terminus of BsLA as functionally important ("mutationally sensitive") and evolutionary coupled (Fig. 1), highlighting the complementarity of both methods. Based on this data, the N-terminus of BsLA was chosen as the most promising site for fusion of the CitAP sensory PAS domain expected to result in perturbation of BsLA function by ligand-binding induced conformational changes in the sensory domain. The presented strategy yielded a well folded artificial two-domain enzyme (CitAP-BsLA), whose function could readily be controlled by citrate binding in the fused sensory domain. Hereby, CitAP-BsLA showed decreased activity with increasing citrate concentrations (Fig. 2b). When purified CitAP-BsLA is stored for extended periods of time at 20 °C, proteolytic cleavage of the two domains is observed (Supplementary Figure 11). In consequence, in samples stored for 9 days at 20 °C, the covalent linkage between the CitAP-PAS and BsLA domains is to an large extend broken and the corresponding functional response is abolished (Supplementary Figure 12). This observation provides additional evidence for signal-relay between the citrate binding CitAP-PAS domain and BsLA. With respect to the mechanism of inactivation, studies using the detergent TX100 suggested that citrate represents an indirect modulator of CitAP-BsLA inhibition by TX100 rather than a direct allosteric inhibitor (Fig. 2c).

Using dose-response data, an apparent K_D for citrate of $32 \pm 8\,\mu M$ and a Hill coefficient n_H of 0.94 ± 0.11 were determined for CitAP-BsLA, both indicative of specific non-cooperative binding. Hereby, the K_D value of CitAP-BsLA is slightly larger than the one of the isolated CitAP sensor domain, for which a K_D of 11.1 at pH 8.0 was determined by isothermal titration calorimetry[37]. This discrepancy could for example arise from an altered citrate accessibility of the CitAP-PAS domain in the fusion protein, i.e. by a direct interaction between the two domains or by modulation of the quality or magnitude of the conformational change that is induced by citrate binding to the sensor CitAP PAS domain. Complementary, studies using different citrate analogs revealed a high specificity of the ligand-binding controlled enzyme with very similar properties as the isolated CitAP sensor domain[37]. Thus, CitAP-BsLA clearly represents an example of a designed artificial, highly active, yet very specific ligand-binding controlled enzyme.

The membrane-bound SHK CitA that constitutes the sensory receptor of the CitA/CitB two-component system (TCS) of *K. pneumoniae*, is responsible for induction of citrate fermentation genes under anoxic conditions in the presence of environmental citrate[50]. Citrate-binding to the periplasmic CitAP PAS domain constitutes the trigger for structural-changes within the sensory domain that are transmitted *via* the connecting TM2 helix to the effector HK, eventually leading to quaternary structural changes within the CitA dimer which probably influences HK autophosphorylation[46]. Based on nuclear magnet resonance (NMR) spectroscopic and X-ray data, obtained for the citrate-free and citrate-bound form of the isolated CitAP PAS domain, it was suggested that citrate-binding to CitAP PAS results in closing/bending of the PAS β-scaffold by a rearrangement of the minor (residues 99–104) and major loop (residues 68–90)[46]. Moreover, the citrate-free structure of CitAP PAS lacks electron density in the surface exposed major loop, indicative of increased flexibility[46]. This hypothesis is fully corroborated by our 100 ns MD simulation of the citrate-free form of the isolated CitAP-PAS domain, where we observed a large-scale rearrangement of surface exposed loops of the citrate binding site and a stretching/flattening of the central β-scaffold in the absence of bound citrate (Fig. 6a; Supplementary Figure 6). In terms of global structure, it is assumed that full-length CitA possesses an elongated parallel dimeric structure with gross structural similarity to other soluble PAS domain containing SHKs like bacteriophytochromes[51] or the artificial light-dependent HK YF1[42]. According to our simulations and SAXS data, it is unlikely that CitAP-BsLA adopts such an elongated parallel dimer structure. In the best model obtained from MD simulations and SAXS envelope reconstructions, we observed a dimeric arrangement of the CitAP-PAS domain with the BsLA domain being arranged parallel to the CitAP-PAS dimer flanking the sensory module on both sites (Figs 5c and 6b). The catalytic triad of BsLA is accessible in both subunits of the dimer as well as in both monomer models (Fig. 6b), enabling robust lipolytic activity of the fusion protein. While we believe that the overall subunit arrangement revealed by the SAXS-guided MD simulations is physically feasible, detailed structural questions cannot be addressed using the present model. In particular, the citrate-induced structural changes of CitAP-BsLA appear globally too subtle to be modelled accurately from SAXS data, since both the MD-derived models as well as the SAXS envelopes of the dimer in the citrate-free and citrate-bound form are very similar. To better understand the mechanism of the citrate-dependent functional response of CitAP-BsLA a number of complementary biochemical and biophysical techniques were used, which together hint at global tertiary/quaternary structural changes associated with citrate binding and hence ligand-binding dependent control of CitAP-BsLA. Several mechanistic scenarios could account for the observed citrate-dependent modulation of the TX100 inhibition of CitAP-BsLA. Based on our data, the most likely explanation is a small-scale structural rearrangement of the two domains relative to each other (illustrated in Fig. 6c), which would by congruent with the observed differences in Trp/Tyr fluorescence, interpreted as different Tyr/Trp FRET efficiencies in the presence and absence of citrate (Fig. 3a) as well with the small change in compactness of the molecule observed by AUC and SAXS (Table 2). This rearrangement results in the exposure of additional hydrophobic surface patches (marked by asterisks in Fig. 6c), as evidenced by bis-ANS binding studies (Fig. 3c,d) and an decreased K_D for Triton X-100 in the presence of citrate (Fig. 2c), which allows increased binding of the non-ionic detergent TX100 facilitating increased inactivation of the BsLA domain in the presence of citrate. Likewise, this rearrangement could impose strain to the BsLA structure relayed by the Jα-linker to first β-strand (β3) of the BsLA domain, which was identified by our computational and mutagenesis studies as mutationally sensitive (Fig. 1), onto the active site, in turn inactivating the enzyme. However, based on current data and without a crystal structure of the fusion protein, it is impossible to delineate between these scenarios.

In conclusion our study highlights the complementarity of evolutionary coupling analyses and site-saturation mutagenesis in identifying functionally important residues and potential pathways of information flow within proteins. As exemplified here for a small bacterial lipase, this information can be exploited for the construction

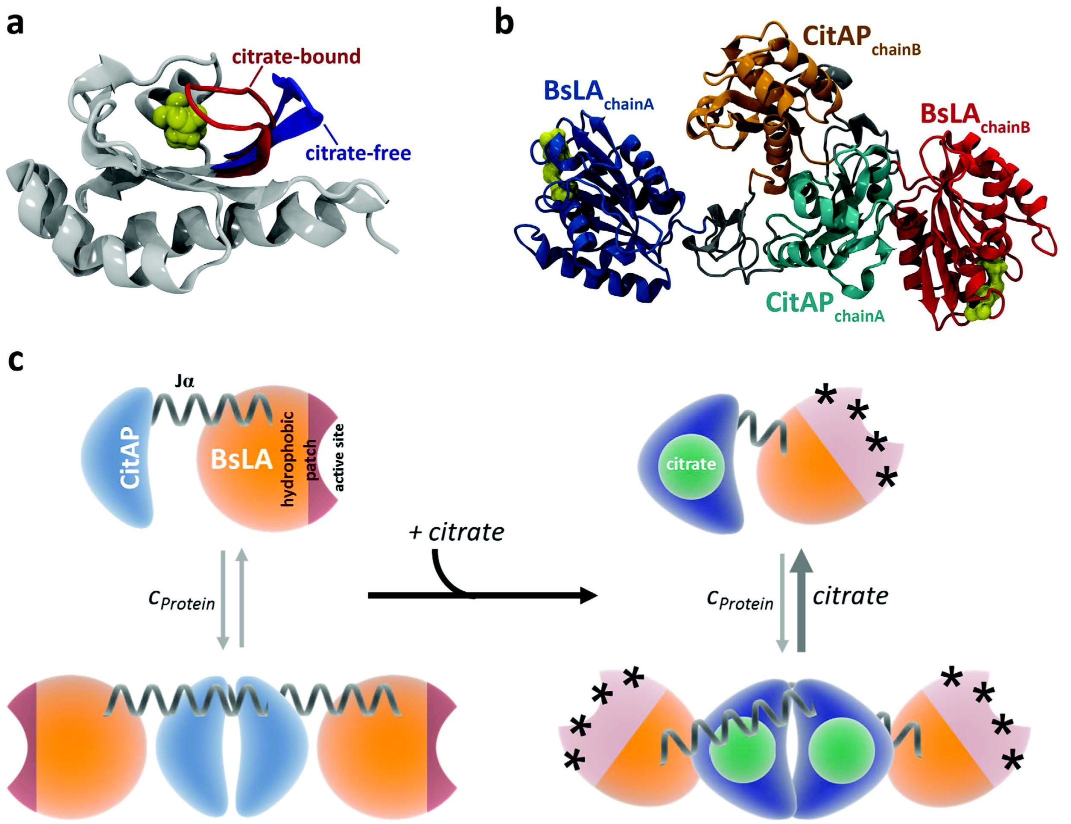


Figure 6. (**a**) Citrate-bound and citrate-free structure of the CitAP-PAS domain, as used for model-building. The citrate-bound structure (PDB ID: 2J80) is shown in grey, with the closed lid, containing the minor loop (residues 99–104), colored in red. Citrate is depicted as yellow van der Waals (vdW) surface. The corresponding citrate-free structure was obtained from a 100 ns MD simulation of the citrate-bound structure. For clarity, only the opened lid is shown (colored in blue). (**b**) Best model obtained from MD simulations. The BsLA and CitAP-PAS domains of the dimer are colored as in Fig. 5c. The catalytic triad of the lipase is shown in yellow. (**c**) Schematic illustration of a potential mechanism for the regulation of BsLA activity. The citrate-induced rearrangement of the CitAP-PAS and BsLA domains relative to each other results in the exposure of additional hydrophobic surface patches (marked by asterisks) which allows for increased binding of Triton-X100 to BsLA, which acts as an inhibitor of BsLA.

of artificially controlled multidomain proteins. The simplicity of the here employed fusion strategy poses the interesting question if the catalytic mechanism of some enzymes is evolutionary optimized in a way that allows it to be easily perturbed by small conformational changes and/or non-natural protein-protein interactions. Such an evolutionary design could easily be realized by domain fusion and could account for the ubiquitous presence of allostery and multidomain sensory receptors.

Methods

Molecular biological and microbiological methods. Details about general molecular biological methods, site-saturation mutagenesis, fusion protein construction, expression of gene fusions and protein purification can be found in the Supplementary Materials.

Evolutionary coupling analysis. Evolutionary coupling analysis was carried out for lipases using the BsLA sequence (Uniprot ID: P37957) as input sequence for the EVcouplings webserver (www.evfold.org). For the generation of the alignment the JackHHMer software (5 interations)[52], implemented as part of the EVcouplings webserver, was utilized, to search the Uniprot database[53] for sequences similar to BsLA. We ran an unrestrained search not limiting the number of sequences in the alignment, which retrieved 149.524 sequences with an E-value cutoff of 10E-3, covering 168 out of 181 residues of the query BsLA sequence. In a subsequent restrained run we limited the number of sequences in the alignment to 20.000 while using the same E-value cutoff. This search produced an alignment containing 20.000 sequences covering 176 out of 181 residues of the query sequence. Covariation information was inferred employing the plmDCA (pseudolikelihood maximization for Potts models with direct coupling analysis algorithm)[54], implemented in the EVcouplings webserver. Evolutionary constraints (EC) values were mapped onto the B-factor field of the BsLA X-ray structure (PDB ID: 1I6W) and visualized by using Pymol v1.7.0.0 (Schrödinger Inc., NY, USA).

High-throughput lipase assay and determination of the mutational sensitivity of BsLA. BsLA was used in a previous study as a model protein to assess the full protein landscape towards ionic liquid resistance[34] and detergent tolerance[33]. All variant genes were expressed in *E.coli* BL21(DE3) and fused to a PelB secretion signal, which led to an unspecific release into the culture supernatant. The mean activity against *p*-nitrophenylbutyrate (*p*-NPB) of 96 *E. coli* BL21(DE3) clones harbouring the pET22b(+) vector with no insert was used to determine the experimental background and standard deviation (σ). All variants with an activity lower than the mean of the experimental background ±3 σ were considered as inactive. The B-factor of the pdb file (PDB-ID: 1I6W) was replaced with the absolute number of inactive variants for each of the 181 BsLA amino acid positions to generate the representation shown in Fig. 1.

Determination of citrate-dependent lipase activity. BsLA lipolytic activity was measured using *p*-NPB as the substrate at 37 °C. Activity measurements were carried out in 1 cm disposable cuvettes with 100 mM 3-(N-morpholino)propanesulfonic acid (MOPS) buffer, pH 7,5 supplemented with 50 μM Triton X-100 (TX100) as assay buffer. Substrate stock solutions were prepared in acetonitrile containing 16 mM *p*-NPB. A suitable volume of enzyme was pipetted into the cuvette placed into a Beckman DU650 UV/Vis spectrophotometer temperature controlled to 37 °C. Assay buffer was heated to 37 °C in a thermo-block. Immediately before the activity measurement, the assay buffer was mixed with the substrate stock solution to yield an assay substrate concentration of 0.8 mM. This mixture was vortexed briefly and then added to the enzyme solution in the cuvette. Hydrolysis of *p*-NPB was monitored by measuring the release of *p*-nitrophenolate (*p*-NP) at 410 nm over 60 seconds. The lipolytic activity of the constructs was calculated using the molar extinction coefficient of *p*-NP (15.000 M^{-1} cm^{-1}). All measurements were carried out in triplicate. For the determination of the citrate-dependent lipolytic activity of BsLA and CitAP-BsLA sodium citrate was added to the reaction mixture in concentrations up to 10 mM. Dose-response curves were obtained by plotting the relative lipolytic activity against the logarithmic citrate concentration. Dose-response data was fitted using Origin 9 G employing a four parameter logistic dose-response model according to the following equation:

$$y = A_{min} + \frac{A_{max} - A_{min}}{1 + 10^{(\log K_D - C) \times n_H}} \tag{1}$$

With A_{max} and A_{min} representing the top and bottom asymptotic activity values, K_D the apparent dissociation constant, C the citrate concentration and n_H the Hill slope.

Determination of the effect of Triton-X100 (TX100) on the citrate-dependent activity of CitAP-BsLA. The TX100 dependence of the citrate-response of CitAP-BsLA was determined using the same experimental setup as described for the citrate-dependent lipase activity assay. The sodium citrate concentration was kept constant at 0 mM, 0.2.mM and 1 mM while the TX100 concentration was varied from 0 mM to 160 μM. All measurements were performed in triplicate and the data was analysed as described above.

Tryptophan fluorescence. The fluorescence of aromatic amino acids was monitored in the presence and absence of 1 mM citrate for CitAP-BsLA and wild-type BsLA. For all measurements 1 cm quartz cuvettes (Hellma Analytics, Müllheim, Germany) were used employing a Cary Eclipse™ spectrofluorimeter (Varian GmbH, Darmstadt, Germany) temperature controlled to 37 °C. A bandwidth of 5 nm was used in both the excitation and emission. CitAP-BsLA and wild-type BsLA were diluted to 3 μM in 10 mM glycin buffer pH 10 supplemented with 10 mM NaCl. Tryptophan fluorescence emission was measured from 300 nm to 400 nm while exciting the sample at 295 nm. When the sample is excited at 278 nm, both Trp and Tyr sidechains are excited and hence contribute to the observed fluorescence emission spectra which were recorded from 280 nm to 400 nm. The influence of citrate on the emission properties of the aromatic amino acids of the protein was determined by adding 1 mM of sodium citrate to the same protein sample.

4,4′-dianilino-1,1′-binaphthyl-5,5′-disulfonic acid (bis-ANS) fluorescence. Bis-ANS binding studies were carried out in 1 cm quartz cuvettes (Hellma Analytics, Müllheim, Germany) using a Cary Eclipse™ spectrofluorimeter (Varian GmbH, Darmstadt, Germany) temperature controlled to 37 °C. Bis-ANS was dissolved in acetonitrile and added to protein solutions to a final concentration of 6 μM. Protein samples were diluted to 3 μM with 10 mM glycine buffer pH 10 supplemented with 10 mM NaCl. The influence of citrate on the emission properties of bis-ANS was determined by adding 1 mM of sodium citrate (dissolved in 10 mM glycine buffer pH 10 supplemented with 10 mM NaCl) to the sample containing the dye and the respective protein. Bis-ANS emission spectra were recorded from 400 nm to 600 nm by exciting the dye at 385 nm (emission and excitation band-width: 5 nm).

Circular dichroism (CD) spectroscopy. Far-UV circular dichroism (CD) spectra were recorded using 2 mm quartz cuvettes (Hellma Analytics, Müllheim, Germany) using a JASCO J-810 spectropolarimeter temperature controlled to 37 °C. All protein samples were diluted in 10 mM glycine buffer (pH 10) supplemented with 10 mM NaCl to a final concentrations of 0.1 mg/ml (approx. 3 μM). CD spectra were collected between 190 and 250 nm in 1 nm intervals with a scan speed of 50 nm/min. Ten spectra were averaged to obtain the final CD spectrum of the respective sample. The influence of citrate on the far-UV CD spectra of CitAP-BsLA was determined by adding 1 mM sodium citrate (dissolved in 10 mM glycine buffer pH 10 supplemented with 10 mM NaCl) to the protein sample. Additionally the influence of 50 μM Triton X-100 was tested. Near-UV CD spectra were recorded from 250 nm to 370 nm using the same setup. Samples were diluted to a final concentration of 1 mg/ml (approx. 30 μM) using 10 mM glycine buffer (pH 10) supplemented with 10 mM NaCl.

Analytical ultracentrifugation (AUC). Freshly thawed CitAP-BsLA solutions at a concentration of 0.5 mg/ml dissolved in 10 mM glycine buffer pH 10 supplemented with 10 mM NaCl (±citrate) were filled into custom-produced titanium centerpieces with sapphire windows and optical pathlengths of 20 mm. Upon inserting the cells into the rotor, optical alignment along the centrifugal field is ensured by the application of a custom-made cell alignment tool (Nanolytics). Sedimentation velocity experiments were carried out on a BeckmanCoulter XL-A/XL-I Analytical Ultracentrifuge using absorbance optics (l = 275 nm) at 25 °C and an angular velocity of 40 krpm. The data were analyzed with the standard c(s) model in SEDFIT version 12.5 (https://sedfitsedphat.nibib.nih.gov/software/default.aspx) using Bayesian prior expectations for weighting the regularization. Buffer density and viscosity were calculated incrementally using Sednterp 2.0 according to the given composition. Likewise, the partial specific volume (0.734 mL/g) was calculated incrementally according to the amino acid composition. After completing a conventional c(s) analysis with uniform prior, the $c^{(P\delta)}(s)$ distributions were calculated as a secondary analysis, based on the prior expectation that the protein sample exclusively contains monodisperse species resulting in sharp peaks[55]. Two major peaks (monomer and dimer) as well as up to two minor peaks representing higher oligomers were automatically detected from an existing c(s) distribution. For each, a numerical representation of a delta-peak (width = 0.1 S) is placed at the weight-average s-value integrated across the peak. From this $c^{(P\delta)}(s)$ distribution the relative peak concentrations were calculated. Since no material outside the peaks was assigned by the $c^{(P\delta)}(s)$ distribution, the validity of the prior expectation is demonstrated. The corresponding frictional ratios (f/f_0) are related to the diffusion coefficient and were calculated from the respective sedimentation coefficient and the molecular mass of the species using the Svedberg equation. All plots of AUC raw data, best fits and residuals were created with the software GUSSI, which can be downloaded from the MBR Software Page (http://biophysics.swmed.edu/MBR/software.html). Data plots of c(s) and $c^{(P\delta)}(s)$ distributions were created by in-house developed software.

Small angle X-ray scattering (SAXS). SAXS was measured of CitAP-BsLA (0.5 to 5.0 mg/mL, 10 mM glycine buffer pH 10, 10 mM NaCl (±citrate), 10 °C sample temperature) at the beamline BM29 at the ESRF[56]. Measured data were scaled by the concentration. The excluded Porod volume was calculated with the program DATPOROD and the molecular mass was estimated by using the reported protein density of 0.588 g/mL[47]. The distance distribution function $P(r)$ was determined using the program DATGNOM. In total 20 *ab initio* models were generated using the program DAMMIF, averaged and the filtered model was used. The envelope function was determined using the SITUS package[57].

CitAP-BsLA model building and molecular dynamic (MD) simulations. The detailed strategy for modelling of the dimeric CitAP-BsLA complex is summarized in the Supplementary Materials. CitAP-BsLA monomer models were built with the program BUNCH[58] of the ATSAS package[47]. In all cases, template coordinates were taken from the PDB structures with IDs 2J80[46] (CitAP) and 1I6W[30] (BsLA). The linker connecting the CitAP and BsLA domains in the monomer, the His6 tag and all other remaining missing residues of the fusion protein were modelled as Cα traces during the fitting procedure. Afterwards, the Cα traces were extended to all-atom models with the web server MaxSprout[59]. As the all-atom extension of prolines failed, these were modelled by a superposition of a template proline residue onto the proline backbone obtained from MaxSprout. Dimer models were either built manually, by superimposing the corresponding monomer models onto the dimeric crystal structure of CitAP-PAS (PDB-ID 2J80), or were assembled *ab initio* by oligomerizing the monomer models using the program SASREF[58] optimizing the dimer orientation against SAXS data at high protein concentration (100% dimer). Further details are given in the Supplementary Material and Supplementary Table 2. The quality of all models was evaluated with the program CRYSOL[60]. CRYSOL computes theoretical scattering curves and compares these to the experimental data. As quality indicator for each model the χ values computed by CRYSOL were used, which present a measure for the discrepancy between theoretical and experimental curves. In order to improve the initial models, a 100 ns molecular dynamics (MD) simulation was performed for each dimeric assembly and a theoretical scattering curve was calculated for every 200 ps snapshot of each trajectory and fitted against the experimental data using CRYSOL. Details can be found in the Supplementary Materials.

References

1. Liang, J., Kim, J. R., Boock, J. T., Mansell, T. J. & Ostermeier, M. Ligand binding and allostery can emerge simultaneously. *Protein Sci* **16,** 929–937, doi: 10.1110/ps.062706007 (2007).
2. Ostermeier, M. & Benkovic, S. J. Evolution of protein function by domain swapping. *Advances in protein chemistry* **55,** 29–77 (2000).
3. Vogel, C., Bashton, M., Kerrison, N. D., Chothia, C. & Teichmann, S. A. Structure, function and evolution of multidomain proteins. *Current opinion in structural biology* **14,** 208–216, doi: 10.1016/j.sbi.2004.03.011 (2004).
4. de Souza, S. J. Domain shuffling and the increasing complexity of biological networks. *BioEssays: news and reviews in molecular, cellular and developmental biology* **34,** 655–657, doi: 10.1002/bies.201200006 (2012).
5. West, A. H. & Stock, A. M. Histidine kinases and response regulator proteins in two-component signaling systems. *Trends in biochemical sciences* **26,** 369–376 (2001).
6. Taylor, B. L. & Zhulin, I. B. PAS domains: internal sensors of oxygen, redox potential, and light. *Microbiology and molecular biology reviews: MMBR* **63,** 479–506 (1999).
7. Henry, J. T. & Crosson, S. Ligand-binding PAS domains in a genomic, cellular, and structural context. *Annual review of microbiology* **65,** 261–286, doi: 10.1146/annurev-micro-121809-151631 (2011).
8. Kramer, J. *et al.* Citrate sensing by the C4-dicarboxylate/citrate sensor kinase DcuS of Escherichia coli: binding site and conversion of DcuS to a C4-dicarboxylate- or citrate-specific sensor. *Journal of bacteriology* **189,** 4290–4298, doi: 10.1128/JB.00168-07 (2007).
9. Reinelt, S., Hofmann, E., Gerharz, T., Bott, M. & Madden, D. R. The structure of the periplasmic ligand-binding domain of the sensor kinase CitA reveals the first extracellular PAS domain. *J Biol Chem* **278,** 39189–39196, doi: 10.1074/jbc.M305864200 (2003).
10. Zhou, Y. F. *et al.* C-4-Dicarboxylates Sensing Mechanism Revealed by the Crystal Structures of DctB Sensor Domain. *J Mol Biol* **383,** 49–61, doi: 10.1016/j.jmb.2008.08.010 (2008).
11. Delgado-Nixon, V. M., Gonzalez, G. & Gilles-Gonzalez, M. A. Dos, a heme-binding PAS protein from Escherichia coli, is a direct oxygen sensor. *Biochemistry-Us* **39,** 2685–2691, doi: 10.1021/Bi991911s (2000).

12. Gilles-Gonzalez, M. A., Ditta, G. S. & Helinski, D. R. A haemoprotein with kinase activity encoded by the oxygen sensor of Rhizobium meliloti. *Nature* **350,** 170–172, doi: 10.1038/350170a0 (1991).
13. Key, J., Hefti, M., Purcell, E. B. & Moffat, K. Structure of the redox sensor domain of Azotobacter vinelandii NifL at atomic resolution: Signaling, dimerization, and mechanism. *Biochemistry-Us* **46,** 3614–3623, doi: 10.1021/bi0620407 (2007).
14. Taylor, B. L. Aer on the inside looking out: paradigm for a PAS-HAMP role in sensing oxygen, redox and energy. *Molecular microbiology* **65,** 1415–1424, doi: 10.1111/j.1365-2958.2007.05889.x (2007).
15. Christie, J. M. *et al.* Arabidopsis NPH1: a flavoprotein with the properties of a photoreceptor for phototropism. *Science* **282,** 1698–1701 (1998).
16. Pellequer, J. L., Wager-Smith, K. A., Kay, S. A. & Getzoff, E. D. Photoactive yellow protein: A structural prototype for the three-dimensional fold of the PAS domain superfamily. *P Natl Acad Sci USA* **95,** 5884–5890, doi: 10.1073/pnas.95.11.5884 (1998).
17. Zoltowski, B. D. *et al.* Conformational switching in the fungal light sensor Vivid. *Science* **316,** 1054–1057, doi: 10.1126/science.1137128 (2007).
18. Lee, J. *et al.* Surface sites for engineering allosteric control in proteins. *Science* **322,** 438–442, doi: 10.1126/science.1159052 (2008).
19. Strickland, D., Moffat, K. & Sosnick, T. R. Light-activated DNA binding in a designed allosteric protein. *Proc Natl Acad Sci USA* **105,** 10709–10714, doi: 10.1073/pnas.0709610105 (2008).
20. Wu, Y. I. *et al.* A genetically encoded photoactivatable Rac controls the motility of living cells. *Nature* **461,** 104–108, doi: 10.1038/nature08241 (2009).
21. Gasser, C. *et al.* Engineering of a red-light-activated human cAMP/cGMP-specific phosphodiesterase. *Proc Natl Acad Sci USA* **111,** 8803–8808, doi: 10.1073/pnas.1321600111 (2014).
22. Guntas, G., Mitchell, S. F. & Ostermeier, M. A molecular switch created by *in vitro* recombination of nonhomologous genes. *Chemistry & biology* **11,** 1483–1487, doi: 10.1016/j.chembiol.2004.08.020 (2004).
23. Guntas, G. & Ostermeier, M. Creation of an allosteric enzyme by domain insertion. *J Mol Biol* **336,** 263–273 (2004).
24. Möglich, A., Ayers, R. A. & Moffat, K. Design and signaling mechanism of light-regulated histidine kinases. *J Mol Biol* **385,** 1433–1444, doi: 10.1016/j.jmb.2008.12.017 (2009).
25. Möglich, A., Ayers, R. A. & Moffat, K. Addition at the molecular level: signal integration in designed Per-ARNT-Sim receptor proteins. *J Mol Biol* **400,** 477–486, doi: 10.1016/j.jmb.2010.05.019 (2010).
26. Guntas, G., Mansell, T. J., Kim, J. R. & Ostermeier, M. Directed evolution of protein switches and their application to the creation of ligand-binding proteins. *P Natl Acad Sci USA* **102,** 11224–11229, doi: 10.1073/pnas.0502673102 (2005).
27. Lockless, S. W. & Ranganathan, R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* **286,** 295–299 (1999).
28. Marks, D. S., Hopf, T. A. & Sander, C. Protein structure prediction from sequence variation. *Nature biotechnology* **30,** 1072–1080, doi: 10.1038/nbt.2419 (2012).
29. Suel, G. M., Lockless, S. W., Wall, M. A. & Ranganathan, R. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nature structural biology* **10,** 59–69, doi: 10.1038/nsb881 (2003).
30. van Pouderoyen, G., Eggert, T., Jaeger, K. E. & Dijkstra, B. W. The crystal structure of Bacillus subtilis lipase: a minimal alpha/beta hydrolase fold enzyme. *J Mol Biol* **309,** 215–226 (2001).
31. Marks, D. S. *et al.* Protein 3D structure computed from evolutionary sequence variation. *PloS one* **6,** e28766, doi: 10.1371/journal.pone.0028766 (2011).
32. Morcos, F. *et al.* Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci USA* **108,** E1293–1301, doi: 10.1073/pnas.1111471108 (2011).
33. Fulton, A. *et al.* Exploring the Protein Stability Landscape: Bacillus subtilis Lipase A as a Model for Detergent Tolerance. *Chembiochem* **16,** 930–936, doi: 10.1002/cbic.201402664 (2015).
34. Frauenkron-Machedjou, V. J. *et al.* Towards Understanding Directed Evolution: More than Half of All Amino Acid Positions Contribute to Ionic Liquid Resistance of Bacillus subtilis Lipase A. *Chembiochem* **16,** 937–945, doi: 10.1002/cbic.201402682 (2015).
35. Möglich, A. & Moffat, K. Structural basis for light-dependent signaling in the dimeric LOV domain of the photosensor YtvA. *J Mol Biol* **373,** 112–126, doi: 10.1016/j.jmb.2007.07.039 (2007).
36. Tiller, G. E., Mueller, T. J., Dockter, M. E. & Struve, W. G. Hydrogenation of Triton X-100 Eliminates Its Fluorescence and Ultraviolet-Light Absorption While Preserving Its Detergent Properties. *Anal Biochem* **141,** 262–266, doi: 10.1016/0003-2697(84)90455-X (1984).
37. Kaspar, S. *et al.* The periplasmic domain of the histidine autokinase CitA functions as a highly specific citrate receptor. *Molecular microbiology* **33,** 858–872, doi: 10.1046/j.1365-2958.1999.01536.x (1999).
38. Acharya, P. & Rao, N. M. Stability studies on a lipase from Bacillus subtilis in guanidinium chloride. *J Protein Chem* **22,** 51–60, doi: 10.1023/A:1023067827678 (2003).
39. Kamal, M. Z., Ali, J. & Rao, N. M. Binding of bis-ANS to Bacillus subtilis lipase: A combined computational and experimental investigation. *Bba-Proteins Proteom* **1834,** 1501–1509, doi: 10.1016/j.bbapap.2013.04.021 (2013).
40. Hawe, A., Sutter, M. & Jiskoot, W. Extrinsic fluorescent dyes as tools for protein characterization. *Pharm Res-Dordr* **25,** 1487–1499, doi: 10.1007/s11095-007-9516-9 (2008).
41. Dago, A. E. *et al.* Structural basis of histidine kinase autophosphorylation deduced by integrating genomics, molecular dynamics, and mutagenesis. *Proc Natl Acad Sci USA* **109,** E1733–1742, doi: 10.1073/pnas.1201301109 (2012).
42. Diensthuber, R. P., Bommer, M., Gleichmann, T. & Möglich, A. Full-length structure of a sensor histidine kinase pinpoints coaxial coiled coils as signal transducers and modulators. *Structure* **21,** 1127–1136, doi: 10.1016/j.str.2013.04.024 (2013).
43. Ferris, H. U. *et al.* Mechanism of regulation of receptor histidine kinases. *Structure* **20,** 56–66, doi: 10.1016/j.str.2011.11.014 (2012).
44. Hulko, M. *et al.* The HAMP domain structure implies helix rotation in transmembrane signaling. *Cell* **126,** 929–940, doi: 10.1016/j.cell.2006.06.058 (2006).
45. Matthews, E. E., Zoonens, M. & Engelman, D. M. Dynamic helix interactions in transmembrane signaling. *Cell* **127,** 447–450, doi: 10.1016/j.cell.2006.10.016 (2006).
46. Sevvana, M. *et al.* A ligand-induced switch in the periplasmic domain of sensor histidine kinase CitA. *J Mol Biol* **377,** 512–523, doi: 10.1016/j.jmb.2008.01.024 (2008).
47. Petoukhov, M. V. *et al.* New developments in the ATSAS program package for small-angle scattering data analysis. *J Appl Crystallogr* **45,** 342–350, doi: 10.1107/S0021889812007662 (2012).
48. Novinec, M. *et al.* A novel allosteric mechanism in the cysteine peptidase cathepsin K discovered by computational methods. *Nature communications* **5,** 3287, doi: 10.1038/ncomms4287 (2014).
49. McLaughlin, R. N., Poelwijk, F. J., Raman, A., Gosal, W. S. & Ranganathan, R. The spatial architecture of protein function and adaptation. *Nature* **491,** 138–U163, doi: 10.1038/nature11500 (2012).
50. Bott, M., Meyer, M. & Dimroth, P. Regulation of anaerobic citrate metabolism in Klebsiella pneumoniae. *Molecular microbiology* **18,** 533–546 (1995).
51. Yang, X., Kuk, J. & Moffat, K. Crystal structure of Pseudomonas aeruginosa bacteriophytochrome: photoconversion and signal transduction. *Proc Natl Acad Sci USA* **105,** 14715–14720, doi: 10.1073/pnas.0806718105 (2008).
52. Eddy, S. R. Accelerated Profile HMM Searches. *Plos Comput Biol* **7,** doi: ARTNe100219510.1371/journal.pcbi.1002195 (2011).
53. Apweiler, R. *et al.* Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res* **41,** D43–D47, doi: 10.1093/nar/gks1068 (2013).

54. Ekeberg, M., Lovkvist, C., Lan, Y. H., Weigt, M. & Aurell, E. Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Phys Rev E* **87,** doi: Artn01270710.1103/Physreve.87.012707 (2013).
55. Brown, P. H., Balbo, A. & Schuck, P. Using prior knowledge in the determination of macromolecular size-distributions by analytical ultracentrifugation. *Biomacromolecules* **8,** 2011–2024, doi: 10.1021/bm070193j (2007).
56. Pernot, P. *et al.* Upgraded ESRF BM29 beamline for SAXS on macromolecules in solution. *J Synchrotron Radiat* **20,** 660–664, doi: 10.1107/S0909049513010431 (2013).
57. Wriggers, W. Conventions and workflows for using Situs. *Acta Crystallogr D* **68,** 344–351, doi: 10.1107/S0907444911049791 (2012).
58. Petoukhov, M. V. & Svergun, D. I. Global rigid body modeling of macromolecular complexes against small-angle scattering data. *Biophysical journal* **89,** 1237–1250, doi: 10.1529/biophysj.105.064154 (2005).
59. Holm, L. & Sander, C. Database algorithm for generating protein backbone and side-chain co-ordinates from a C alpha trace application to model building and detection of co-ordinate errors. *J Mol Biol* **218,** 183–194 (1991).
60. Svergun, D., Barberato, C. & Koch, M. H. J. CRYSOL - A program to evaluate x-ray solution scattering of biological macromolecules from atomic coordinates. *J Appl Crystallogr* **28,** 768–773, doi: 10.1107/S0021889895007047 (1995).

Acknowledgements

The excellent technical assistance of Doreen Freund (Nanolytics) is acknowledged. We acknowledge the beamline scientists Drs. Adam Round and Martha Brennich for help during the SAXS experiments at the beamline BM29 at the ESRF (Grenoble, France). We gratefully acknowledge the computing time granted by the JARA-HPC Vergabegremium and provided on the JARA-HPC partition part of the RWTH Bull Cluster in Aachen (Grant Number JARA0122). U.K., M.K. and K.-E.J. acknowledge support provided by the International NRW Research School BioStruct, granted by the Ministry of Innovation, Science and Research of the State North Rhine-Westphalia, the HHU Düsseldorf, and the Entrepreneur Foundation of the HHU of Düsseldorf. U.K., T.F., and K.-E.J. also acknowledge funding by the Federal Ministry of Education and Research (BMBF) in the framework of the collaborative research project “OptoSys” (FKZ 031A16).

Author Contributions

M.K. generated fusion constructs, carried out all biochemical characterizations of CitAP-BsLA and analyzed the biochemial data; O.S. performed molecular dynamics simulations and analyzed the data, overseen by B.S.; T.F. and C.N. contributed to the biochemical characterization of CitAP-BsLA; F.K. carried out and analyzed AUC experiments; A.F. constructed and analyzed BsLA mutant libraries; A.S. performed SAXS measurements and analyzed the data; U.K. conceived the study. B.S., K-E.J. and U.K. coordinated and oversaw the project; M.K., O.S., A.S., B.S. and U.K. wrote the paper. All authors discussed the results and commented on the manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Kaschner, M. *et al.* A combination of mutational and computational scanning guides the design of an artificial ligand-binding controlled lipase. *Sci. Rep.* **7**, 42592; doi: 10.1038/srep42592 (2017).

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

3.2 Conformational Polymorphism in Autophagy-Related Protein GATE-16

3.2.1 Summary

The autophagy-related protein GATE-16 (section 2.3.2) was studied by X-ray crystallography (section 2.1.1), NMR spectroscopy (section 2.1.2) and MD simulations (section 2.2.1).

Different X-ray models of GATE-16 show different conformational states of the C-terminus. In addition, NMR S^2 order parameters also indicate a flexible C-terminus. It is known that the C-terminus is crucial for GATE-16 processing. It is proteolytically cleaved by ATG4 family proteases and subsequently lipidated to enable membrane tethering.

The presented research attempts to answer the questions: What is the nature of GATE-16's C-terminal dynamics and how is it important for its function?

We discovered that Hamiltonian replica exchange MD simulations (HREMD, section 2.2.3) sample multiple C-terminal states. Structural indicators, such as hydrophobic contacts, side chain orientations and salt-bridges were used to classify the C-terminal conformations into discrete states. From their populations a three-state model of GATE-16's C-terminal dynamics is proposed:

1. Newly synthesized GATE-16 is in a closed state.

2. The C-terminus of GATE-16 exists in a conformational equilibrium between multiple states. It assumes an extended, detached conformation about 15% of the time.

3. When GATE-16 is bound to ATG4 family proteases, a completely extended state of the C-terminus is stabilized by a loop protruding from the enzyme. It shields the hydrophobic groove of GATE-16 formerly accommodating Phe115 or Phe117 and, at the same time, sterically prevents the C-terminus from reverting to the closed state. Finally, proteolytic cleavage of the C-terminus and lipidation can take place.

3.2.2 Contribution

I performed the MD simulations, including Hamiltonian Replica Exchange MD, analysed the MD data, computed the S^2 order parameters and identified the dominant structural states of GATE-16.

I wrote the manuscript sections on MD simulation methods and MD data analysis. Furthermore, I produced Figure 3 and Table 2. In total, I contributed approximately 40 % of the complete manuscript.

3.2.3 Publication

This section contains a complete reprint of the publication [97]. The supporting information to this article is located in section 6.2.

Reprinted with permission. Copyright 2015 American Chemical Society.

BIOCHEMISTRY
including biophysical chemistry & molecular biology

Article

pubs.acs.org/biochemistry

Conformational Polymorphism in Autophagy-Related Protein GATE-16

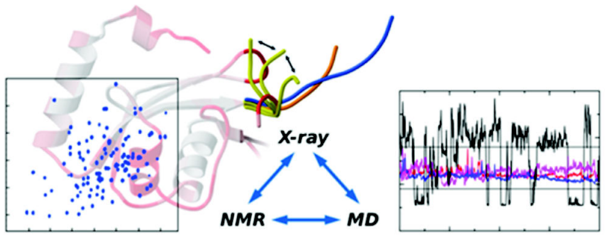
Peixiang Ma,[†,‡,§] Oliver Schillinger,[†] Melanie Schwarten,[†] Justin Lecher,[†] Rudolf Hartmann,[†] Matthias Stoldt,[†,‡] Jeannine Mohrlüder,[†] Olujide Olubiyi,[†,∥] Birgit Strodel,[†] Dieter Willbold,[†,‡] and Oliver H. Weiergräber*,[†]

[†]Institute of Complex Systems, ICS-6 (Structural Biochemistry), Forschungszentrum Jülich, 52425 Jülich, Germany

[‡]Institut für Physikalische Biologie und BMFZ, Heinrich-Heine-Universität Düsseldorf, 40225 Düsseldorf, Germany

Supporting Information

ABSTRACT: Autophagy is a fundamental homeostatic process in eukaryotic organisms, fulfilling essential roles in development and adaptation to stress. Among other factors, formation of autophagosomes critically depends on proteins of the Atg8 (autophagy-related protein 8) family, which are reversibly conjugated to membrane lipids. We have applied X-ray crystallography, nuclear magnetic resonance spectroscopy, and molecular dynamics simulations to study the conformational dynamics of Atg8-type proteins, using GATE-16 (Golgi-associated ATPase enhancer of 16 kDa), also known as GABARAPL2, as a model system. This combination of complementary approaches provides new insight into a structural transition centered on the C-terminus, which is crucial for the biological activity of these proteins.



Autophagy is an evolutionarily conserved protein and organelle degradation system in eukaryotic cells, which plays important roles in cellular homeostasis, differentiation, and stress response. In fact, dysregulation of autophagy has been implicated in widespread diseases like cancer and neurodegenerative disorders, as well as in aging. The core autophagy machinery in yeast requires some 15 autophagy-related (Atg) proteins to accomplish autophagosome formation and maturation.[1] Atg8 is one of these, playing a vital role in the specific recruitment of cargo proteins destined for autophagic degradation and promoting autophagosome maturation. During eukaryotic evolution, the primordial Atg8-encoding gene underwent repeated duplication to yield eight human homologues; on the basis of amino acid sequence similarity, the corresponding proteins have been classified into two groups: the microtubule-associated protein 1 light chain 3 (MAP1LC3 or LC3) subfamily comprises LC3A, LC3B, LC3B2, and LC3C, whereas the γ-aminobutyric acid (GABA)$_A$ receptor-associated protein (GABARAP) subfamily contains GABARAP itself together with several GABARAP-like (GABARAPL) proteins: GABARAPL1 (GEC1, glandular epithelial cell protein 1), GABARAPL2 (GATE-16, Golgi-associated ATPase enhancer of 16 kDa), and GABARAPL3, which corresponds to a putative pseudogene.[2,3] Note that GATE-16 is sometimes considered a subfamily of its own because of its divergence from other GABARAP-like proteins. The biological relevance of the expansion of the Atg8 family in higher eukaryotes is poorly understood; however, recent evidence indicates that subfamilies are involved in different steps along the autophagic process. LC3s are essential for elongation of the phagophore membrane, whereas the GABARAP/GATE-16 subfamily engages in a later stage of autophagosome maturation.[4] Despite these functional differences, all homologues undergo the same post-translational modification steps: newly synthesized human Atg8 family proteins are cleaved at their C-termini by ATG4 proteases, yielding a product termed form I, which exposes a conserved terminal glycine residue. The E1-like enzyme ATG7 forms thioester intermediates with Atg8 homologues, which are transferred to the E2-like component ATG3 and finally conjugated to phosphatidylethanolamine (PE), resulting in form II anchored to the autophagosome membrane.[5]

Over the past 15 years, our laboratory has been investigating the three-dimensional structures of Atg8-related proteins[6,7] and their complexes with model ligands as well as cellular binding partners.[8−11] Together with data published by others, our structural studies revealed very similar tertiary folds, comprising a C-terminal ubiquitin-like domain (a β-grasp fold) that is preceded by an N-terminal helical extension. On the other hand, human Atg8 proteins display some intriguing differences with regard to patterns of surface-exposed side chains. For example, the first α-helix in LC3 is basic, whereas in GATE-16 and GABARAP, it has an acidic nature. The surface of the second α-helix is acidic in LC3, neutral in GATE-16, and basic in GABARAP. Whereas conservation of certain features across

Received: April 8, 2015
Revised: July 6, 2015
Published: August 18, 2015

DOI: 10.1021/acs.biochem.5b00366

subfamilies is likely to reflect common interactions within the core autophagy machinery, such as binding to the conjugation enzymes, structural differences between various Atg8 proteins are expected to confer specificity toward target proteins, indicating functional diversification. In many Protein Data Bank (PDB) entries for Atg8-related proteins, the C-terminal tail contacts the body of the β-grasp domain, giving rise to a more or less "closed" conformation. During the C-terminal modification process, however, this segment will need to adopt a more extended fold. Indeed, the X-ray structure of the ATG4B–LC3B complex[12] revealed that the LC3B C-terminus points away from the core of the molecule, inserting into the catalytic center of ATG4B for cleavage. Similarly, the lipidated (membrane-attached) state of Atg8 proteins is probably incompatible with a fully closed conformation. In line with this assumption, we could show that residues spatially close to the C-terminus display the largest nuclear magnetic resonance (NMR) chemical shift differences after attachment to a membrane.[13] Detailed knowledge of the dynamics of these proteins will therefore be crucial for understanding the regulation of their C-terminal modification. With molecular dynamics (MD) simulations, the time-dependent behavior and dynamics of a molecular system can be studied *in silico*, using conditions closely mirroring the physiological state. Detailed information about the conformational changes characterizing a protein's folding and association landscapes as well as the thermodynamics and kinetics of macromolecular processes can be obtained by this technique. With recent hardware developments and algorithmic improvements, simulations on the nanosecond to microsecond time scale have become feasible for solvated systems, further enhancing the relevance of this method for the study of macromolecular structures.

In this work, we focus on human GATE-16 as a model to investigate the conformational polymorphism of Atg8 proteins. We obtained crystals under close-to-native conditions without using conventional precipitants, and the respective X-ray structure revealed a conformation at the C-terminus different from that previously observed. Protein dynamics were investigated using solution NMR spectroscopy and MD simulations. We describe conformational transitions between open and closed states of the C-terminal segment. These findings are discussed with respect to the biological functions and the life cycle of Atg8 family proteins in the cell.

EXPERIMENTAL PROCEDURES

Expression and Purification of GATE-16. A cDNA encoding full-length human GATE-16 (UniProt entry P60520) was amplified by polymerase chain reaction from a human cDNA library and cloned into vector pGEX-4T-2 (GE Healthcare) using BamHI and NotI restriction sites. The integrity of the construct was verified by DNA sequencing. *Escherichia coli* C43 cells were transformed with the plasmid DNA and grown at 37 °C in appropriate media; protein expression for crystallization was performed in Luria broth, whereas ^{13}C- and ^{15}N-labeled material was produced in M9 medium supplemented with [^{13}C]glucose and [^{15}N]ammonium chloride. In all cases, expression was induced at an OD$_{600}$ of 0.6–0.8 by adding 1 mM isopropyl β-D-thiogalactopyranoside, and the cells were further grown overnight. The glutathione *S*-transferase–GATE-16 fusion protein was purified from the soluble cell extract by affinity chromatography on glutathione sepharose 4B (Amersham Biosciences). Thrombin (Merck) cleavage yielded full-length GATE-16 with additional glycine and serine residues at the N-terminus. For final purification, the sample was applied to a Superdex 75 prep-grade size exclusion chromatography column (Amersham Biosciences), and the GATE-16-containing fractions were pooled and concentrated by ultrafiltration.

Protein Crystallization and X-ray Data Collection. Purified human GATE-16 was observed to crystallize from a concentrated protein solution (approximately 400 μM) at 4 °C without addition of conventional precipitants. This condition was subjected to optimization using vapor diffusion experiments in a hanging drop setup at 20 °C. Well-diffracting crystals were obtained with a reservoir buffer containing 100 mM Na_2HPO_4/NaH_2PO_4 (pH 7.0), 50 mM KCl, and 10 mM dithiothreitol (DTT), and a protein concentration of 370 μM. X-ray diffraction experiments were performed at 100 K. Prior to being cryocooled, crystals were soaked in reservoir buffer containing 35% (v/v) glycerol for 10 min. Native data were recorded on beamline ID14-4 of the European Synchrotron Radiation Facility (ESRF, Grenoble, France) tuned to a wavelength of 1.009 Å on an ADSC Q315r detector. Data processing, including reflections up to 2.0 Å resolution, was conducted using MOSFLM[14] as well as SCALA[15] and TRUNCATE,[16] which are part of the CCP4[17] software suite.

Structure Determination. Crystals of GATE-16 belonged to space group $P2_1$, with a monoclinic angle equal to 90°. Initial phases were obtained by molecular replacement using MOLREP[18] with a single native diffraction data set and a search model derived from a previous GATE-16 structure (PDB entry 1EO6[19]). The unit cell was found to contain two copies of the molecule per asymmetric unit, corresponding to a Matthews coefficient of 2.1 Å^3/Da and a solvent content of 40%. For improvement of the model, reciprocal and real space refinement in PHENIX[20] was alternated with manual rebuilding in COOT.[21] Because intensity statistics indicated the presence of nearly perfect pseudomerohedral twinning, refinement was conducted using a twinned target function. Statistics on data collection and refinement are listed in Table 1. The refined model contains amino acids 1–117 of GATE-16 along with part of the N-terminal cloning artifact. Validation with MOLPROBITY[22] and COOT revealed good geometry with all of the residues in the allowed regions of the Ramachandran plot and no rotamer outliers. Atomic coordinates and structure factor amplitudes have been deposited as PDB entry 4CO7.

NMR Spectroscopy. NMR experiments were performed on a Varian 600, 800, or 900 MHz spectrometer equipped with triple-resonance (^{1}H, ^{13}C, ^{15}N) cryoprobes and shielded z-gradients. The data were collected at 25 °C at a sample concentration of 230 or 410 μM in 25 mM sodium phosphate (pH 6.5), 50 mM KCl, 10 mM DTT, 50 μM ethylenediaminetetraacetic acid, and 5% (v/v) ^2H_2O. Chemical shifts were referenced with 2,2-dimethyl-2-silapentane-5-sulfonate. All NMR spectra were processed with NMRPipe[23] and analyzed with CcpNmr.[24] Complete sequential backbone resonance assignment was achieved by two-dimensional (2D) (^{1}H–^{15}N)-HSQC, ct-(^{1}H–^{13}C)-HSQC, three-dimensional (3D) BEST-HNCA, HNCACB, HNcoCA, HNCO, HNcaCO, HBHAcoNH, (^{1}H–^{1}H–^{15}N)-NOESY-HSQC, and H(C)CH–COSY experiments (Figure S1). GATE-16 backbone assignments have been deposited at the Biological Magnetic Resonance Data Bank (entry 18827).

Relaxation Analysis. R_1, R_2, and heteronuclear (het) {^{1}H}–^{15}N nuclear Overhauser effect (NOE) relaxation experiments were conducted on 600 and 800 MHz NMR

Table 1. X-ray Data Collection and Refinement Statistics

Data Collection[a]	
space group	$P2_1$
unit cell parameters (T = 100 K)	
a, b, c (Å)	28.73, 67.44, 58.72
β (deg)	90.0
resolution range (Å)	29.36−2.00 (2.11−2.00)
beamline	ID14-4 (ESRF)
detector	ADSC Q315r
wavelength (Å)	1.009
no. of unique reflections	14406
completeness (%)	94.8 (94.8)
average multiplicity	2.2 (2.2)
R_{merge}	0.04 (0.12)
$\langle I/\sigma(I)\rangle$	14.2 (5.8)
Wilson B factor (Å^2)	23.3
Refinement	
R_{work} (%)	18.1
R_{free} (%)	21.2
twinning statistics	
operator	h, $-k$, $-l$
twinning fraction	0.49
contents of the asymmetric unit	
no. of atoms	2073
no. of protein residues	236
no. of water molecules	151
average B factor (Å^2)	
protein	23.5
water	24.8
rmsd	
bond lengths (Å)	0.005
bond angles (deg)	0.817
Ramachandran statistics (%)	
favored	97.8
allowed	2.2
outliers	0

[a] Values for the highest-resolution shell are given in parentheses.

spectrometers. The recycle delay was set to 6 s for R_1 and R_2 experiments, and the relaxation-caused magnetization decay was sampled at 10, 20, 60, 120, 250, 350, 500, 700, 900, 1200, and 1800 ms for longitudinal relaxation and at 10, 30, 50, 70, 90, 110, 130, 150, and 170 ms for transverse relaxation. For the hetNOE, the saturation and recycle delays were set to 3 and 6 s, respectively. Resonances were automatically picked and fitted using CcpNmr. Peak intensities were then fitted using an exponential decay model implemented in relax.[25,26] The hetNOE ratio was determined after fitting of the resonances and extraction of the intensities in each spectrum. The obtained relaxation rates and the heteronuclear NOEs were used as input for relax, with the coordinates of our GATE-16 crystal structure (chains A and B) serving as a reference.[25−32]

MD Simulations. For exploration of the conformational ensemble of GATE-16, we performed Hamiltonian replica exchange molecular dynamics (HREMD) simulations,[33] while constant-temperature MD simulations were subsequently run for the determination of S^2 order parameters. All simulations were performed using Gromacs, version 5.0.4 for MD simulations and version 4.6.7 combined with the PLUMED plugin (a special branch of version 2.1 that can be obtained from https://github.com/GiovanniBussi/plumed2) for the HREMD simulations. In the more common temperature REMD, one runs N copies of the system at different temperatures, and after each replica has completed a certain number of MD steps, one exchanges configurations at different temperatures based on the Metropolis criterion. The purpose of the replicas at higher temperatures is to accelerate sampling. In HREMD, different replicas evolve according to different Hamiltonians, mimicking different temperatures, which is more efficient because the number of required replicas is much smaller if the Hamiltonian of only the solute (in this case GATE-16), including solute−solvent interactions, is modified. This protocol has been tested for the Trp-cage miniprotein and a β-hairpin,[33] showing a significantly lower computational cost and better sampling compared to those of temperature replica exchange. The Amber99SB-ILDN[34] protein force field was chosen to represent GATE-16 as this force field was recently found to be one of the best protein force fields currently available.[35] In ref 35, 11 recent all-atom force fields in combination with five solvent models were quantitatively evaluated against 524 diverse NMR measurements on dipeptides, tripeptides, tetraalanine, and ubiquitin. Given that GATE-16 comprises a C-terminal ubiquitin-like domain, the performance of the force fields for ubiquitin and not so much for the short peptides was of relevance to us. In this regard, Amber99SB-ILDN together with the TIP3P water model[36] produces convincing results. Only the Amber99SB-ILDN-NMR force field performs equally well or even slightly better, especially when combined with the computationally more costly TIP4P-EW water model. However, when simulating more flexible proteins than GATE-16 and ubiquitin, we found that Amber99SB-ILDN substantially outperforms Amber99SB-ILDN-NMR, based on the comparison of NMR measurements (unpublished results). On the basis of this reasoning, we decided to use Amber99SB-ILDN combined with TIP3P to model GATE-16 in solution. The HREMD simulation was initiated from the same starting structure for all replicas; chain B of the GATE-16 structure determined in this study was selected for this purpose. The protein was centered in a dodecahedral simulation box, allowing for a 1.2 nm distance between the protein and the nearest edge of the box. The box was solvated using the TIP3P[36] explicit water model, and counterions (Na^+ and Cl^-) were added to neutralize any net protein charge as well as to achieve a salt concentration close to 100 mM (15 sodium and 16 chloride ions have been added, amounting to an ion concentration of 102.9 mM). All bonds present in the system were constrained using the LINCS[37] algorithm. Short-range nonbonded interactions were truncated at 1.2 nm, while the particle mesh Ewald method[38] was employed for treating the long-range component of electrostatic interactions in combination with periodic boundary conditions. Initially, energy minimization using the steepest descent algorithm was performed until the largest force acting on any atom decreased below 100 kJ mol^{-1} nm^{-1}. The system was then equilibrated for 5 ns in the NPT (constant particle number, pressure, and temperature) ensemble, using a velocity rescaling thermostat[39] set to 298 K and a Berendsen barostat[40] at 1 bar. During the equilibration, the heavy atoms were position-restrained with a harmonic force to relax the solvent while preventing significant protein rearrangements at this stage. The restraining force was decreased every 500 ps to allow for a gradual relaxation of intramolecular interactions to prevent strain that could lead to large unphysical dynamics if the position restraints were switched off instantaneously. The sequence of force constants was 1000, 800, 600, 400, 200, 150,

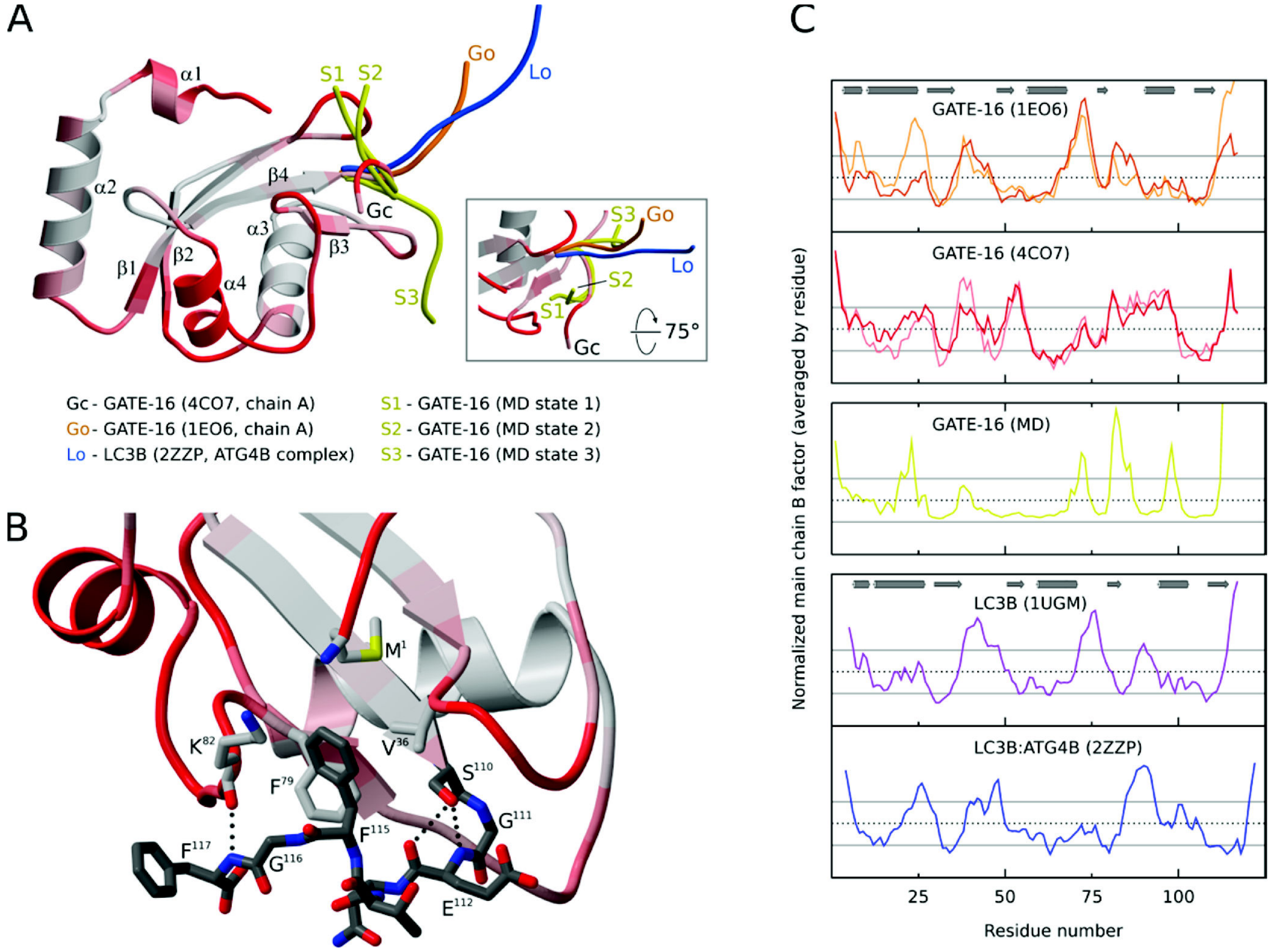

A
α1
α2
β1
β2
β3
β4
α3
α4
S1 S2
Go
Lo
Gc
S3
S3 Go
Lo
S2
S1
Gc
75°
Gc - GATE-16 (4CO7, chain A)
Go - GATE-16 (1EO6, chain A)
Lo - LC3B (2ZZP, ATG4B complex)
S1 - GATE-16 (MD state 1)
S2 - GATE-16 (MD state 2)
S3 - GATE-16 (MD state 3)
B
M1
K82
V36
F79
S110
G111
F115
F117
G116
E112
C
GATE-16 (1EO6)
GATE-16 (4CO7)
GATE-16 (MD)
LC3B (1UGM)
LC3B:ATG4B (2ZZP)
Normalized main chain B factor (averaged by residue)
25
50
75
100
Residue number

Figure 1. (A) Ribbon diagram of the GATE-16 crystal structure described in this paper, with Cα *B* factors indicated by colors ranging from white [≤mean − 1 standard deviation (SD) = 20.0 Å^2] to red (≥mean + 2SD = 28.6 Å^2). The conformation of the C-terminal tail is also depicted for representative snapshots corresponding to the three most populated states (1−3) in our HREMD simulation (residues 110−117 each, yellow), as well as for a previous GATE-16 X-ray structure (residues 110−116, orange) and the structure of LC3B in a complex with ATG4B (residues 114−122, blue). (B) Close-up view of the C-terminus of GATE-16 emphasizing its noncovalent interactions with the remainder of the molecule. (C) Normalized main chain *B* factors derived from a selection of X-ray structures of Atg8 family proteins. For the calculation of the mean and SD, residues preceding helix α1 or following strand β4 were excluded. GATE-16 structures 1EO6 and 4CO7 each comprise two chains per asymmetric unit, as signified by light (chain A) and dark coloring (chain B), respectively. In each plot, an interval of one SD about the mean *B* factor is indicated. See the text for details.

100, 70, 30, and 10 kJ mol^{-1} nm^{-1}. The final state of this equilibration procedure was the starting structure of the HREMD simulation. Each replica was simulated in the *NPT* ensemble at 298 K and 1 bar using a Nose−Hoover thermostat[41] and a Parrinello−Rahman barostat.[42]

We simulated 20 replicas for which the Hamiltonian scaling factors were optimized in preceding test runs to achieve good exchange rates between neighboring replicas. The resulting scaling factors were exponentially distributed between 1.00 and 0.66 (exact scaling factors of 1.00, 0.98, 0.96, 0.94, 0.91, 0.89, 0.88, 0.86, 0.84, 0.82, 0.80, 0.78, 0.77, 0.75, 0.73, 0.72, 0.70, 0.69, 0.67, and 0.66). These scaling factors correspond to temperatures between 298 and 457 K. Care has been taken that the minimal scaling factor corresponds to a temperature below the folding temperature of GATE-16. For temperatures of ≥500 K, we observed complete unfolding of the protein, which we did not wish to study here. An exchange between neighboring replicas was attempted every 5 ps. The average acceptance ratio for exchanges was 0.46. The total simulation time per replica was 500 ns, amounting to an accumulated HREMD simulation time of 10 μs. Coordinates and energies were saved every 10 ps for each replica.

For analysis of the HREMD simulations, only structures sampled for the target replica with a scaling factor of 1.0 (i.e., for the unmodified Hamiltonian) were considered. The structures were clustered according to the following criteria. (1) Phe115 is considered to be in the hydrophobic groove formed by Met1, Val36, Phe79, and the aliphatic portion of Lys82 if at least 8 of the 14 side chain atoms of Phe115 are in contact with these four residues. Here, a contact is defined on the basis of the minimal distance between each of Phe115's side chain atoms and any of the atoms of the other four hydrophobic residues. If this minimal distance is ≤4 Å, a contact is present. Multiple contacts between a given side chain atom of Phe115 and the hydrophobic groove were counted as one; thus, the maximal number of hydrophobic contacts is 14. (2) Phe117 is considered to reside in the hydrophobic groove using the same definition as for Phe115. (3) When residing in the groove, Phe115 can point into two different directions. We distinguish these directions on the basis of the minimal distance between the Cζ atom of Phe115 and any of the Trp3 atoms. If this distance is ≤7.2 Å, Phe115 is oriented toward the N-terminus; otherwise, it points away. (4) The existence of a salt bridge between the N- and C-termini is evaluated via the distance between the N atom of Met1 and the closest of the two carboxylate oxygens of Phe117. If this distance is ≤4 Å, the salt bridge is considered to be formed; otherwise, it is not. The selection and definition of these criteria were made on the basis of histograms we computed for the corresponding observables of the target replica (see Figure S3). We also determined the ensemble-averaged structural fluctuations in terms of *B* factors for the structures of the target replica. The *B* factors were computed by averaging the mean square positional fluctuations of the main chain atoms of each residue, after the replica structures had been superimposed on the starting structure of the HREMD simulation.

For the calculation of NMR order parameters, we performed regular MD simulations at 298 K for 20 ns starting from representative structures of the four dominant states obtained from the HREMD simulation. Here, the number of simulation runs per state was chosen to represent the relative thermodynamic weight of the state in question (i.e., the state population in the target replica). Thus, we ran 15, 6, 5, and 3 simulations for states 1−4, respectively, amounting to 29 × 20 ns = 580 ns of additional simulation time. The MD simulation conditions were identical to those of the target replica from the HREMD simulation. All simulations for a given state started from the same representative structure, though the initial velocities were individually generated from a Maxwell−Boltzmann distribution at 298 K. To calculate the squared order parameter S^2 from these 29 simulations, time correlation functions of the N−H bond vectors were computed according to[31]

$$C(\tau) = \langle P_2[\vec{\mu}(t)\cdot\vec{\mu}(t+\tau)]\rangle$$

where $P_2(x)$ is the second Legendre polynomial, $\vec{\mu}$ describes the N−H bond unit vector, and angular brackets indicate averaging over time *t*. The order parameter is then defined as

$$S^2 = C_\infty = \lim_{\tau\to\infty} C(\tau)$$

Before the calculation of S^2 from the MD trajectories, the overall rotational motion of the protein was removed by superposing all structures of each 20 ns trajectory to the initial conformation of that trajectory. Here, it should be noted that the trajectory length of 20 ns is well above the rotational correlation time of GATE-16, which was determined to be 7.43 ns from our NMR experiments. Furthermore, the TIP3P water model used in the MD simulations is known to underestimate the shear viscosity of real water by a factor of 2.18 ± 0.07.[43] As the rotational correlation time is directly proportional to the viscosity of the solvent, the rotational correlation time of GATE-16 in our simulations is correspondingly smaller. On the other hand, a trajectory length of 20 ns is short enough not to overrate long time internal fluctuations compared to motions faster than tumbling, which would otherwise lead to an underestimation of the predicted S^2 values.[44] S^2 order parameters per residue were then computed as the average of the convergence values of the respective time correlation functions. Here, all nonconverged correlation functions were discarded, and convergence was judged by comparing the mean values and standard deviations of the third and fourth quarters of the correlation functions. As convergence criteria, a difference in means of less than 0.02 and a standard deviation of ≤0.02 were applied. The average of the mean values of the fourth quarters of $C(t)$ was then used as an estimate for C_∞.

Molecular Graphics. Representations of the GATE-16 structure determined in this study are based upon the coordinates of chain A. Panels A and B of Figure 1 were generated with MOLSCRIPT[45] and RASTER3D[46] using secondary structure assignments provided by DSSP.[47] Figure 3 was generated with Matplotlib,[48] Seaborn,[49] and VMD.[50]

RESULTS

The three-dimensional structure of human GATE-16 was determined by X-ray crystallography. Crystals of the protein were observed to form in a concentrated solution at near-physiological ionic strength and in the absence of typical precipitating agents (see Experimental Procedures). While the initial evaluation of diffraction data extending to 2.0 Å resolution suggested Laue group *mmm*, structure determination was unsuccessful in all candidate orthorhombic space groups. Instead, the true space group was found to be $P2_1$, with near-perfect pseudomerohedral twinning accounting for additional symmetry in the diffraction data.

Overall, the three-dimensional fold of GATE-16 in this crystal form is consistent with structures of Atg8 family proteins determined previously; in addition to a β-grasp fold, which is characteristic of ubiquitin superfamily proteins, it comprises an N-terminal helical extension attached to the central β-sheet. The two copies of the protein chain related by noncrystallographic symmetry (NCS) turned out to be very similar, with a root-mean-square distance of 0.37 Å for Cα atoms and 0.79 Å for all non-hydrogen atoms. Figure 1A shows this GATE-16 structure in ribbon representation, with *B* factors of Cα atoms encoded by a color ramp. In general, the structure is well-ordered; elevated *B* factors indicating enhanced flexibility are observed at the N- and C-termini, as well as in the $\beta1-\beta2$, $\beta2-\alpha3$, and $\beta3-\alpha4$ loops and the $\alpha4$ helix. These segments are mostly located on one side of the molecule, distant from the well-established hydrophobic patches accounting for most of the macromolecular interactions described for this protein family. A comparison with other crystal structures of Atg8 homologues revealed that the C-terminal residues following strand $\beta4$ have been found in different conformations. In the original GATE-16 structure,[19] for instance, chain A adopts an extended conformation; i.e., the C-terminus points away from the body of the molecule (Figure 1A, orange), whereas in chain B it is positioned closer to the β-grasp domain (not shown). Notably, our crystal structure displays an even more compact fold, in which the C-terminal segment is held in place by numerous polar and hydrophobic interactions (Figure 1B). Specifically, the approximately 90° bend following $\beta4$ and centered on Gly111 is stabilized by hydrogen bonds between the side chain of Ser110 and main chain atoms of Glu112; Phe115 is inserted into a hydrophobic groove formed by Met1, Val36, Phe79, and the aliphatic portion of Lys82, and finally, Phe117 forms a main chain hydrogen bond with Lys82. The other extremity of the spectrum is exemplified by the structure of LC3B in complex with the cysteine protease ATG4B.[12] Here, the C-terminus is extended completely, pointing toward the active site of the enzyme (Figure 1A, blue).

In X-ray crystallography, *B* factors are used to parametrize the displacement of atoms from their mean positions using an isotropic model.[51] The major sources of such displacement are

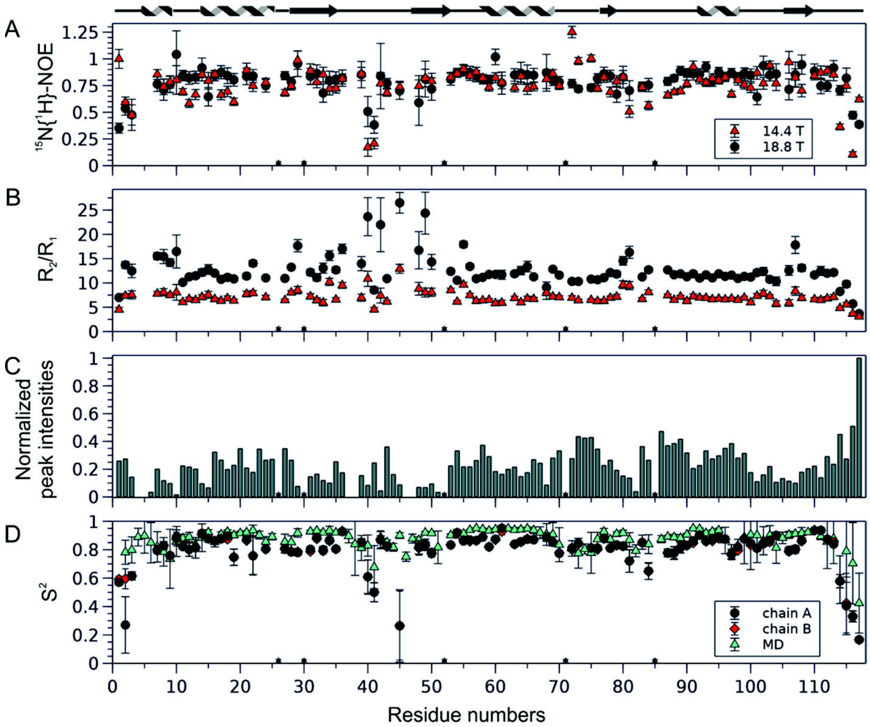


Figure 2. Relaxation data for backbone amide ^{15}N nuclei of GATE-16, {^{1}H}−^{15}N NOE (A) and R_2/R_1 (B), at 14.1 T (red triangles) and 18.8 T (black circles). Normalized peak intensities from HNCO experiments are plotted vs residue number (C). Comparison of order parameters (D): S^2 derived from model-free analysis of the relaxation data based on chain A (black circles) or B (red diamonds) and S^2 calculated from MD trajectories (turquoise triangles). Secondary structure elements are depicted above the plots; asterisks indicate proline residues.

thermal motion and (mostly short-range) disorder in the lattice, which are expected to correlate with flexibility in solution.[52] We therefore investigated the distribution of *B* factors along the polypeptide chains of selected Atg8 family proteins (Figure 1C). While these plots reveal a number of shared features, including high-*B* regions as outlined above for GATE-16, significant differences do exist. Generally, these may either reflect intrinsic properties of individual homologues or result from different lattice contacts imposing restraints on the mobility of certain parts of the molecules. This ambiguity can be addressed if alternative sources of structural information are available, such as additional crystal structures of the same proteins with different packing interactions. For instance, the large difference in *B* factors of the $\alpha3-\beta3$ loop observed between the original GATE-16 structure and the one described in this work is readily explained by a lack of lattice interactions in the former, whereas we found this loop to be hydrogen-bonded to a similar region of an NCS-related molecule.

While crystallographic *B* factors can be interpreted as indicators of flexibility in the context of the crystal lattice, direct experimental evidence of the behavior in solution can be obtained by NMR spectroscopy. For this purpose, protein backbone mobility was investigated by measuring the two ^{15}N spin relaxation rates R_1 and R_2, as well as heteronuclear ^{1}H−^{15}N NOEs at 298 K (Figure 2A,B). A complete mobility analysis of GATE-16 was performed on the basis of the Lipari−Szabo approach using a model-free analysis implemented in relax. R_1, R_2, and heteronuclear NOE data could be fitted to the two chains of our GATE-16 crystal structure by assuming an anisotropy ellipsoid as a diffusion model with a correlation time τ_c of 7.45 ± 0.01 ns for protein chain A and 7.41 ± 0.01 ns for chain B. Thus, the correlation time calculated from relaxation data by relax is slightly shorter compared to the value computed by HYDRONMR[53] based on our crystal structure (7.66 and 7.70 ns for chains A and B, respectively). This implies a monomeric state of GATE-16 in solution, at the concentration used in our NMR experiments. In contrast, using the coordinates of the previously published GATE-16 structure (PDB entry 1EO6), HYDRONMR predicts correlation times of 8.14 and 7.83 ns for the two chains, which is even larger than what we observe. For the relaxation data, the mobility analysis yields the square of the generalized order parameter (S^2), which is shown in Figure 2D. For the majority of residues in GATE-16, the near-unity S^2 values indicate that the molecule is globally rigid and displays restrained internal motion on the pico- to nanosecond time scale. On the other hand, lower order parameters were observed for the N- and C-termini, as well as two loop regions (residues 40, 41, and 45 in the $\beta1-\beta2$ loop and residues 81 and 84 in the $\beta3-\alpha4$ loop). For the N-terminal segment, flexibility is confirmed by the deviation between the two chains in the crystal asymmetric unit, and the weak or undetectable NH signals in NMR (Figure 2C). Notably, our X-ray structure reveals only minor differences between chains in the C-terminal region, because of restraints imposed by very similar lattice contacts, yet such differences are very pronounced in PDB entry 1EO6, which exhibits quite different packing environments for the two chains. Residues 37−40 in the $\beta1-\beta2$ loop (adjacent to the 40−45 segment mentioned above) show significant NCS deviation in the new GATE-16 structure. As in the case of the N-terminal segment, such differences correlate with broadened or weak NH signals in solution NMR, which are even undetectable for residues 37 and 38, indicating these residues can adopt different conformations with intermediate exchange rates on the NMR time scale. While we could not obtain accurate dynamics information for residue 44 because of overlap, the absence of observable resonances for residues 46 and 47 and the weak signal of residue 48 (Figure 2C), presumably because of exchange broadening, support conformational motion of this region. The two chains in the asymmetric unit of our crystal structure are virtually identical in the $\beta3-\alpha4$ loop region containing residues 81−84; nonetheless, flexibility in solution is indicated by the relatively low S^2 value of residues 81 and 84.

To further investigate these conformational changes, we performed an HREMD simulation of GATE-16 in solution. Only structures collected for the target replica were considered for analysis. With regard to the conformation of the C-terminus, the simulations revealed significant fluctuations, in accordance with data acquired by other methods. These fluctuations are reflected by the large *B* factors obtained from the HREMD simulation (Figure 1C), which are directly correlated to the mean fluctuations shown in Figure S5. For further characterization of the C-terminal motion, we clustered the structures as explained in detail above. In short, we tested each structure for the existence of the salt bridge between the N-terminal residue Met1 and the C-terminal residue Phe117, coverage of the hydrophobic groove by Phe115, Phe117, or both, and, in case Phe115 is in the hydrophobic pocket, whether the side chain points toward the N-terminus. Application of these criteria results in 12 distinct states, which are listed in Table 2.

Table 2. Characteristics of the 12 GATE-16 States Determined from the HREMD Simulation

state	population (%)	Phe115 contacts h.p.[a]	Phe115 toward the N-terminus	Phe117 contacts h.p.[a]	Met1−Phe117 salt bridge
1	41.2	yes	yes	no	yes
2	15.7	yes	yes	no	no
3	14.7	no	n/a	no	no
4	8.4	yes	no	no	no
5	4.3	yes	no	yes	no
6	4.1	yes	no	no	yes
7	3.5	no	n/a	yes	yes
8	2.5	yes	no	yes	yes
9	2.3	no	n/a	yes	no
10	1.5	yes	yes	yes	no
11	1.1	yes	yes	yes	yes
12	0.8	no	n/a	no	yes

[a] h.p., hydrophobic pocket.

The salt bridge between Met1 and Phe117 has about equal probability of being open or closed as in 53.2% of all structures this salt bridge is formed. A more stable interaction derives from the hydrophobic contact between Phe115 and the hydrophobic groove formed by Met1, Val36, Phe79, and the aliphatic portion of Lys82. This contact is present in 78.8% of all sampled conformations, including the two most populated states, which account for 56.9% of all structures. Comparison of states 1 and 2 indicates that Phe115 can cover the hydrophobic groove regardless of whether the salt bridge is formed. Representative structures for these two states are depicted in panels A and B of Figure 3. Conformations similar to state 2 are found in state 4 (Figure 3D), the only difference being the orientation of Phe115 within the hydrophobic groove. In state 2, it is oriented toward the N-terminus, while it points away in state 4. Phe117 can also form contacts with the hydrophobic groove yet has a weaker tendency to do so than Phe115. Only in 15.2% of all structures does Phe117 reside within the hydrophobic groove. Interestingly, in more than half of these structures (61.8%), both Phe115 and Phe117 are inside the hydrophobic groove (states 5, 8, 10, and 11), while in states 7 and 9, only Phe117 is there. An interesting conformation, though of minor importance as it is populated with only 0.8% probability, is given by state 12, in which neither Phe115 nor Phe117 covers the hydrophobic groove, yet the Met1−Phe117 salt bridge is formed. In this case, the salt bridge prevents complete extension of the C-terminus, which in most of the structures is secured by the hydrophobic interactions of Phe115 and/or Phe117. When none of these three stabilizing interactions is present, the C-terminus becomes even more flexible and detaches from the core of the protein, exposing both Phe115 and Phe117 to the solvent. This situation is present in state 3, for which a representative structure is shown in Figure 3C. The biological relevance of the conformations with extended C-termini, which account for 14.8% of the structures sampled in our HREMD simulation, is discussed below. Together, the top four states include 80% of all structures. Snapshots of the other eight states representing the remaining 20% of the C-terminal conformations are shown in Figure S4.

The flexibility of the C-terminus is reflected in the S^2 order parameter, which we calculated for each residue from additional simulations. To this end, we performed multiple 20 ns MD simulations using structures from the top four states as starting conformations, where the number of MD runs for each state was derived from the probabilities listed in Table 2. A similar approach has been suggested by Blackledge and co-workers.[54] They found that running multiple but short all-atom explicit solvent MD simulations exploring the different conformational substates sampled from a preceding accelerated MD simulation resulted in better reproduction of order parameters compared to the same number of simulations starting from the relaxed crystal structure. In Figure 2D, the averages of the computed S^2 data together with standard deviations are shown. In general, the theoretical S^2 values display a distribution very similar to the distribution of those derived from our NMR data. Both simulation and experiment indicate that the C-terminal residues Thr114−Phe117 represent the most flexible part of the protein, which is reasonable considering that we observed 12 different conformers for the C-terminus (Table 2). Noteworthy deviations are observed for only the five N-terminal residues, for Asp45, which is located in the $\beta1-\beta2$ loop, and for Val84 in the middle of the $\beta3-\alpha4$ loop. In either case, a higher flexibility is predicted by NMR.

To test the dependence of our results on the force field, we performed three 250 ns MD simulations with the CHARMM27[55] force field with TIP3P water. Our decision to use CHARMM27 was motivated by two reasons. (i) We wanted to use a force field not belonging to the Amber family, which is based on a completely different parametrization scheme to avoid any possible force field bias. (ii) Taking (i) and the force field comparison in ref 35 into account, we opted for CHARMM27 as it also performs well for ubiquitin. To warrant direct comparability between our Amber99SB-ILDN and CHARMM27 calculations of S^2, we kept TIP3P as a water model. The CHARMM27 simulations were initiated from chain B of the GATE-16 structure determined in this study. All other simulation conditions were identical to those of the 20 ns MD

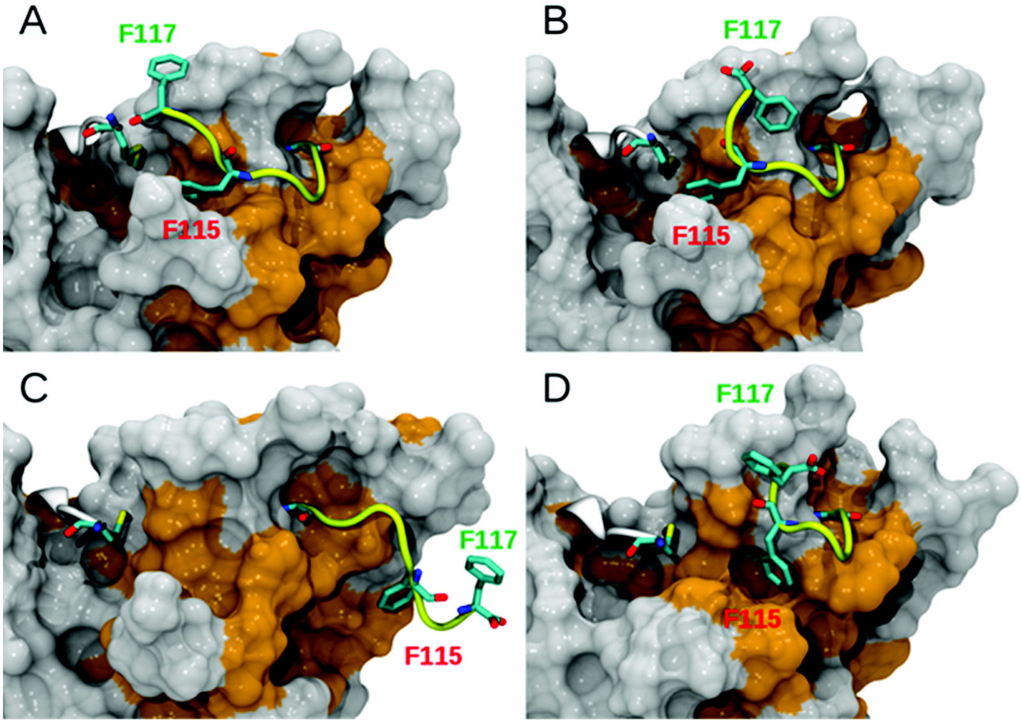


Figure 3. Representative structures for states 1−4 (A−D, respectively) obtained from the HREMD simulation. The structures from panels A−C correspond to traces S1−S3, respectively, in Figure 1A. Most residues of GATE-16 are shown in surface representation with hydrophobic residues highlighted in orange. Residues Met1−Trp3 and Gly111−Phe117 are shown as cartoons, and the C-terminus is colored yellow (as in Figure 1A). The side chains of the residues of particular interest (Met1, Phe115, and Phe117) are shown explicitly.

simulations with Amber99SB-ILDN. For the analysis, only the last 200 ns of each simulation was considered. The root-mean-square fluctuations and S^2 values averaged over the three runs agree equally well to the experiment and those obtained with Amber99SB-ILDN (Figures S5 and S6). The conformations sampled during two of these simulations belong to states 1 and 2 determined from the HREMD simulations. This is not surprising as the top two states represent more than half (56.9%) of all structures; i.e., they are considerably stable. In the third MD simulation, structures belonging to state 7 were exclusively sampled. Representative structures from these three simulations are shown in Figure S7, together with the evolution of the Met1−Phe117 distance and the number of hydrophobic contacts of Phe115 and Phe117. The results in panel A of this figure reveal that of these three interactions, the state of the surface-exposed Met1−Phe117 salt bridge can change most quickly. It can spontaneously form and break again, as happened multiple times between 50 and 120 ns where the lifetime of the salt bridge is short. The instability of the salt bridge is a consequence of the flexibility of the C-terminus, which is mainly brought about by the presence of a Gly residue at position 116.

■ DISCUSSION

Various functions of biological macromolecules are critically related to dynamic properties, covering all structural levels from single atoms to entire domains, and time scales ranging from picoseconds to seconds. In the case of proteins, collective motions of large regions are of particular interest, because these are often required during the catalytic cycle of enzymes or for switching the functional state of a protein. The two prevailing methods of protein structure determination, X-ray crystallography and solution NMR spectroscopy, not only differ by the physical state of the samples to be investigated but also provide disparate parametrizations of coordinate uncertainty: while a crystallographic structure is represented by mean atomic coordinates and associated atomic displacement parameters, NMR yields an ensemble of models compatible with a given set of restraints. Because they provide different views on a common subject, combining results from both methods can be instructive.

In this work, we have investigated the conformational polymorphism of GATE-16 by (1) comparison of different crystal structures, including a newly determined one, (2) direct assessment of dynamics using solution NMR spectroscopy, and (3) MD simulations in an explicitly solvated system.

The X-ray structure of GATE-16 described in this study has been determined from crystals grown in the absence of typical precipitating agents and should therefore suffer less from artifacts caused by high concentrations of such compounds. As expected, the overall fold of the protein is not altered significantly during enhanced conformational sampling provided by an HREMD simulation with a total of 10 μs of simulation time. Major fluctuations were mostly observed for the last six residues (Glu112−Phe117), which appear to occupy a number of transient states (Figure 3 and Table 2). These states are characterized by the presence or absence of a salt bridge between Phe117 and Met1, and different hydrophobic contacts mediated by the Phe115 and Phe117 side chains. It is important to note that the bend at Gly111 is conserved in all states excluding state 3, corresponding to 85.3% of the sampled population. In these conformers, the C-terminal segment following strand β4 thus remains attached to the body of the molecule. Both our experimental data and *in silico* results support the notion that Phe115 plays a crucial role in stabilizing these six residues against complete extension; the aromatic side chain engages in a stable interaction with its hydrophobic groove (Figure 1B). This stabilization is of special importance when the C-terminus is not confined by the Met1−Phe117 salt bridge (Figure 3B). In a minority of the conformers, Phe117 engages as an alternative hydrophobic anchor.

To allow proteolytic cleavage of their scissile bonds (a prerequisite for lipidation), the C-termini of Atg8 family proteins need to protrude into the catalytic center of their cognate ATG4 proteases. This extended conformation is documented by a set of crystal structures featuring rat LC3B in complex with human ATG4B,[12] and its general properties are expected to be valid for all Atg8 homologues. While the GATE-16 crystal structure features a sharp bend following strand β4, the corresponding segment of LC3B takes a straight path in the ATG4B complex. Notably, the C-terminal stretch appears to be displaced from its position by a loop of ATG4B, with Leu232 of the protease occupying the binding pocket of Phe119 (corresponding to Phe115 in GATE-16), while the segment containing Phe119 interacts with the substrate channel of ATG4B. Because of the exposure of additional hydrophobic side chains, the extended conformation is not expected to be a favorable state for an isolated GATE-16 molecule in solution; it can, however, be stabilized by a binding partner such as ATG4B. In the GATE-16 structure determined previously,[19] chain A (included in Figure 1A) represents an extended conformation akin to LC3B in complex with ATG4B. In this case, the neighboring GATE-16 molecule (chain B) takes the role of ATG4B in stabilizing this conformation. Specifically, the anchoring residue Phe115 is accommodated by a hydrophobic groove normally involved in ligand binding, whereas Ile55 in the $\beta2-\alpha3$ loop shields the Phe115 groove. Intriguingly, our HREMD simulation of isolated GATE-16 does capture detachment of the C-terminal segment (characterized by the absence of the Met1−Phe117 salt bridge, solvent exposure of Phe115 and Phe117 side chains, and loss of the Gly111 bend) in one significantly populated state (state 3, accounting for 14.7% of the conformers). Together with published data, this observation confirms the notion that the swing-out movement of the C-terminus is a built-in capability of the GATE-16 molecule and that the extended conformation found in the ATG4 complex is mostly selected, rather than induced, by the protease.

On the basis of these considerations, we propose a three-state model for the C-terminal dynamics of GATE16. (1) In newly synthesized GATE-16, the preferred conformation of the C-terminal segment is a closed state, which is characterized by a bend of the chain at Gly111, anchorage of the side chain of Phe115 to a hydrophobic groove on the ubiquitin-like domain, and a salt bridge keeping Phe117 in the vicinity of Met1. (2) This state is in dynamic equilibrium with a number of alternate conformations, in which the C-terminal residues are displaced to a varying extent. Most of the time, the Phe115 or Phe117 hydrophobic anchor is preserved, which ≈50% of the time is further stabilized by the Met1−Phe117 salt bridge. Full detachment with loss of the bend at Gly111 is possible but occurs with a lower probability. (3) When GATE-16 is bound to ATG4 family proteases, a completely extended state of the C-terminus is stabilized by a loop protruding from the enzyme, which shields the hydrophobic groove formerly accommodating Phe115 or Phe117 and, at the same time, sterically prevents the C-terminus from reverting to an "attached" state.

Lipidation of truncated GATE-16 (form I) at Gly116 involves sequential interaction with the E1-like enzyme ATG7 and the E2-like ATG3. In the respective complexes, an extended conformation of the C-terminal segment is expected (as confirmed experimentally by the crystal structure of the yeast Atg7−Atg8 complex[56]). With regard to the membrane-associated PE conjugate (form II), a similar displacement of the C-terminus is supported by our previous NMR experiments with the GATE-16 homologue GABARAP.[13] Here, coupling of residue 116 to a nanodisc membrane was shown to result in significant chemical shift changes for (among others) Tyr115 and Asn82, which correspond to Phe115 and Lys82 in GATE-16, respectively, suggesting that anchorage of the aromatic side chain to its groove is at least partially lost (Figure S2). Finally, delipidation of form II by ATG4 should cause the C-terminus to revert to its previous equilibrium, with the caveat that a Met1−Phe117 salt bridge will no longer be possible. In this context, we note that our MD simulation suggests Asp81 as a substitute interaction partner of the N-terminal amino group.

Analysis of primary and tertiary structures of human LC3 subfamily proteins reveals that all of them contain an aromatic amino acid corresponding to Phe115 in GATE-16 as well as apolar residues at the positions of Val36 and Phe79, suggesting that the mechanism anchoring the C-terminus to the body of the protein may be conserved throughout the Atg8 family.

In an effort to characterize the relevance of Arg68 for the structure and dynamics of LC3B, Liu et al. performed MD simulations of the wild-type protein as well as a splice variant lacking this residue. In contrast to our results with GATE-16, they did not detect any stable conformation of the C-terminal segment for wild-type LC3B.[57] It should be noted, however, that their simulation time was limited to 13 ns, while our approach (a combination of HREMD and conventional full-atom MD for extended periods of time) allows for much more exhaustive sampling of conformational space. Moreover, the starting model employed in the LC3B study was constructed by grafting residues 114−116 of GATE-16 (PDB entry 1EO6) into rat LC3B (PDB entry 1UGM[58]), possibly leading to a mixed conformation requiring more time for convergence to a low-energy state. It is therefore tempting to speculate that, with comparable simulation times, an equilibrium similar to that found for GATE-16 might also materialize for LC3B.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.biochem.5b00366.

Original 2D NMR spectra as well as additional simulation data (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*Phone: +49 2461 61-2028. E-mail: o.h.weiergraeber@fz-juelich.de.

Present Addresses

§P.M.: Shanghai Institute for Advanced Immunochemical Studies (SIAIS), ShanghaiTech University, Shanghai 201210, China.

‖O.O.: Department of Pharmacology and Therapeutics, College of Medicine and Health Sciences, Afe Babalola University Ado-Ekiti, Nigeria.

Funding

This work was supported by a research grant from the Deutsche Forschungsgemeinschaft (Wi 1472/5).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We acknowledge the ESRF for provision of synchrotron radiation facilities and thank the staff for assistance in using beamline ID14-4. Computational support and infrastructure was provided by the Centre for Information and Media Technology (ZIM) at the University of Düsseldorf (Düsseldorf, Germany).

■ ABBREVIATIONS

Atg or ATG, autophagy-related; DTT, dithiothreitol; GABA, γ-aminobutyric acid; GABARAP, GABA_A receptor-associated protein; GABARAPL, GABARAP-like protein; GATE-16, Golgi-associated ATPase enhancer of 16 kDa; GEC1, glandular epithelial cell protein 1; het, heteronuclear; MAP1LC3, light chain 3 of microtubule-associated protein 1; MD, molecular dynamics; NCS, noncrystallographic symmetry; NMR, nuclear magnetic resonance; NOE, nuclear Overhauser effect; *NPT*, constant particle number, pressure, and temperature; PE, phosphatidylethanolamine; rmsd, root-mean-square deviation; SD, standard deviation.

■ REFERENCES

(1) Nakatogawa, H., Suzuki, K., Kamada, Y., and Ohsumi, Y. (2009) Dynamics and diversity in autophagy mechanisms: lessons from yeast. *Nat. Rev. Mol. Cell Biol. 10*, 458−467.

(2) Shpilka, T., Weidberg, H., Pietrokovski, S., and Elazar, Z. (2011) Atg8: an autophagy-related ubiquitin-like protein family. *Genome Biol. 12*, 226.

(3) Weiergräber, O. H., Mohrlüder, J., and Willbold, D. (2013) Atg8 family proteins—Autophagy and beyond. In *Autophagy: A Double-Edged Sword—Cell Survival or Death?* (Bailly, Y., Ed.) InTech: Rijeka, Croatia. pp 13−45.

(4) Weidberg, H., Shvets, E., Shpilka, T., Shimron, F., Shinder, V., and Elazar, Z. (2010) LC3 and GATE-16/GABARAP subfamilies are both essential yet act differently in autophagosome biogenesis. *EMBO J. 29*, 1792−1802.

(5) Kabeya, Y., Mizushima, N., Yamamoto, A., Oshitani-Okamoto, S., Ohsumi, Y., and Yoshimori, T. (2004) LC3, GABARAP and GATE16 localize to autophagosomal membrane depending on form-II formation. *J. Cell Sci. 117*, 2805−2812.

(6) Stangler, T., Mayr, L. M., and Willbold, D. (2002) Solution structure of human GABA_A receptor-associated protein GABARAP: implications for biological function and its regulation. *J. Biol. Chem. 277*, 13363−13366.

(7) Schwarten, M., Stoldt, M., Mohrlüder, J., and Willbold, D. (2010) Solution structure of Atg8 reveals conformational polymorphism of the N-terminal domain. *Biochem. Biophys. Res. Commun. 395*, 426−431.

(8) Weiergräber, O. H., Stangler, T., Thielmann, Y., Mohrlüder, J., Wiesehan, K., and Willbold, D. (2008) Ligand binding mode of GABA_A receptor-associated protein. *J. Mol. Biol. 381*, 1320−1331.

(9) Thielmann, Y., Weiergräber, O. H., Mohrlüder, J., and Willbold, D. (2009) Structural framework of the GABARAP-calreticulin interface−implications for substrate binding to endoplasmic reticulum chaperones. *FEBS J. 276*, 1140−1152.

(10) Thielmann, Y., Weiergräber, O. H., Mohrlüder, J., and Willbold, D. (2009) Structural characterization of GABARAP-ligand interactions. *Mol. BioSyst. 5*, 575−579.

(11) Ma, P., Schwarten, M., Schneider, L., Boeske, A., Henke, N., Lisak, D., Weber, S., Mohrlüder, J., Stoldt, M., Strodel, B., Methner, A., Hoffmann, S., Weiergräber, O. H., and Willbold, D. (2013) Interaction of Bcl-2 with the autophagy-related GABA_A receptor-associated protein (GABARAP): biophysical characterization and functional implications. *J. Biol. Chem. 288*, 37204−37215.

(12) Satoo, K., Noda, N. N., Kumeta, H., Fujioka, Y., Mizushima, N., Ohsumi, Y., and Inagaki, F. (2009) The structure of Atg4B-LC3 complex reveals the mechanism of LC3 processing and delipidation during autophagy. *EMBO J. 28*, 1341−1350.

(13) Ma, P., Mohrlüder, J., Schwarten, M., Stoldt, M., Singh, S. K., Hartmann, R., Pacheco, V., and Willbold, D. (2010) Preparation of a functional GABARAP-lipid conjugate in nanodiscs and its investigation by solution NMR spectroscopy. *ChemBioChem 11*, 1967−1970.

(14) Leslie, A. G. W. (1992) Recent changes to the MOSFLM package for processing film and image plate data. *Joint CCP4/ESF-EAMCB Newsletter on Protein Crystallography*, Vol. *26*.

(15) Evans, P. R. (2006) Scaling and assessment of data quality. *Acta Crystallogr., Sect. D: Biol. Crystallogr. D62*, 72−82.

(16) French, G. S., and Wilson, K. S. (1978) On the treatment of negative intensity observations. *Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr. A34*, 517−525.

(17) Winn, M. D., Ballard, C. C., Cowtan, K. D., Dodson, E. J., Emsley, P., Evans, P. R., Keegan, R. M., Krissinel, E. B., Leslie, A. G., McCoy, A., McNicholas, S. J., Murshudov, G. N., Pannu, N. S., Potterton, E. A., Powell, H. R., Read, R. J., Vagin, A., and Wilson, K. S. (2011) Overview of the CCP4 suite and current developments. *Acta Crystallogr., Sect. D: Biol. Crystallogr. D67*, 235−242.

(18) Vagin, A., and Teplyakov, A. (1997) MOLREP: an automated program for molecular replacement. *J. Appl. Crystallogr. 30*, 1022−1025.

(19) Paz, Y., Elazar, Z., and Fass, D. (2000) Structure of GATE-16, membrane transport modulator and mammalian ortholog of autophagocytosis factor Aut7p. *J. Biol. Chem. 275*, 25445−25450.

(20) 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., and Zwart, P. H. (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr., Sect. D: Biol. Crystallogr. D66*, 213−221.

(21) Emsley, P., Lohkamp, B., Scott, W. G., and Cowtan, K. (2010) Features and development of Coot. *Acta Crystallogr., Sect. D: Biol. Crystallogr. D66*, 486−501.

(22) Chen, V. B., Arendall, W. B., 3rd, Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S., and Richardson, D. C. (2010) MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr., Sect. D: Biol. Crystallogr. D66*, 12−21.

(23) Delaglio, F., Grzesiek, S., Vuister, G. W., Zhu, G., Pfeifer, J., and Bax, A. (1995) NMRPipe. A multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR 6*, 277−293.

(24) Vranken, W. F., Boucher, W., Stevens, T. J., Fogh, R. H., Pajon, A., Llinas, M., Ulrich, E. L., Markley, J. L., Ionides, J., and Laue, E. D. (2005) The CCPN data model for NMR spectroscopy. Development of a software pipeline. *Proteins: Struct., Funct., Genet. 59*, 687−696.

(25) d'Auvergne, E. J., and Gooley, P. R. (2008) Optimisation of NMR dynamic models I. Minimisation algorithms and their performance within the model-free and Brownian rotational diffusion spaces. *J. Biomol. NMR 40*, 107−119.

(26) d'Auvergne, E. J., and Gooley, P. R. (2008) Optimisation of NMR dynamic models II. A new methodology for the dual optimization of the model-free parameters and the Brownian rotational diffusion tensor. *J. Biomol. NMR 40*, 121−133.

(27) Clore, G. M., Szabo, A., Bax, A., Kay, L. E., Driscoll, P. C., and Gronenborn, A. M. (1990) Deviations from the simple 2-parameter

model-free approach to the interpretation of N-15 nuclear magnetic-relaxation of proteins. *J. Am. Chem. Soc. 112*, 4989−4991.

(28) d'Auvergne, E. J., and Gooley, P. R. (2003) The use of model selection in the model-free analysis of protein dynamics. *J. Biomol. NMR 25*, 25−39.

(29) d'Auvergne, E. J., and Gooley, P. R. (2006) Model-free model elimination: A new step in the model-free dynamic analysis of NMR relaxation data. *J. Biomol. NMR 35*, 117−135.

(30) d'Auvergne, E. J., and Gooley, P. R. (2007) Set theory formulation of the model-free problem and the diffusion seeded model-free paradigm. *Mol. BioSyst. 3*, 483−494.

(31) Lipari, G., and Szabo, A. (1982) Model-free approach to the interpretation of nuclear magnetic-resonance relaxation in macromolecules I. Theory and range of validity. *J. Am. Chem. Soc. 104*, 4546−4559.

(32) Lipari, G., and Szabo, A. (1982) Model-free approach to the interpretation of nuclear magnetic-resonance relaxation in macromolecules II. Analysis of experimental results. *J. Am. Chem. Soc. 104*, 4559−4570.

(33) Bussi, G. (2014) Hamiltonian replica exchange in GROMACS: a flexible implementation. *Mol. Phys. 112*, 379−384.

(34) Lindorff-Larsen, K., Piana, S., Palmo, K., Maragakis, P., Klepeis, J. L., Dror, R. O., and Shaw, D. E. (2010) Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Struct., Funct., Genet. 78*, 1950−1958.

(35) Beauchamp, K. A., Lin, Y.-S., Das, R., and Pande, V. S. (2012) Are protein force fields getting better? A systematic benchmark on 524 diverse NMR measurements. *J. Chem. Theory Comput. 8*, 1409−1414.

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

(37) Hess, B., Bekker, H., Berendsen, H. J. C., and Fraaije, J. G. E. M. (1997) LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem. 18*, 1463−1472.

(38) Darden, T., York, D., and Pedersen, L. (1993) Particle mesh Ewald: an N·log(N) method for Ewald sums in large systems. *J. Chem. Phys. 98*, 10089−10092.

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

(40) Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A., and Haak, J. R. (1984) Molecular dynamics with coupling to an external bath. *J. Chem. Phys. 81*, 3684−3690.

(41) Hoover, W. G. (1985) Canonical dynamics: equilibrium phase-space distributions. *Phys. Rev. A: At., Mol., Opt. Phys. 31*, 1695−1697.

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

(43) Mao, Y., and Zhang, Y. (2012) Thermal conductivity, shear viscosity and specific heat of rigid water models. *Chem. Phys. Lett. 542*, 37−41.

(44) Maragakis, P., Lindorff-Larsen, K., Eastwood, M. P., Dror, R. O., Klepeis, J. L., Arkin, I. T., Jensen, M. Ø., Xu, H., Trbovic, N., Friesner, R. A., Palmer, A. G., III, and Shaw, D. E. (2008) Microsecond molecular dynamics simulation shows effect of slow loop dynamics on backbone amide order parameters of proteins. *J. Phys. Chem. B 112*, 6155−6158.

(45) Kraulis, P. J. (1991) MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr. 24*, 946−950.

(46) Merritt, E. A., and Bacon, D. J. (1997) Raster3D: photorealistic molecular graphics. *Methods Enzymol. 277*, 505−524.

(47) Kabsch, W., and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers 22*, 2577−2637.

(48) Hunter, J. D. (2007) Matplotlib: A 2D graphics environment. *Comput. Sci. Eng. 9*, 90−95.

(49) Waskom, S. Seaborn: Statistical data visualization (http://stanford.edu/~mwaskom/software/seaborn/).

(50) Humphrey, W., Dalke, A., and Schulten, K. (1996) VMD - visual molecular dynamics. *J. Mol. Graphics 14*, 33−38.

(51) Debye, P. (1913) Interferenz von Röntgenstrahlen und Wärmebewegung. *Ann. Phys. 348*, 49−92.

(52) Karplus, P. A., and Schulz, G. E. (1985) Prediction of chain flexibility in proteins. *Naturwissenschaften* 72, 212−213.

(53) García de la Torre, J., Huertas, M. L., and Carrasco, B. (2000) HYDRONMR: prediction of NMR relaxation of globular proteins from atomic-level structures and hydrodynamic calculations. *J. Magn. Reson. 147*, 138−146.

(54) Markwick, P. R. L., Bouvignies, G., and Blackledge, M. (2007) Exploring multiple timescale motions in protein GB3 using accelerated molecular dynamics and NMR spectroscopy. *J. Am. Chem. Soc. 129*, 4724−4730.

(55) Mackerell, A. D., Feig, M., and Brooks, C. L. (2004) Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J. Comput. Chem. 25*, 1400−1415.

(56) Noda, N. N., Satoo, K., Fujioka, Y., Kumeta, H., Ogura, K., Nakatogawa, H., Ohsumi, Y., and Inagaki, F. (2011) Structural basis of Atg8 activation by a homodimeric E1, Atg7. *Mol. Cell 44*, 462−475.

(57) Liu, C., Ma, H., Wu, J., Huang, Q., Liu, J. O., and Yu, L. (2013) Arginine68 is an essential residue for the C-terminal cleavage of human Atg8 family proteins. *BMC Cell Biol. 14*, 27.

(58) Sugawara, K., Suzuki, N. N., Fujioka, Y., Mizushima, N., Ohsumi, Y., and Inagaki, F. (2004) The crystal structure of microtubule-associated protein light chain 3, a mammalian homologue of Saccharomyces cerevisiae Atg8. *Genes Cells 9*, 611−618.

3.3 Molecular Dynamics Simulations Reveal Key Roles of the Interleukin-6 Alpha Receptor in the Assembly of the Human Interleukin-6 Receptor Complex

3.3.1 Summary

The cytokine Interleukin 6 (IL-6, section 2.3.3) stimulates immune responses after burns, tissue damage or infections. It needs to bind its associated α receptor first, after which it forms a complex with the β receptor gp130, and this trimer then forms a membrane-anchored hexamer with a second identical trimer. Interestingly, a 25% sequence identical IL-6 variant from human herpesvirus 8 does not need to bind the α receptor, but can directly bind the β-receptor gp130, leading to a tetramer involving two viral IL-6 and two gp130 proteins.

In this research project, we answered the questions: What is the function of the α receptor for human IL-6 and why is it not needed by viral IL-6?

To this end, we studied receptor bound and apo states of IL-6 with extensive MD simulations. We validated the simulation results against experimental S^2 order parameters and chemical shifts obtained by NMR spectroscopy. We found out that the α receptor pivots around the four-helix bundle of IL-6 and subsequently stabilizes an exceptionally flexible part of the AB-loop in human IL-6, which interacts with the D1 domain of the gp130 receptor. This stabilization is presumably required to bind gp130, especially, as it orients several key residues to present ideal binding partners to the gp130 D1 interface.

Viral IL-6, however, has a much more hydrophobic AB-loop, which is significantly more rigid at its N-terminus and very flexible in the center, compared to human IL-6. The more rigid portion can readily interact with the D1 domain of gp130, especially as it lacks a lysine residue that is blocking the interaction in the apo state of human IL-6. The extremely flexible portion is at the same time more hydrophobic than the corresponding section in human IL-6 and therefore the thermodynamic advantage of interactions with the D1 domain of gp130, which reduce the solvent accessible area, is large enough to allow gp130 to bind in the absence of the α receptor.

3.3.2 Contribution

I modelled the starting structures, performed the MD simulations, analyzed the data, co-developed the hypothesis, wrote approximately 90 % of the manuscript and created all figures.

3.3.3 Publication

This section contains a complete reprint of the publication [98]. The supporting information to this article is located in section 6.3.

THE JOURNAL OF PHYSICAL CHEMISTRY B

Article

pubs.acs.org/JPCB

Molecular Dynamics Simulations Reveal Key Roles of the Interleukin-6 Alpha Receptor in the Assembly of the Human Interleukin-6 Receptor Complex

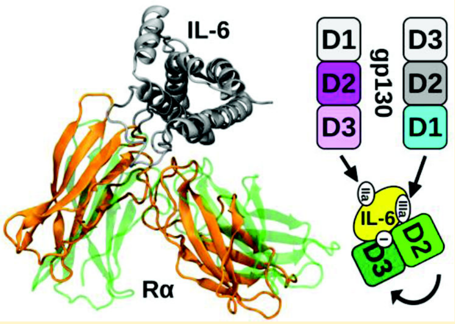
Oliver Schillinger,[†,‡] Vineet Panwalkar,[†,‡] Birgit Strodel,[†,§] and Andrew J. Dingley*[†,‡]

[†]Institute of Complex Systems, Structural Biochemistry (ICS-6), Forschungszentrum Jülich, 52425 Jülich, Germany

[‡]Institut für Physikalische Biologie, Heinrich-Heine-Universität, 40225 Düsseldorf, Germany

[§]Institut für Theoretische Chemie und Computerchemie, Heinrich-Heine-Universität, 40225 Düsseldorf, Germany

Supporting Information



ABSTRACT: Human interleukin-6 (hIL-6) is a pleiotropic cytokine with three distinct receptor epitopes, termed sites I, II, and III, which function to assemble a signaling complex. hIL-6 signals via a glycoprotein 130 (gp130) homodimer after initially forming a heterodimer with the nonsignaling α-receptor (IL-6Rα). The molecular description of the assembly of the hIL-6 signaling complex remains elusive because available structures provide descriptions of hIL-6 in its free and fully bound receptor forms, but not for intermediate steps that are crucial in the stepwise assembly of the signaling complex. In this report, molecular dynamics simulations provide atomic details describing the functional role of the initial hIL-6/IL-6Rα complex in facilitating subsequent interactions with gp130, which have not been previously shown. IL-6Rα binding to hIL-6 rigidifies the flexible N-terminus of the hIL-6 AB-loop through interactions with the D2 domain of IL-6Rα. This rigidification combined with repositioning of residues involved in gp130 receptor recognition promotes gp130 binding at site III. Binding of gp130 receptors at sites II and III is coupled with the release of the hIL-6 N-terminal AB-loop interaction and a pivoting of IL-6Rα around the hIL-6 helix bundle to the state of the hIL-6/IL-6Rα/gp130 complex.

INTRODUCTION

Human interleukin-6 (hIL-6) is a pleiotropic cytokine, which regulates biological functions that include the immune response, hematopoiesis, the acute-phase response and inflammation,[1−3] and dysfunctional hIL-6 signaling, is linked to acute and chronic inflammation, autoimmune diseases, and neoplastic disorders.[4−6] hIL-6 signals through cis- or trans-mediated pathways and the pathways differ substantially in their cellular distribution.[7] Cis-signaling is mediated by the membrane-bound alpha-receptor (IL-6Rα), which is expressed on a limited number of cell types, whereas for trans-signaling the soluble form of the IL-6Rα can stimulate all cells.[8] IL-6 adopts the canonical four-helix bundle fold of cytokines with the helices connected by two long (A−B and C−D) and one short (B−C) interhelical loops.[9,10] Three distinct receptor binding epitopes, termed sites I, II, and III, exist on hIL-6. Residues at site I of hIL-6 interact with residues of the extracellular domains 2 (D2) and 3 (D3) of IL-6Rα (i.e., cytokine binding homology region) to form a heterodimer.[11] This hIL-6/IL-6Rα complex is essential for recruiting two ubiquitously expressed membrane-bound glycoprotein 130 (gp130) signaling receptors at sites II and III (Figure 1);[12−14] because neither hIL-6 nor IL-6Rα alone has sufficient affinity to form a stable complex with gp130. Dimer formation of gp130 activates Janus kinase/signal transducer and activator of transcription intracellular signaling pathways.[15]

The structure of the IL-6/IL-6Rα/gp130 complex confirmed previous studies that suggested a 2:2:2 stoichiometry for the signaling complex (Figure 1A),[16] and the hexameric complex is postulated to be required for both cis- and trans-mediated signaling.[17] This structure showed that a stable hexameric complex is formed because the IL-6−gp130 interfaces at sites II and III are extended by interactions between gp130 and IL-6Rα. In this hexameric assembly, sites II and III consist of two energetically coupled composite sites each, with sites involving hIL-6 and gp130 termed IIa and IIIa, and sites IIb and IIIb representing the interfaces between IL-6Rα and gp130 (Figure 1A,B). Interestingly, the Kaposi's sarcoma-associated herpesvirus functional homologue of human IL-6, HHV-8 IL-6 (vIL-6) shows modest sequence identity to hIL-6 (∼25%; Figure S1),[18] and binds directly to gp130 to activate signaling cascades without the requirement of a precomplex with IL-6Rα.[19,20] The sequence divergence between hIL-6 and vIL-6 has been proposed to define the differences in the assembly of the IL-6/gp130 signaling complexes, with vIL-6 interacting with gp130

Received: June 12, 2017
Revised: July 31, 2017
Published: August 7, 2017

DOI: 10.1021/acs.jpcb.7b05732

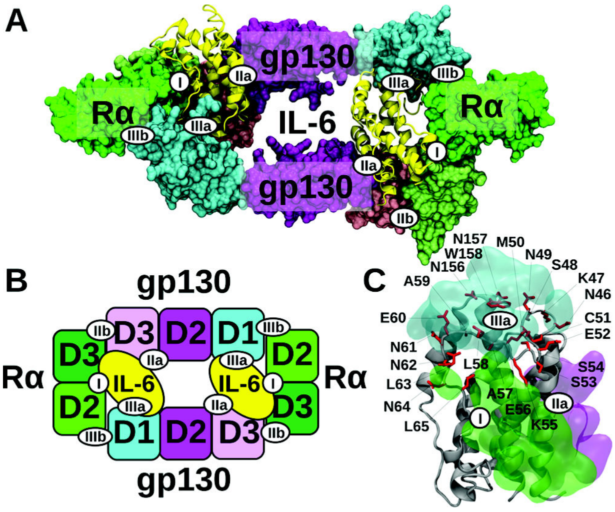


Figure 1. Human IL-6 signaling complex. (A) The crystal structure of the hexameric hIL-6/IL-6R*α*/gp130 signaling complex. hIL-6 is shown in yellow, the IL-6R*α* in shades of green, and the gp130 receptor in cyan, purple, and rose to indicate the individual subdomains. The three binding sites of hIL-6 are indicated with labeled circles. (B) Schematic representation of the hexameric assembly. (C) The binding epitopes of hIL-6 are shown with transparent surface representation with the same color-coding as panels (A) and (B). Side chains involved in receptor interactions are shown as stick representations and labeled with assignment information.

through a larger number of hydrophobic contacts when compared with the hIL-6 interaction interfaces (i.e., sites IIa and IIIa).[21] Binding site IIIa has been divided further into three subsites based on their position in the protein sequence[22] (Figure S2). Site IIIaA includes the C-terminus of helix A and the N-terminus of the AB-loop (residues 40−62), site IIIaB includes the C-terminus of the CD-loop (154−167), and site IIIaC includes residues in the BC-loop (103−115) and is not in direct contact with the D1 immunoglobulin-like domain of gp130 (gp130-D1) based on structural data, yet shown to be crucial for IL-6R*α*-independent binding to gp130 by vIL-6.[22]

Although the molecular assembly of the IL-6 signaling complex has been described at atomic resolution, a detailed molecular account of the assembly process remains undefined. This is because current structural data provide a ground-state description of IL-6 in its free and fully bound forms, but not for any of the possible intermediate steps or structural excursions into higher energy states that are potentially crucial in the assembly process, but are not easily detected experimentally. Here, computational methods, in particular molecular dynamics (MD) simulations have proven to be valuable and augment experimental data to reveal motions and conformations that describe protein functions at the molecular level.[23,24]

We have demonstrated previously that fast time scale (ps-ns) dynamics of apo-hIL-6 show heterogeneity of the calculated backbone order parameter [square of the generalized model-free order parameter (S^2)] among clusters of residues in hIL-6.[25] In particular, the N-terminal region of the long AB-loop, which corresponds to half of the buried surface area that constitutes the IL-6/gp130 site IIIa interaction (Figures 1 and S2), experiences large fluctuations along the conformation of the backbone (S^2 = 0.3−0.8) that are absent at site I or II. Thus, we hypothesize that dynamic properties of the AB-loop inhibit the direct interaction of hIL-6 with gp130 in the absence of the IL-6R*α*, and that binding of IL-6R*α* at site I reduces the dynamics of the AB-loop to favor interaction with gp130 at site IIIa. We have attempted to acquire dynamic data of hIL-6 in complex with a soluble form of IL-6R*α*; however, these efforts were complicated by the molecular size of the complex, yielding spectra with large resonance line widths and significant spectral overlap. In addition, our attempts to study the dynamics of vIL-6 to have comparative data between hIL-6 and vIL-6 have failed because of self-association of bacterially produced recombinant vIL-6, which corroborates a previous study.[19] Therefore, in the absence of stable protein material, computer simulations offer an alternative approach to study IL-6 dynamics.

In this report, motional characteristics of hIL-6 in complex with the cytokine binding homology region of IL-6R*α* (IL-6R*α*-D2D3; Figure 1) were investigated by MD simulations. We show that IL-6R*α*-D2D3 interacts with side chains of residues in the N-terminal region of the AB-loop, which has not been observed previously. Such interactions rigidify this region of hIL-6. Additionally, we report that along with rigidification of this region, conformational changes in IL-6 side chain orientations upon IL-6R*α*-D2D3 binding position these residues for specific interactions with the side chains of gp130-D1 that would not occur in the absence of IL-6R*α*. Based on our MD simulations, we hypothesize that rigidification of this region facilitates interaction of the D1 domain of gp130 at site III and upon interaction the D2 domain of IL-6R*α* disengages with the AB-loop residues to facilitate the full docking of gp130 molecules. Such processes have not been observed previously and rationalize the stepwise assembly process of the hIL-6 receptor-signaling complex. MD simulations of vIL-6 are also presented and augment structural data that describes receptor assembly.

■ METHODS

Residue Numbering. The sequence of the hIL-6 NMR structure (PDB ID: 1IL6[10]) is 165 residues long (numbered from 20 to 185). All IL-6 residue numbers in this report correspond to the residue numbering of this structure. Residue numbers of vIL-6 (PDB ID: 1I1R)[19] and of hIL-6 in the IL-6/IL-6R*α*/gp130 complex (PDB ID: 1P9M) are shifted by −14 and −1, respectively. For example, Leu34 of apo hIL-6 corresponds to Leu20 in vIL-6 and to Leu33 of hIL-6 in the hexameric complex.

MD Simulations. To assess the effect of IL-6R*α* binding to hIL-6, MD simulations of hIL-6 were run in the apo state and the IL-6R*α*-D2D3 bound state. The starting structure of the apo state simulation was the NMR minimized average structure (PDB ID: 1IL6).[10] As the unordered part of the AB-loop (residues Ser48, Asn49, Met50) is missing in the crystal structure of the receptor bound hIL-6 (PDB ID: 1P9M),[16] a homology model was built with the SWISS-MODEL Web server.[26] Thus, the hIL-6 structure from the receptor bound hIL-6 structure was used as the template and only the coordinates of the missing residues were added into the crystal structure. The resulting model is very similar to the template with a backbone root-mean-square deviation (RMSD; C*α*, C′, O, and N atoms) between them of 1.0 Å after energy minimization and equilibration of the model. The simulation of vIL-6 was run with the coordinates taken from the gp130 receptor bound tetramer (PDB ID 1I1R).[19] All simulations were performed with GROMACS version 4.6.7,[27] the Amber99SBnmr-ILDN force field,[28] and the TIP3P water model[29] in periodic dodecahedral boxes with a minimum solute

DOI: 10.1021/acs.jpcb.7b05732

to wall distance of 12 Å. Sodium and chloride ions were added as counterions to neutralize the total charge of the system and mimic a salt concentration of approximately 200 mM. The numbers of ions, protein, and solvent atoms, along with further details on the MD simulations are provided in Table S1. Electrostatic interactions were treated with the Particle-Mesh-Ewald algorithm[30] with a Fourier grid spacing of 1.0 Å and cubic interpolation. Short-range electrostatic and van-der-Waals interactions were computed up to a 12 Å cutoff and treated with a Verlet-buffer whose list was updated every 20 time steps. Long range dispersion corrections were applied to the van-der-Waals energy calculations. Moreover, it was tested that the minimum distance between the protein and its periodic images was at all times significantly larger than the cutoff distance for the short-range interactions, ensuring that the protein did not interact with its images during the simulations (Table S1). The systems were energy minimized down to a maximum force of 10 kJ/mol/Å. Equilibration was first done for 100 ps in the NVE ensemble (constant number of particles N, constant volume V, and total energy E) with position restraints of 100 kJ/mol/Å on the heavy backbone atoms, and then for 4 ns in the NPT ensemble (constant pressure P and temperature T) with gradually decaying position restraints ranging from 100 to 1 kJ/mol/Å in magnitude. The purpose of the latter was to allow for a gradual relaxation of the side chains at the interface between IL-6 and IL-6Rα. The integration time step was set to 2 fs. During the equilibration runs a velocity rescale thermostat[31] was used to keep the temperature at 298 K and a Berendsen barostat[32] to keep the pressure close to 1 bar. The human and viral IL-6 simulations were subsequently simulated for 1 μs. Given the larger size of the α-receptor bound system (hIL-6/IL-6Rα-D2D3), three independent MD simulations were run with identical starting structures but randomized velocities for 1.1 μs each. All simulations were performed in the NPT ensemble at the same pressure and temperature as during equilibration (298 K, 1 bar), but with the more accurate Nosé−Hoover thermostat[33] and Parinello-Rahman barostat with coupling time constants of 2 ps.[34] All bond lengths were constrained with the LINCS algorithm.[35] Atom positions were stored every 60 ps.

To estimate the hydrophobic surface area buried at site IIIa due to gp130-D1 binding, one short MD simulation of 1 ns per system was performed for the complexes hIL-6/IL-6Rα/gp130-D1 and vIL-6/gp130-D1 and for both systems in the absence of gp130-D1. These were necessary because residues at the site IIIa interface were not optimally packed in the crystal structures of the IL-6 complexes, resulting in an overestimation of the solvent accessible surface area. All systems were prepared and simulated as described above and the surface area was computed with the Shrake-Rupley algorithm[36] with a probe radius of 1.4 Å. For hIL-6, the buried hydrophobic surface area was determined as the difference in mean hydrophobic surface area of hIL-6 in the hIL-6/IL-6Rα complex and in the hIL-6/IL-6Rα/gp130-D1 complex. For vIL-6, the computation was equivalent except for the absence of IL-6Rα.

Experimental Data. Backbone ^{1}H, ^{15}N, and ^{13}C chemical shifts are from published work.[25]

Analysis. Sequence alignment of human and viral IL-6 was performed with Clustal Omega,[37] using the default settings. The sequences used for the alignment were taken from the UniProtKB[38] database with accession codes P05231, sequence positions 47 to 212 (hIL-6), and Q98823, sequence positions 29 to 195 (vIL-6).

For analysis of the MD simulations, the first 100 ns were additional equilibration time for all simulations and not included in the analysis. Chemical shifts were computed with SHIFTX2[39] under the aid of the MDTRAJ[40] package. For comparison, chemical shifts were also computed with SPARTA+.[41] Both methods gave very similar results (Table S2). Therefore, only the predictions of SHIFTX2 are reported. Prediction of chemical shifts was performed for each frame of the apo hIL-6 trajectory and averaged.

Amide bond vector S^2 order parameters were computed after removal of overall protein tumbling by a superposition of C$_\alpha$ atoms of each frame on the crystal structure:[42]

$$S^2 = \frac{\xi}{2} \sum_{i=1}^{3} \sum_{j=1}^{3} \langle \mu_i \mu_j \rangle^2 - 1 \quad (1)$$

where, μ_i denotes the x, y, and z components of the N−H bond vector scaled to unit length and angular brackets denote averaging over time t. The scaling factor $\xi = (1.02/1.04)^6 \approx 0.89$ accounts for zero point vibrational motions of the bond vectors not captured by MD simulations.[43] S^2 order parameters were separately computed for subtrajectories of 50 ns in length and subsequently averaged to obtain error estimates.

Root mean square fluctuations (RMSFs) were computed over the whole length of the trajectories (900 ns for the apo-state and 1 μs for the complex simulations), after superimposing the helix backbone atoms (i.e., Cα, C′, O, and N atoms) of all frames onto the first frame (i.e., those that exist in a helix in the hIL-6 structure). RMSFs were computed for all backbone atoms as the root of the mean squared displacement of each atom over the course of a trajectory, with respect to the mean structure of each trajectory. To obtain RMSFs per residue, the RMSFs of the backbone atoms of each residue were averaged. Finally, averages and standard deviations over the three independent trajectories were computed for the hIL-6/IL-6Rα simulations. Hence, RMSFs reflect the mobility of each residue on a 1 μs time scale. For the analysis of the binding between different proteins in the complexes, inter-residue contacts were computed. A contact, as for the IL-6Rα binding interfaces, is defined if the distance between any pair of atoms for two residues is <5 Å (the higher cutoff for attractive London-van der Waals forces).[44] In addition, mean (and standard deviation) interaction energies of IL-6 AB-loop residues with IL-6Rα binding interfaces I to III (see Results Section) were determined as the sum of the Lennard-Jones and Coulomb energies. Only the short-range part of the interaction energies up to the cutoffs used in the MD simulation (12 Å) were considered, as the long-range component computed with the Particle-Mesh-Ewald method cannot be computed for separate groups of the system individually. The energies were computed with the GROMACS energy tool and the atoms of the individual groups specified in an index file. For the analysis of the outputted energies, an interaction energy of 0 kJ/mol corresponds to the absence of interaction at infinite distance, a positive energy to a repulsive and a negative energy to an attractive interaction.

The hydrophobic packing of the AB-loop to the hydrophobic core of the cytokine was determined for hIL-6, vIL-6, and the leukemia inhibitory factor (LIF, PDB ID: 1EMR). The amino acid sequence of LIF from the end of helix A to the beginning of helix B was aligned manually to both IL-6 proteins. Residues Val65, Phe68, and Pro69 in the AB-loop of LIF correspond to Leu63, Leu65, and Pro66 in hIL-6, and Leu63, Phe65, and

DOI: 10.1021/acs.jpcb.7b05732

Pro66 in vIL6. These residues are located at the N-terminus of the AB-loop and function as hydrophobic anchors that interact with the hydrophobic core of the cytokine four-helix bundle. The N-terminus of the AB-loop of IL-6 consists of residues 48 to 62, which corresponds to residues 49−64 in LIF. Distances were computed between all side chain heavy atoms of two groups of residues: hydrophobic residues at the N-terminus of the AB-loop (group 1) and all hydrophobic residues with a minimum distance in sequence of two to this region (group 2). Contacts were defined as distances <5 Å.

All plots were produced with VMD,[45] Matplotlib,[46] and seaborn.[47]

RESULTS AND DISCUSSION

Sequence Alignment of Human and Viral IL-6 Shows Low Sequence Identity. The amino acid sequences of human and viral IL-6 align without any insertions or deletions over a length of 166 residues (Figure S1). The alignment is also structurally justified (Figure S3), as it superimposes the four helices and most parts of the connecting loops (backbone RMSD of 1.76 Å for helices and 4.65 Å for the loops, where superposition was performed on the backbone of the helices). However, the sequence identity of both structures of only 24.7% (41 residues) is remarkably low considering the gapless alignment of both sequences. An additional 48 residues are strongly similar and 21 residues weakly similar, yielding a strong sequence similarity of 53.6% (including identical residues) and a weak sequence similarity of 66.3% (including identical and strongly similar residues).

Chemical Shifts of Backbone Nuclei of hIL-6 Derived from the Simulated Data Match Experimental Values. The quality of the MD simulations to sample the relevant ensemble of states was assessed by comparing the MD-calculated backbone chemical shifts with experimental chemical shift data.[25] The experimental and predicted C_α chemical shifts of apo hIL-6 are shown in Figure 2. In addition, comparisons between experimental and predicted backbone C′, N and H_N chemical shifts for apo hIL-6 were also determined (Figures S4−S6). The predicted C_α shifts match the experimental shifts with an RMSD of only 1.12 ppm, which is similar to the RMSD value reported in the original SHIFTX2 validation set of proteins.[39] Larger RMSD deviations (>2 standard deviations of the distribution of residuals of experimental and predicted C_α shifts, equal to 2.15 ppm; Figure 2B) are observed for only 10 residues (Thr21, Cys45, Glu52, Leu65, Phe75, Phe95, Arg105, His165, Ile167, and Phe171). The predictions of the chemical shifts of the C′ and N nuclei (Figures S4 and S5) follow the trends of the experimental values. For the backbone C′ atoms, there are a few nuclei with differences between predicted and experimental chemical shifts >2 standard deviations (2.7 ppm; Cys51, Leu93, Val116, Thr139, and Asp161). The H_N atom shifts display larger RMSDs (Figure S6), and the relative standard deviation over the trajectory, as indicated by the error bars, is also significantly larger when compared with that of the other nuclei. Overall, the predicted chemical shifts of the apo hIL-6 simulation match the experimental shifts remarkably well, even in many flexible regions, such as most residues in the first part of the loop connecting helices A and B (AB-loop) and the CD-loop formed by residues 129 to 140. Hence, the simulations sample the equilibrium structure of hIL-6.

Longer MD Simulations Improve the Calculation of S^2 Data. To further validate the MD simulations, experimentally derived S^2 data[25] were compared with MD calculated S^2 values (Figure S7). The MD-derived S^2 values match closely the experimental values for residues in the four helices (RMSD = 0.06, n = 88), but with a lower variance in the data. In the loop regions, the standard deviation of the S^2 predictions across subtrajectories is much larger than the error in the experimental data, in particular for the C-terminal residues of the AB-loop. As previously reported for a much shorter MD simulation of hIL-6 (i.e., 73.5 ns),[25] the overestimation of the degree of flexibility in the long AB-loop has been attributed to the force field,[42,48] which confers too much flexibility in loops and most likely accounts for the calculated higher flexibility in the loops. Nonetheless, the MD-derived S^2 values in this study are in better overall agreement with the experimental S^2 values when compared with the previous study,[25] indicating that the longer simulation captures more realistically motions on the ns time scale that hIL-6 undergoes.

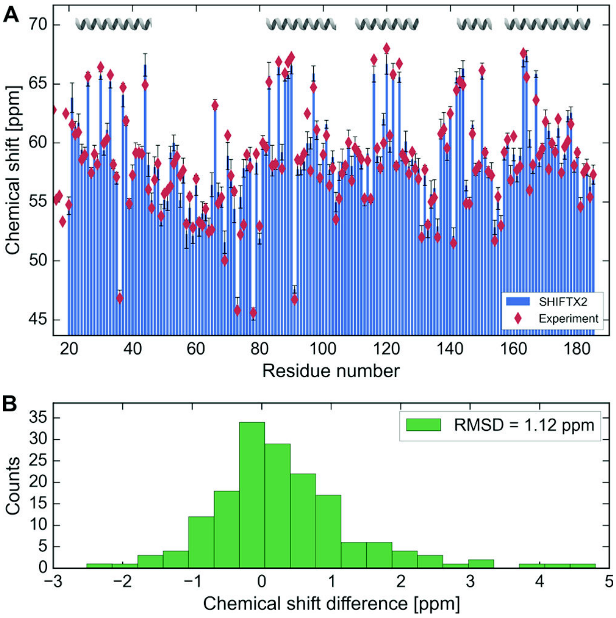


Figure 2. Experimental and MD calculated C_α chemical shifts. (A) Backbone C_α chemical shifts measured by NMR (red diamonds) and predicted with SHIFTX2 using the MD simulations data (blue bars). The error bars represent one standard deviation over all simulation time steps. RMSD denotes the root-mean-square deviation of predicted and experimental shifts. (B) Deviation of predicted and experimental shifts per residue. Twenty evenly spaced bins between the minimum and maximum values were chosen to illustrate well the near Gaussian distribution of the data with distant outliers. The positions of the α-helices are indicated at the top.

The Flexible Site III Located at the N-terminus of the AB-loop of IL-6 rigidifies Upon IL-6Rα Binding. To test our hypothesis that interaction of IL-6Rα with hIL-6 regulates the dynamic state of residues that function in receptor assembly, we have used the MD simulations of hIL-6, the hIL-6/IL-6Rα-D2D3 complex, and vIL-6 to measure dynamic fluctuations of the backbone and changes in the spatial position of side chains.

Figure 3 shows a superposition of 100 evenly spaced frames from each simulation, providing a visual impression of the flexibility in different regions of IL-6, which is further encoded by coloring residues according to their RMSF. Residues located in the four-helix bundle in each simulation, as expected, show the lowest RMSFs, whereas loop regions show greater RMSFs

DOI: 10.1021/acs.jpcb.7b05732

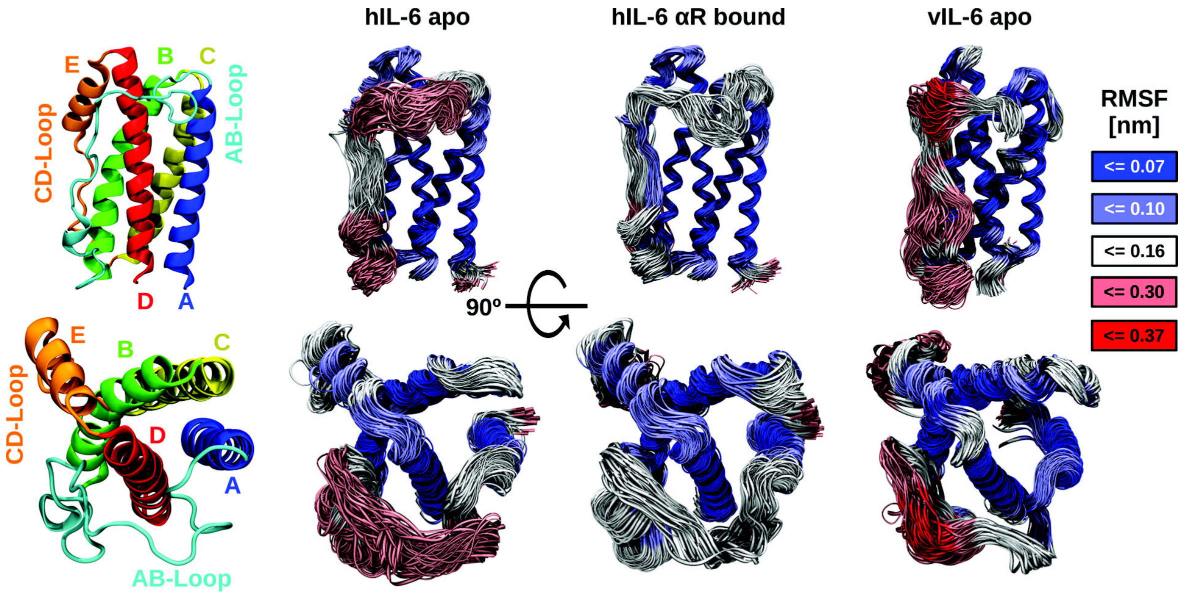


Figure 3. Color-coded IL-6 structure in two orientations is presented in the left column. Columns 2 to 4: superpositions (on helix backbone atoms) of evenly spaced 100 frames from each MD simulation. The color-coding represents the RMSF per residue in five discrete categories from blue to red. The upper bounds of each category are given in the legend, and the lower bounds are the upper bounds of the previous category. The lower panel shows a rotation of the molecule by 90° around the horizontal axis.

with residues in the long AB-loop (35 residues) fluctuating the most. Binding of the hIL-6Rα at site I leads to an expected modest reduction in the RMSFs for residues at the C-terminus of the AB-loop. Interestingly, however, is the observation that residues at the N-terminus of the AB-loop undergo larger restrictions in motions, indicating that binding of IL-6Rα to residues at the C-terminus of this loop restricts motion along the entire AB-loop. RMSFs for vIL-6 are the largest in the AB-loop, except for residues 45 to 57 that form the initial part of this loop and are part of the gp130-D1 interaction interface (i.e., site III).

The RMSFs per residue for the three simulations are plotted in Figure 4. All frames have been superimposed to minimize the backbone RMSD of the helices, and hence these regions have the lowest RMSF values. Consequently, an RMSF of ~0.5 Å is considered to be the highest rigidity achievable in a well-folded protein and represents an RMSF baseline. This RMSF is consistent with MD simulations of isolated α-helices with restrained backbone hydrogen bonds.[49] The largest RMSF across all three systems is observed for Ala59 of vIL-6 with a value of 3.7 Å. Starting at residue 42 (Figure 4B), two full turns before the end of helix A, the RMSFs of apo and α-receptor bound hIL-6 increase steadily, except for Cys51, which has limited mobility because it forms a disulfide bond with Cys45. Residues in the AB-loop of hIL-6 (residues 56 to 63) that comprise site IIIaA are flexible in the apo state, but are rigidified in the presence of hIL-6Rα, which binds these residues from the opposing side in the MD simulation, as explained below. The RMSFs for residues of apo vIL-6 indicate less flexibility than either apo or bound hIL-6 at the end of helix A and the N-terminal part of the AB-loop that comprises part of site III of vIL-6 (42–55; Figure 4, yellow shading). This lower flexibility in this region of vIL-6 compared with that of hIL-6 can be explained partly by the difference in hydrophobic contacts. Here, the much larger number of hydrophobic residues in the AB-loop of vIL-6 when compared with that of hIL-6 pack tightly with hydrophobic residues of the protein core (Figure 5). The number of contacts (averaged over all frames) between hydrophobic residues located in this region of the loop, i.e., residues 48 to 62, of vIL-6 (Ile50, Ile54, Leu55, Pro57, Ala58, Ala59, Ile60, and F61) during the apo-vIL-6 MD simulation was calculated to be 83.4 ± 12.1, whereas for hIL-6 (M50, A57, L58, and A59) this value was calculated to be 20.4 ± 9.1. This difference in hydrophobic packing confers rigidity to this region of the AB-loop of vIL-6 (Figure 4B). Hydrophobic contacts were also determined for LIF (residues 49 to 64) because this cytokine, like vIL-6, does not require an α-receptor to bind to site I for gp130-mediated signaling,[50] and experimental S^2 values revealed that the AB-loop is rigid.[51] Although, the

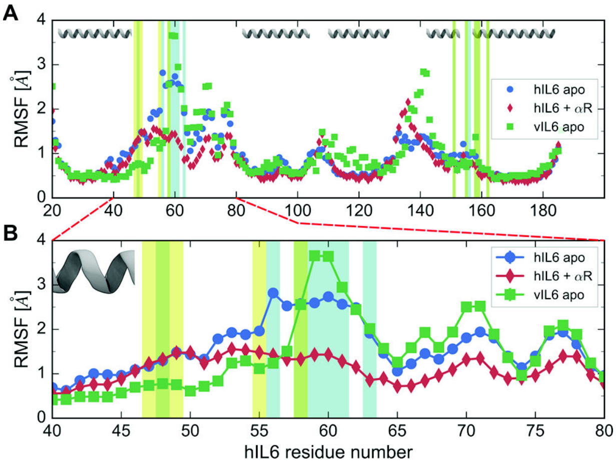


Figure 4. Flexibility of hIL-6 and vIL-6. (A) RMSFs for residues of apo hIL-6 (blue), apo vIL-6 (green), and hIL-6/IL-6Rα (red). (B) RMSFs for residues that constitute the AB-loop (residues 40–80; demarcated by the red dashed lines). The shaded areas denote binding site IIIa residues in hIL-6 (cyan), vIL-6 (yellow), and both (green).

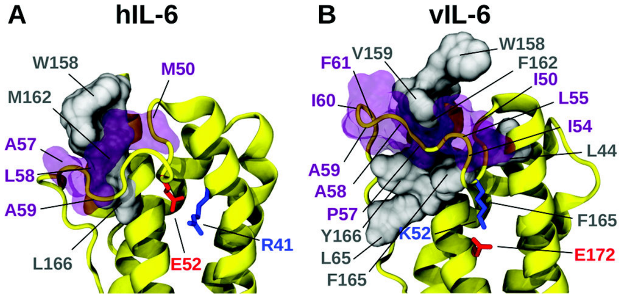


Figure 5. Interactions stabilizing the N-terminal part of the AB-loop for hIL-6 (A), vIL-6 (B), and LIF (C). Cytokines are shown in yellow cartoon representation. Hydrophobic residues of the N-terminal AB-loop (48−62) are depicted in transparent purple surface representation and hydrophobic residues in contact with these residues in the AB-loop are shown in gray space-filled representation. Residues that form a salt bridge at position 52 are shown as stick representations in hIL-6 and vIL-6.

number of hydrophobic contacts of 50 (taken from PDB ID: 1EMR; Figure 5C) was lower than that determined for vIL-6, this value is noticeably higher than that of hIL-6, suggesting that the hydrophobic packing of the N-terminal residues of the AB-loop in LIF explains the rigidity of this loop. This rigidity observed in LIF is likely to be a functionally relevant feature, because the LIF receptor recognizes site III of LIF via an immunoglobulin-like domain,[50] which is similar to the interaction between hIL-6 and gp130-D1, suggesting that the rigid N-terminal AB-loop of LIF precludes the requirement of receptor binding at site I to change the rigidity of site III.

From residue 58 the AB-loop of vIL-6 is more flexible, showing very high flexibility at Ala59 and Ile60 (numbered 45 and 46 in PDB ID: 1I1R). The remainder of the AB-loop is characterized by relatively large RMSFs for hIL-6 and vIL-6, with binding of IL-6R*α* to hIL-6 leading to noticeable reduction in RMSFs. Exceptions to this trend in both apo proteins are Leu65, whose rigidity is explained by hydrophobic packing toward helix D, and Cys74, which forms a disulfide bond with Cys84.

The IL-6R*α* Pivots Around Site I of IL-6 to Form Interactions with the N-terminal Region of the AB-loop that Are Crucial for Enabling gp130 Interactions at Site III. During the initial 15 ns of each of the three MD simulations of the hIL-6/IL-6R*α*-D2D3 complex, both the D2 and D3 domains of the IL-6R*α* pivot around site I of hIL-6 leading to interactions between the N-terminal region of the AB-loop (residues 55 to 70) of hIL-6 and the D2 domain of IL-6R*α* (Figure 6A). These interactions have not been observed in structural studies and explain a requirement for the formation of the IL-6/IL-6R*α* heterodimer prior to interaction with gp130. The average angles of this rotation are (17.5 ± 6.9)° for the D2 domain and (14.0 ± 4.9)° for the D3 domain. As this rotation is observed immediately in all three trajectories, it is not an artifact of the simulation, but rather a key feature necessary for the assembly of the signaling receptor complex. Based on these MD simulations, we propose that IL-6R*α* adopts the rotated state upon binding to IL-6 to facilitate a tight interaction with residues in the AB-loop (Figure 6B,C).

Subsequent binding of gp130-D1 at site III causes rotation of the IL-6R*α* to the position found in the structure of the signaling complex as a requirement for this interaction (i.e., to avoid steric clashes between IL-6R*α* and gp130-D1 if the D2 domain of IL-6R*α* did not release the AB-loop interactions), and this may be facilitated by binding of domains D2 and D3 of the second gp130 receptor at site II because the D3 domain of IL-6R*α* interacts with the D3 domain of the second gp130 through site IIb (Figure 1). Alternatively, binding of gp130-D2D3 at site II might occur before gp130-D1 binding at site III, as suggested by results using soluble receptor constructs and

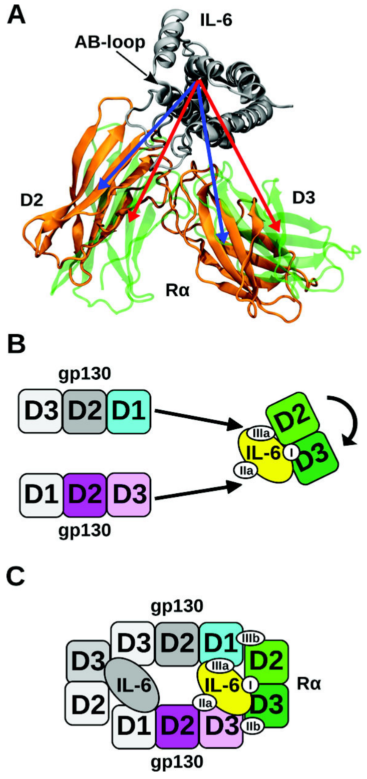


Figure 6. Proposed assembly mechanism of one heterotrimer of the hIL-6 receptor complex. (A) Illustration of the pivoting by ∼16° of IL-6R*α* (orange) around hIL-6 (gray) with respect to the position at the beginning of the simulation (green, transparent), which corresponds to the IL-6R*α* position in the hIL-6/IL-6R*α*/gp130 complex. The rotation angles of the D2 and D3 domains of IL-6R*α* are defined as the angles between vectors extending from the hIL-6 center of mass (COM) to the domain COM at the initial (red) and final time frames (blue), after superposition of the hIL-6 helix bundle. (B) IL-6R*α* is bound at site I in the rotated conformation in the preformed IL-6/IL-6R*α* heterodimer, with the D2 domain of the IL-6R*α* interacting with residues that are located at site IIIaA. Interactions with gp130-D1 at site IIIa and interactions between the D3 domains of IL-6R*α* and gp130 cause IL-6R*α* to rotate into the conformation observed in the hIL-6/IL-6R*α*/gp130 structure. (C) The final hexameric receptor assembly corresponding to Figure 1B.

isothermal titration calorimetry analysis,[16] or both gp130-D2D3 binding at site II and gp130-D1 binding at site III of a preformed membrane-tethered gp130 dimer[52,53] might bind simultaneously. Nonetheless, the pivoting mechanism remains in the three possible gp130 receptor binding scenarios.

IL-6R*α*-D2 possesses three binding interfaces for the AB-loop (Figure 7). Interface I (Figure 7B) is a predominantly hydrophilic stretch created by Thr186, Gln187, Thr188, Phe189, and Gln190, where the side chains of Gln187 and Phe189 are buried, and only the backbone is exposed on the surface. The most important contacts of this interface involve: (i) Ser54 of hIL-6 (67% of the time during the three independent MD simulations, mean interaction energy with interface 1 of (−13.9 ± 14.5) kJ/mol), which docks against the backbone of Phe189, and the backbone and the side chain of Thr188; and (ii) Lys55 of hIL-6 (96%, (−27.9 ± 22.2) kJ/mol), which interacts by undirected charge interactions with Thr186 and Thr188. The large standard deviations associated with the interaction energies are due to the interaction energy not being normally distributed, but rather continuous, sometimes multimodal distributions. Hence, these standard deviations do not imply that the attractive interactions become frequently repulsive; in fact they rarely do. In contrast, Lys55 of apo hIL-6 has an extremely flexible side chain (Figure S8) adopting conformations that are exclusively at the interface, which is relevant for binding with gp130-D1. In the apo state simulations, this residue either points toward Trp158 (hIL-6), a hot spot residue crucial in gp130-D1 interactions,[14,54] or masks three hydrophobic residues Ala57, Leu58, and Ala59, which are part of site IIIaA (Figures 1 and S2). Lys55 does not have a salt bridge partner, neither in the apo state simulation, nor in the receptor bound crystal structure, in which the Lys side chain points away from the site IIIa binding interface. The positive charge of Lys55 likely repels Lys46 of gp130-D1 and hampers gp130 binding in the apo state (Figures S8−S12). The interactions with IL-6R*α* interface I sequester Lys55 from the gp130-D1 binding interface to enable gp130 binding at site IIIa.

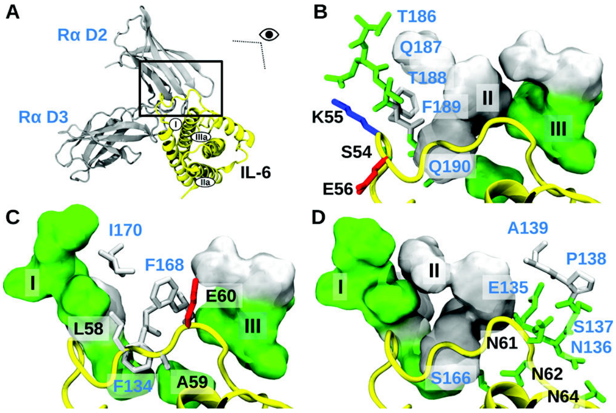


Figure 7. IL-6R*α*-D2 possesses three binding interfaces for the AB-loop of hIL-6. (A) An overview of the viewing angle in this figure. The eye icon indicates the viewing angle of the right IL-6 molecule in the hexameric assembly in Figure 1A. (B−D) Show the binding interfaces I−III of IL-6R*α* (surface representation) and IL-6 (yellow cartoon representation). IL-6R*α* residues are annotated in blue, IL-6 residues in black. Each panel shows one binding interface and the interfacing residues of IL-6 are presented in stick model representation to highlight the details of each interface.

Glu56, which is conserved between viral and human IL-6, forms only sporadic contact with interface I (50%, (−7.2 ± 11.9) kJ/mol), but its conformations are heavily influenced by IL-6R*α* binding, positioning the side chain toward the incoming gp130-D1 (Figures S9−S11). With an RMSF of 2.8 Å it is the most flexible residue in the MD simulation of apo hIL-6 and significantly more rigid (RMSF of 1.2 Å) in the vIL-6 simulation. Glu56 does not form salt bridges in the apo state in either the viral or human IL-6 MD simulations. Moreover, Glu56 does not form a salt bridge in the vIL-6 receptor bound crystal structure; however, it does form a salt bridge with Lys46 of gp130-D1 in the signaling complex structure. This interaction stabilizes the bound state with gp130 in hIL-6.

Binding interface II is entirely hydrophobic and constituted by the side chains of Phe134, Phe168, and Ile170 of IL-6R*α* (Figure 7C). These residues interact with Ala57 (45%, −2.9 ± 2.3 kJ/mol) of IL-6 by hydrophobic packing, but most importantly with hydrophobic contacts to the side chains of Leu58 (88%, (−12.9 ± 9.6) kJ/mol) and Ala59 (98%, (−15.1 ± 7.4) kJ/mol), which dock to Phe168 and Ile170. In addition, Phe134 positions the side chain of Glu60 (88% of the time in contact, (−12.6 ± 8.3) kJ/mol) toward the gp130-D1 domain binding interface, where it forms a salt bridge with Lys31 upon gp130-D1 domain binding (Figure S9−S11). Interestingly, in the receptor bound structure of the hIL-6/IL-6R*α*/gp130 complex, Phe134, Phe168, and Ile170 of IL-6R*α* interact with gp130 instead of hIL-6. The transition of interaction partners from hIL-6 to gp130-D1 happens during the rotation of IL-6R*α* from the position sampled predominantly in the MD simulations to the conformation found in the crystal structure complex (Figure 6).

Binding interface III (Figure 7D) is hydrophilic in nature and consists of residues Glu135, Asn136, Ser137, Pro138, Ala139, and Ser166. Interface III functions to stabilize the AB-loop by interacting with residues Asn61 (47%, (−10.8 ± 17.2) kJ/mol), Asn62 (88%, (−44.1 ± 25.0) kJ/mol), the backbone of Leu63 (82%, (−9.2 ± 8.1) kJ/mol), and Asn64 (81%, (−27.4 ± 25.8) kJ/mol) of hIL-6. Moreover, Asn61 and Leu63 form part of site IIIaA, suggesting that interaction with IL-6R*α* reduces their positional flexibility to facilitate their interaction with gp130-D1.

In summary, the interactions with the three binding interfaces of IL-6R*α* rigidify residues 56 to 64 in the AB-loop and reduce their RMSFs from ∼2.5 to ∼1.5 Å. The stabilized N-terminal region of the AB-loop adopts conformations that sequester Lys55 from the gp130-D1 binding interface, and reorient Glu56 and Glu60 toward the binding interface where they engage in salt bridges with gp130-D1. In the hIL-6 crystal structure, Glu56 interacts with Lys29 and Lys31 of gp130, and Glu60 with Lys46 of gp130, but it is also possible that Glu60 interacts with His49 of gp130. Thus, the proper binding pose of the D1 domain is favored by the directional interactions of the salt bridges formed by Glu56 and Glu60.

Viral IL-6 gp130-D1 Binding at Site IIIa Is Dominated by Stronger Hydrophobic Effects and Greater Rigidity of the Binding Interface in the Apo State When Compared with that of hIL-6. Binding of gp130-D1 at site IIIa involves different interactions in the human (Figure S12) and viral variants of IL-6 (Figure S13).[16,21] In both species, binding of IL-6 to gp130-D1 at site IIIa is dominated by a single conserved tryptophan (Trp158).[14,54] The hydrophobic surface area of site IIIa was determined to be not optimally packed in the crystal structures of the complexes and therefore short MD simulations of 1 ns were performed to relax side chain orientations. These MD simulations revealed a burial of solvent accessible hydrophobic surface area because of gp130-D1 binding of (284 ± 14) Å^2 for human and (339 ± 16) Å^2 for viral IL-6 (uncertainties denote one standard error of the mean over 51 simulation snapshots). Hence, hydrophobicity plays a smaller role in hIL-6 binding, as it contains a smaller hydrophobic surface area at site IIIa, as previously reported.[21]

Binding site IIIaA is large, extending from residue 45 to 63 of IL-6, with contributions from residues at site IIIaB, interacting with residues of the N-terminus of gp130. The conformation of the initial segment of the AB-loop is stabilized by a conserved disulfide bond between Cys45 and Cys51, and a charged residue at position 52 (Glu in hIL-6, Lys in vIL-6), anchoring the loop to the helix bundle by a salt bridge to Arg41 in helix A in hIL-6 and to Glu172 in helix D in vIL-6. The C_α−C_α distance of the residues forming the salt bridge is (10.8 ± 0.9) Å in human and (12.5 ± 0.4) Å in the vIL-6 simulation. Due to the longer C_α-C_α distance and the shorter side chain of lysine in vIL-6 compared to arginine in hIL-6, Lys52 is forced to maintain an extended side chain conformation (stable throughout the complete MD simulation of vIL-6, i.e. closed >95% of the time), which reduces its flexibility and rigidifies the AB-loop until Glu56 (Figure 5B). This rigidification effect is enhanced significantly by the presence of the well packed hydrophobic region of the AB-loop, as aforementioned, and both contributions together account for the lower flexibility of the N-terminal part of the AB-loop of vIL-6 when compared with that of apo hIL-6. In contrast, Glu52 in hIL-6 displays much greater conformational freedom in the intact salt bridge because of the shorter distance to Arg41. Combined with the smaller hydrophobic contact area (Figure 5A) this explains the larger flexibility of this part of the AB-loop as reflected in the RMSF data for residues 52 to 56 (Figure 4; 0.8 to 1.3 Å for vIL-6 versus 1.8 to 2.8 Å for hIL-6).

In vIL-6, the backbone carbonyl oxygens of the C-terminal part of helix A (Cys45, Tyr46, and Arg47 in vIL-6; Cys45, Asn46, and Lys47 in hIL-6) act as hydrogen bond acceptors to gp130. Hydrogen bond donors are His49 and Gln78 from gp130-D1. These IL-6 residues are rigid in vIL-6 with RMSFs of 0.6, 0.7, and 0.7 Å, respectively (Figure 4), whereas in hIL-6, the corresponding residues are more flexible with RMSFs of 1.0 Å (Cys45), 1.1 Å (Asn46), and 1.2 Å (Lys47). The following three residues (Ser48, Asn49, Met50) are not resolved in the human IL-6 receptor bound crystal structure, indicating large flexibility or conformational polymorphism at this location. The RMSFs in the hIL-6 apo MD simulation are elevated (1.3 to 1.5 Å) when compared with the values derived from the MD simulation of vIL-6 (<0.8 Å), but not completely unordered. Hence a more rigid binding interface for gp130-D1 is provided by the C-terminal part of helix A and the N-terminal part of the AB-loop of vIL-6 when compared with that of hIL-6, and this greater rigidity likely yields a smaller entropic cost upon gp130-D1 binding, thereby eliminating the requirement of IL-6Rα binding to provide the necessary enthalpy−entropy compensation that arises for the hIL-6/IL-6Rα interaction.

Residues 57 to 63 are flexible in human and viral IL-6, with RMSF values ≥2.5 Å. The only exceptions are Pro57 in vIL-6 with an RMSF of 1.5 Å and Leu63 (conserved), which is still flexible with an RMSF of 2 Å. Together with three residues from site IIIaB from the CD-loop (hIL-6: Leu152, Gln155, Leu159; vIL-6: Leu152, Leu155, Val159), residues 57 to 63 create a hydrophobic binding pocket for interaction with the N-terminus of gp130-D1 (residues 2 to 5). The binding pocket is much more hydrophobic in vIL-6 (8 hydrophobic residues: Ala58, Ala59, Ile60, Phe61, Leu63, Leu152, Leu155, and Val159) than in human (5 hydrophobic residues: Ala57, Leu58, Ala59, Leu63, Leu152, Leu159). This hydrophobic patch is interrupted in hIL-6 by Glu60 and by His62 in vIL-6. Both residues can form salt bridges with gp130: Glu60 with Lys31, which positions downward from the small helix of gp130-D1, whereas His62 directly interacts with Asp4 at the N-terminus of gp130. Asp4 can also bind to Lys29 of gp130 and hence His62 competes with Lys29 for interaction with Asp4.

The main difference of site IIIaB in hIL-6 is a potential hydrogen bond between Asn156 and Glu12 of gp130, whereas vIL-6 can form a further reaching salt bridge with Lys156 at this position (Figures S12A and S13A).

Remote Effect of the BC-Loop (Site IIIaC) on gp130-D1 Binding. The previous finding using chimeric IL-6 proteins that site IIIaC is required for IL-6Rα independent receptor activation of vIL-6 is surprising,[22] as this region of viral or human IL-6 does not make any direct contacts with gp130 in the IL-6 signaling complex structures. However, as shown in Figure S2, Arg105 at site IIIaC in human IL-6 forms a salt bridge with Asp161 at site IIIaB. When site IIIaB from viral IL-6 is introduced into human IL-6 to create an IL-6Rα independent chimera, the negative charge of Asp161, which is adjacent to the site IIIa binder Met162 in hIL-6 (Phe162 in vIL-6), is replaced by two positively charged residues Arg160 and His161. The repulsion with Arg105 may disrupt site IIIa locally and interfere with gp130-D1 binding. By introducing site IIIaC from viral in combination with IIIaB, the positive charge of Arg105 is replaced by a negatively charged glutamate, which can form a stabilizing salt bridge with the arginine and, to a lesser extent, with the histidine from the viral site IIIaB. Therefore, we hypothesize that the only crucial amino acid exchange required in site IIIaC is the replacement of Arg105 with a negatively charged residue when creating the chimeric IL-6.

CONCLUSIONS

In this study, IL-6Rα was shown to interact with and rigidify the N-terminus of the AB-loop of hIL-6 (Figures 4 and 7). This rigidification, together with a repositioning of three key residues (Lys55, Glu56, and Glu60), promotes gp130-D1 binding at site IIIa, which we postulate simultaneously pivots IL-6Rα around the helix bundle to the state found in the hIL-6/IL-6Rα/gp130-D1 complex (Figure 6). Although reduced flexibility facilitates assembly of this signaling complex, this is not a general mechanism observed in nature. For example, in contrast to rigidification of regions of IL-6 by IL-6Rα facilitating signaling complex formation, MD studies of the dock-lock mechanism of Aβ-peptide oligomer formation have shown that partial disordering of the Aβ-peptide oligomer kinetically favors docking of an Aβ monomer.[55,56] This apparent opposing requirement of greater flexibility to facilitate Aβ-peptide assemblies highlights that the extent of protein or peptide motion is likely to be finely tuned to guide biological processes.

The MD simulations of the hIL-6/IL-6Rα-D2D3 complex are initiated from the ternary complex state rather than starting from a docked hIL-6/IL-6Rα-D2D3 model. Thus, the simulated pivoting of the IL-6Rα (Figure 6), while representing a large conformational rearrangement of the binary complex in detail, does not report on the formation of the binary complex between IL-6 and IL-6Rα, but reports on the final conformational changes that occur in the ternary complex assembly.

DOI: 10.1021/acs.jpcb.7b05732

Investigating the formation of the IL-6/IL-6Rα complex is warranted, but beyond the scope of this work because of the likelihood of numerous starting structures of the binary complex combined with the requirement of long MD simulations to accommodate the time scale of the dynamics to reach equilibrium. In such a study, docking of the hIL-6 and IL-6Rα followed by MD simulations would provide data describing binary complex formation followed by conformational fluctuations that we hypothesize should lead to a complex similar to the one observed after the MD simulations presented herein (i.e., conformation observed at the end of the simulation; Figure 6A).

The gp130-D1 binding mechanism at site IIIa differs between human and viral IL-6. The site IIIa interface of vIL-6 is characterized by a stronger hydrophobicity. The N-terminus of helix A and the beginning of the AB-loop up to residue 57 are more rigid than in human (Figures 3 and 4) because of extensive hydrophobic contacts with the protein core (Figure 5), and provide a well-defined binding interface in the apo state. The hydrophobic residues from position 58 onward are extremely flexible and solvent exposed in apo vIL-6, which enforces an entropically unfavorable ordering of the solvent. When gp130-D1 binds, this region of the AB-loop is rigidified by the N-terminus of gp130 binding and its hydrophobic solvent accessible surface reduced. His62 engages in a salt bridge with Asp4 of gp130-D1 and the additional salt bridge (compared with hIL-6) from Lys156 (vIL-6) to Glu12 (gp130-D1) stabilizes the binding further. In hIL-6, Glu56 and Glu60 play an important role in gp130-D1 binding. These two residues first need to be oriented by IL-6Rα into positions that point toward the incoming gp130-D1 domain. According to the crystal structure of vIL-6 bound to gp130, Glu56 is not involved in binding, and hence IL-6Rα induced reorientation is not crucial, as it is for hIL-6. This leaves an unanswered question to what the function of Glu56 is in vIL-6, because it is conserved across species despite the overall low sequence identity.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jpcb.7b05732.

> Parameters used in the MD simulations and RMSD of chemical shift predictors, the sequence alignment of viral and human IL-6, the three site IIIa subsites in IL-6, superposition of the hIL-6 and vIL-6 structures, comparison of experimental and MD-derived backbone C′, N and H_N chemical shifts, comparison of S^2 values for backbone amide groups of hIL-6 determined experimentally and from the MD simulations, structural representation of binding interface I of hIL-6/IL-6Rα, binding interfaces and distributions of key residues in the site IIIa interface of hIL-6 from multiple perspectives, and structural representation of IL-6 binding interfaces with gp130-D1 (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*Phone: ++49 2461 61 9487; Fax: ++49 2461 61 2023; E-mail: a.dingley@fz-juelich.de.

ORCID

Birgit Strodel: 0000-0002-8734-7765

Andrew J. Dingley: 0000-0002-6838-5803

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Financial support by the Deutsche Forschungsgemeinschaft through funds INST 208/704-1 FUGG to purchase the hybrid computer cluster used in this study is gratefully acknowledged.

■ REFERENCES

(1) Hirano, T.; Yasukawa, K.; Harada, H.; Taga, T.; Watanabe, Y.; Matsuda, T.; Kashiwamura, S.; Nakajima, K.; Koyama, K.; Iwamatsu, A.; et al. Complementary DNA for a Novel Human Interleukin (BSF-2) That Induces B Lymphocytes to Produce Immunoglobulin. *Nature* **1986**, *324*, 73−76.

(2) Kishimoto, T.; Hirano, T. Molecular Regulation of B Lymphocyte Response. *Annu. Rev. Immunol.* **1988**, *6*, 485−512.

(3) Arai, K.; Tsuruta, L.; Watanabe, S.; Arai, N. Cytokine Signal Networks and a New Era in Biomedical Research. *Mol. Cells* **1997**, *7*, 1−12.

(4) Schaper, F.; Rose-John, S. Interleukin-6: Biology, Signaling and Strategies of Blockade. *Cytokine Growth Factor Rev.* **2015**, *26*, 475−487.

(5) Scheller, J.; Chalaris, A.; Schmidt-Arras, D.; Rose-John, S. The Pro- and Anti-Inflammatory Properties of the Cytokine Interleukin-6. *Biochim. Biophys. Acta, Mol. Cell Res.* **2011**, *1813*, 878−888.

(6) Yamamoto, K.; Rose-John, S. Therapeutic Blockade of Interleukin-6 in Chronic Inflammatory Disease. *Clin. Pharmacol. Ther.* **2012**, *91*, 574−576.

(7) Scheller, J.; Garbers, C.; Rose-John, S. Interleukin-6: From Basic Biology to Selective Blockade of Pro-Inflammatory Activities. *Semin. Immunol.* **2014**, *26*, 2−12.

(8) Lin, M.; Rose-John, S.; Grotzinger, J.; Conrad, U.; Scheller, J. Functional Expression of a Biologically Active Fragment of Soluble gp130 as an ELP-Fusion Protein in Transgenic Plants: Purification Via Inverse Transition Cycling. *Biochem. J.* **2006**, *398*, 577−583.

(9) Somers, W.; Stahl, M.; Seehra, J. S. 1.9 Å Crystal Structure of Interleukin 6: Implications for a Novel Mode of Receptor Dimerization and Signaling. *EMBO J.* **1997**, *16*, 989−997.

(10) Xu, G.-Y.; Yu, H.-A.; Hong, J.; Stahl, M.; McDonagh, T.; Kay, L. E.; Cumming, D. A. Solution Structure of Recombinant Human Interleukin-6. *J. Mol. Biol.* **1997**, *268*, 468−481.

(11) Aasland, D.; Oppmann, B.; Grotzinger, J.; Rose-John, S.; Kallen, K. J. The Upper Cytokine-Binding Module and the Ig-Like Domain of the Leukaemia Inhibitory Factor (LIF) Receptor Are Sufficient for a Functional LIF Receptor Complex. *J. Mol. Biol.* **2002**, *315*, 637−646.

(12) Ehlers, M.; Grötzinger, J.; deHon, F. D.; Mullberg, J.; Brakenhoff, J. P.; Liu, J.; Wollmer, A.; Rose-John, S. Identification of Two Novel Regions of Human IL-6 Responsible for Receptor Binding and Signal Transduction. *J. Immunol.* **1994**, *153*, 1744−1753.

(13) Ozbek, S.; Grötzinger, J.; Krebs, B.; Fischer, M.; Wollmer, A.; Jostock, T.; Mullberg, J.; Rose-John, S. The Membrane Proximal Cytokine Receptor Domain of the Human Interleukin-6 Receptor Is Sufficient for Ligand Binding but Not for gp130 Association. *J. Biol. Chem.* **1998**, *273*, 21374−21379.

(14) Paonessa, G.; Graziani, R.; De Serio, A.; Savino, R.; Ciapponi, L.; Lahm, A.; Salvati, A. L.; Toniatti, C.; Ciliberto, G. Two Distinct and Independent Sites on IL-6 Trigger gp 130 Dimer Formation and Signalling. *EMBO J.* **1995**, *14*, 1942−1951.

(15) Behrmann, I.; Wallner, S.; Komyod, W.; Heinrich, P. C.; Schuierer, M.; Buettner, R.; Bosserhoff, A. K. Characterization of Methylthioadenosin Phosphorylase (MTAP) Expression in Malignant Melanoma. *Am. J. Pathol.* **2003**, *163*, 683−690.

(16) Boulanger, M. J.; Chow, D.-c.; Brevnova, E. E.; Garcia, K. C. Hexameric Structure and Assembly of the Interleukin-6/IL-6 α-Receptor/gp130 Complex. *Science* **2003**, *300*, 2101−2104.

(17) Hunter, C. A.; Jones, S. A. IL-6 as a Keystone Cytokine in Health and Disease. *Nat. Immunol.* **2015**, *16*, 448−457.

(18) Neipel, F.; Albrecht, J. C.; Ensser, A.; Huang, Y. Q.; Li, J. J.; Friedman-Kien, A. E.; Fleckenstein, B. Human Herpesvirus 8 Encodes a Homolog of Interleukin-6. *J. Virol.* **1997**, *71*, 839−842.

(19) Chow, D.; He, X.; Snow, A. L.; Rose-John, S.; Garcia, K. C. Structure of an Extracellular gp130 Cytokine Receptor Signaling Complex. *Science* **2001**, *291*, 2150−2155.

(20) Hoischen, S. H.; Vollmer, P.; Marz, P.; Ozbek, S.; Gotze, K. S.; Peschel, C.; Jostock, T.; Geib, T.; Mullberg, J.; Mechtersheimer, S.; et al. Human Herpes Virus 8 Interleukin-6 Homologue Triggers gp130 on Neuronal and Hematopoietic Cells. *Eur. J. Biochem.* **2000**, *267*, 3604−3612.

(21) Boulanger, M. J.; Chow, D. C.; Brevnova, E.; Martick, M.; Sandford, G.; Nicholas, J.; Garcia, K. C. Molecular Mechanisms for Viral Mimicry of a Human Cytokine: Activation of gp130 by HHV-8 Interleukin-6. *J. Mol. Biol.* **2004**, *335*, 641−654.

(22) Adam, N.; Rabe, B.; Suthaus, J.; Grotzinger, J.; Rose-John, S.; Scheller, J. Unraveling Viral Interleukin-6 Binding to gp130 and Activation of Stat-Signaling Pathways Independently of the Interleukin-6 Receptor. *J. Virol.* **2009**, *83*, 5117−5126.

(23) Henzler-Wildman, K. A.; Thai, V.; Lei, M.; Ott, M.; Wolf-Watz, M.; Fenn, T.; Pozharski, E.; Wilson, M. A.; Petsko, G. A.; Karplus, M.; et al. Intrinsic Motions Along an Enzymatic Reaction Trajectory. *Nature* **2007**, *450*, 838−844.

(24) Dror, R. O.; Dirks, R. M.; Grossman, J.; Xu, H.; Shaw, D. E. Biomolecular Simulation: A Computational Microscope for Molecular Biology. *Annu. Rev. Biophys.* **2012**, *41*, 429−452.

(25) Bobby, R.; Robustelli, P.; Kralicek, A. V.; Mobli, M.; King, G. F.; Grotzinger, J.; Dingley, A. J. Functional Implications of Large Backbone Amplitude Motions of the Glycoprotein 130-Binding Epitope of Interleukin-6. *FEBS J.* **2014**, *281*, 2471−2483.

(26) Biasini, M.; Bienert, S.; Waterhouse, A.; Arnold, K.; Studer, G.; Schmidt, T.; Kiefer, F.; Cassarino, T. G.; Bertoni, M.; Bordoli, L.; Schwede, T. Swiss-Model: Modelling Protein Tertiary and Quaternary Structure Using Evolutionary Information. *Nucleic Acids Res.* **2014**, *42*, W252−W258.

(27) Pronk, S.; Páll, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; et al. Gromacs 4.5: A High-Throughput and Highly Parallel Open Source Molecular Simulation Toolkit. *Bioinformatics* **2013**, *29*, 845−854.

(28) Li, D.-W.; Brüschweiler, R. NMR-Based Protein Potentials. *Angew. Chem.* **2010**, *122*, 6930−6932.

(29) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79*, 926−935.

(30) Darden, T.; Perera, L.; Li, L.; Pedersen, L. New Tricks for Modelers from the Crystallography Toolkit: The Particle Mesh Ewald Algorithm and Its Use in Nucleic Acid Simulations. *Structure* **1999**, *7*, R55−R60.

(31) Bussi, G.; Donadio, D.; Parrinello, M. Canonical Sampling through Velocity Rescaling. *J. Chem. Phys.* **2007**, *126*, 014101.

(32) Berendsen, H. J.; Postma, J. v.; van Gunsteren, W. F.; DiNola, A.; Haak, J. Molecular Dynamics with Coupling to an External Bath. *J. Chem. Phys.* **1984**, *81*, 3684−3690.

(33) Hoover, W. G. Canonical Dynamics: Equilibrium Phase-Space Distributions. *Phys. Rev. A: At., Mol., Opt. Phys.* **1985**, *31*, 1695−1697.

(34) Parrinello, M.; Rahman, A. Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method. *J. Appl. Phys.* **1981**, *52*, 7182−7190.

(35) Hess, B. P-Lincs: A Parallel Linear Constraint Solver for Molecular Simulation. *J. Chem. Theory Comput.* **2008**, *4*, 116−122.

(36) Shrake, A.; Rupley, J. Environment and Exposure to Solvent of Protein Atoms. Lysozyme and Insulin. *J. Mol. Biol.* **1973**, *79*, 351−364.

(37) Sievers, F.; Wilm, A.; Dineen, D.; Gibson, T. J.; Karplus, K.; Li, W.; Lopez, R.; McWilliam, H.; Remmert, M.; Söding, J.; Thompson, J. D.; Higgins, D. G. Fast, Scalable Generation of High-Quality Protein Multiple Sequence Alignments Using Clustal Omega. *Mol. Syst. Biol.* **2011**, *7*, 539.

(38) UniProt Consortium. Uniprot: The Universal Protein Knowledgebase. *Nucleic Acids Res.* **2017**, *45*, D158−D169.

(39) Han, B.; Liu, Y.; Ginzinger, S. W.; Wishart, D. S. ShiftX2: Significantly Improved Protein Chemical Shift Prediction. *J. Biomol. NMR* **2011**, *50*, 43−57.

(40) McGibbon, R. T.; Beauchamp, K. A.; Harrigan, M. P.; Klein, C.; Swails, J. M.; Hernández, C. X.; Schwantes, C. R.; Wang, L.-P.; Lane, T. J.; Pande, V. S. Mdtraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **2015**, *109*, 1528−1532.

(41) Shen, Y.; Bax, A. Sparta+: A Modest Improvement in Empirical NMR Chemical Shift Prediction by Means of an Artificial Neural Network. *J. Biomol. NMR* **2010**, *48*, 13−22.

(42) Trbovic, N.; Kim, B.; Friesner, R. A.; Palmer, A. G. Structural Analysis of Protein Dynamics by MD Simulations and NMR Spin-Relaxation. *Proteins: Struct., Funct., Genet.* **2008**, *71*, 684−694.

(43) Case, D. A. Calculations of NMR Dipolar Coupling Strengths in Model Peptides. *J. Biomol. NMR* **1999**, *15*, 95−102.

(44) Sengupta, D.; Kundu, S. Role of Long- and Short-Range Hydrophobic, Hydrophilic and Charged Residues Contact Network in Protein's Structural Organization. *BMC Bioinf.* **2012**, *13*, 142.

(45) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graphics* **1996**, *14*, 33−38.

(46) Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **2007**, *9*, 90−95.

(47) Waskom, M. *Seaborn: Statistical Data Visualization*, version 0.6.0; New York University: New York City, NY, 2015.

(48) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters. *Proteins: Struct., Funct., Genet.* **2006**, *65*, 712−725.

(49) Allison, J. R.; Müller, M.; van Gunsteren, W. F. A Comparison of the Different Helices Adopted by α-and β-Peptides Suggests Different Reasons for Their Stability. *Protein Sci.* **2010**, *19*, 2186−2195.

(50) Boulanger, M. J.; Bankovich, A. J.; Kortemme, T.; Baker, D.; Garcia, K. C. Convergent Mechanisms for Recognition of Divergent Cytokines by the Shared Signaling Receptor gp130. *Mol. Cell* **2003**, *12*, 577−589.

(51) Yao, S.; Smith, D. K.; Hinds, M. G.; Zhang, J. G.; Nicola, N. A.; Norton, R. S. Backbone Dynamics Measurements on Leukemia Inhibitory Factor, a Rigid Four-Helical Bundle Cytokine. *Protein Sci.* **2000**, *9*, 671−682.

(52) Schroers, A.; Hecht, O.; Kallen, K. J.; Pachta, M.; Rose-John, S.; Grotzinger, J. Dynamics of the gp130 Cytokine Complex: A Model for Assembly on the Cellular Membrane. *Protein Sci.* **2005**, *14*, 783−790.

(53) Tenhumberg, S.; Schuster, B.; Zhu, L.; Kovaleva, M.; Scheller, J.; Kallen, K. J.; Rose-John, S. gp130 Dimerization in the Absence of Ligand: Preformed Cytokine Receptor Complexes. *Biochem. Biophys. Res. Commun.* **2006**, *346*, 649−657.

(54) Barton, V. A.; Hudson, K. R.; Heath, J. K. Identification of Three Distinct Receptor Binding Sites of Murine Interleukin-11. *J. Biol. Chem.* **1999**, *274*, 5755−5761.

(55) Kouza, M.; Co, N. T.; Nguyen, P. H.; Kolinski, A.; Li, M. S. Preformed Template Fluctuations Promote Fibril Formation: Insights from Lattice and All-Atom Models. *J. Chem. Phys.* **2015**, *142*, 145104.

(56) Nguyen, P. H.; Li, M. S.; Stock, G.; Straub, J. E.; Thirumalai, D. Monomer Adds to Preformed Structured Oligomers of Aβ-Peptides by a Two-Stage Dock-Lock Mechanism. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 111−116.

DOI: 10.1021/acs.jpcb.7b05732

3.4 Integrated NMR, Fluorescence and MD Benchmark Study of Protein Mechanics and Hydrodynamics

3.4.1 Summary

A protein's function depends critically on its dynamics [99]. While protein structures have been solved for a long time with many methods, a complete picture of protein dynamics is much harder to obtain. In this research, three methods that are inherently suited for the study of macromolecular dynamics are combined in a coherent theoretical framework: Fluorescence spectroscopy (section 2.1.4), NMR spectroscopy (section 2.1.2) and MD simulations (section 2.2.1). The theory behind the nature of S^2 order parameters, obtained from NMR spectroscopy, and the fluorescence anisotropy measured in fluorescence spectroscopy is essentially constructed from the same mathematical tools and hence very similar. We compared data about the dynamics of the protein GABARAP (section 2.3.2) obtained by both experimental methods to corresponding quantities computed from extensive MD simulations. We demonstrated that MD is a valuable tool, aiding the interpretation of data from both methods and explaining their discrepancies. In particular, MD can help to select optimal positions for the attachment of the fluorescent dyes to amino acids.

3.4.2 Contribution

I analysed the MD simulations and discovered the optimal method for the calculation of S^2 order parameters from MD trajectories, for both amino acid side chains and backbone N–H bond vectors. Furthermore, I compared the order parameters from the MD simulations to those derived from experimental data and explained discrepancies and interesting features.

I wrote the methods section on MD simulation for the manuscript, as well as parts of the results section (together approximately 10 %) and I created approximately 40 % of the figures.

3.4.3 Publication

This section contains a complete reprint of the manuscript, which is has been submitted to the Journal of Physical Chemistry B. The supporting information to this article is located in section 6.4.

Integrated NMR, Fluorescence and MD Benchmark Study of Protein Mechanics and Hydrodynamics

Christina Möller,[1,2] Jakub Kubiak,[3#] Oliver Schillinger,[2,4#] Ralf Kühnemuth,[3] Dennis Della Corte,[2] Gunnar Schröder,[1,2] Dieter Willbold,[1,2] Birgit Strodel,[2,4*] Claus A. M. Seidel,[3*] and Philipp Neudecker[1,2*]

[1]Institut für Physikalische Biologie and BMFZ, Heinrich-Heine-Universität Düsseldorf, Germany

[2]Institute of Complex Systems: Structural Biochemistry, Forschungszentrum Jülich, Germany

[3]Lehrstuhl für Molekulare Physikalische Chemie, Heinrich-Heine-Universität Düsseldorf, Germany

[4]Institut für Theoretische Chemie und Computerchemie, Heinrich-Heine-Universität Düsseldorf, Germany

# Contributed equally.

* Corresponding authors.

Abstract

Understanding the mechanisms of protein function usually requires knowledge of its tertiary structure and conformational dynamics at preferentially atomic resolution. Nuclear magnetic resonance (NMR), polarization resolved fluorescence spectroscopy and molecular dynamics (MD) simulations are powerful methods to provide detailed insight into structural dynamics on multiple timescales. Using these techniques, we present an integrated approach to study the dynamics of the autophagy-related protein GABARAP on the pico- to nanosecond timescale together with its rotational and translational diffusion. We determine order parameters and corresponding relaxation times and examine the accuracy of our methods by benchmarking them against each other. All data are analyzed and interpreted within a joint general theoretical description of rotational and translational diffusion with a consistent nomenclature demonstrating the conceptual similarities. We compare GABARAP's fast dynamics determined by ^{15}N spin relaxation in the backbone, fluorescence anisotropy decays and fluorescence correlation spectroscopy of side chains labeled with BODIPY FL, and MD simulations providing movies of the protein motions. We consistently find a global rotational correlation of GABARAP of 8.2 ns ± 0.5 ns at 25°C in sub-µM aqueous solution, which corresponds to a hydrodynamic radius of 20.8 Å ± 0.1 Å including a hydration shell of ≈ 3 Å. Moreover, we identify dynamic motifs based on low order parameters of the N- and C-termini and a loop at I41 and present a method to identify potential hinges, which can lead to large-scale and functionally relevant motions, based on the S^2 backbone order parameters. Residues 27 and 28 of GABARAP are predicted to be such a hinge for its two sub-domains. These findings are a rationale for GABARAP self-assembly and enzymatic processing for membrane anchoring, which are necessary events for autophagy and possibly apoptosis. We conclude that the presented integrated concept is fundamental for characterizing multi-scale protein flexibilities.

1. Introduction

Conformational dynamics is a prerequisite for proteins to be able to fold and exert their physiological functions. Understanding the physicochemical mechanisms underlying folding and function of a protein therefore requires knowledge of both structure as well as dynamics in atomic detail. Protein conformational dynamics is observed on a wide range of time-scales [1]. On the most fundamental level, polypeptide chains in aqueous solution at ambient temperature are highly mobile on the pico- to nanosecond time-scale unless these motions are restricted by stable tertiary interactions. In globular proteins with a stable tertiary fold the residual internal dynamics on the time-scale such as side chain rotations, hydrogen bond formation, and backbone motions in loops and at the termini can be separated conceptually from the overall rotational diffusion of the molecule, which is governed by the hydrodynamic properties of the protein. If the protein is approximately spherical in shape rotational diffusion is isotropic and described by a single global rotational autocorrelation time ρ_{global}, which is typically of the order of several nanoseconds for moderately sized proteins and approximately proportional to the viscosity of the solvent and the molecular weight of the solute according to the Stokes-Einstein law [2]. By contrast, processes such as protein folding or allosteric regulation that involve significant rearrangement of larger secondary structure elements, subdomains or even entire domains relative to each other are often limited by the rate of intra-chain diffusion and by sizeable activation barriers and thus most commonly observed on the micro- to millisecond time-scale or slower.

For a variety of reasons a detailed knowledge of the ps to ns dynamics of a protein is a key element in understanding its structural biology. First, identification of well-ordered and disordered regions of the polypeptide chain is an important step in the high-resolution structure determination by X-ray crystallography and/or NMR spectroscopy because the degree of disorder has serious implications for the interpretation of the experimental data (e. g., lack of crystal formation, electron density map, nuclear Overhauser effect (NOE) restraints) as well as of the resulting structural model (e. g., crystal packing artifacts, precision and accuracy, plasticity). Second, ps to ns dynamics is a major contribution to the conformational entropy of a protein [3-5], and hence to the thermodynamics of biochemical processes such as protein folding, ligand binding, allosteric regulation, enzyme catalysis etc. Moreover, the hydrodynamic properties governing the rotational diffusion are a sensitive probe of the overall shape and oligomerization state of the protein [6] and of environmental effects such as subcellular localization [7]. Most importantly, conformational dynamics on the ps to ns time-scale is often required for function, e. g. to make sites of interest readily accessible for ligand binding [8] posttranslational modifications such as phosphorylation [9] or lipidation [10], or protease degradation [11].

NMR and fluorescence spectroscopy are the most widely used methods providing experimental insight into dynamics on virtually all relevant time-scales from picoseconds to days [1, 12]. Molecular dynamics (MD) simulations, on the other hand, are the only method describing protein dynamics directly at atomic resolution [13] and can nowadays reach micro- and even milliseconds [14], thereby allowing interpretation of experimental data from NMR and fluorescence on the structural level [15, 16].

In this work, we focus on the pico- to nanosecond dynamics of the multifunctional autophagy-related protein GABARAP. The 117-residue GABA$_A$ receptor-associated protein (GABARAP) from *H. sapiens* is a versatile key regulator in autophagy. GABARAP was initially identified as an interaction partner of GABA$_A$ receptors [17]. Further studies revealed that GABARAP interacts with the cytoskeleton through tubulin binding [18], and is implicated in receptor trafficking to the plasma membrane [19]. GABARAP belongs to the ubiquitin-like modifiers and its tertiary structure comprises the C-terminal ubiquitin-like subdomain (ULD) preceded by an N-terminal helical subdomain (NHD) consisting of helices α_1 and α_2, an arrangement that exposes two hydrophobic ligand binding pockets on the molecular surface [10]. It is a cytosolic protein ubiquitously expressed in most tissues and primarily localized to the Golgi apparatus, the endoplasmatic reticulum, and transport vesicles [20]. In addition to its soluble cytosolic forms, GABARAP can also be membrane-anchored via covalent coupling of a phospholipid moiety to G116 by a ubiquitin-like conjugation system [10]. In order for this to occur, the C-terminal L117 of GABARAP must first be cleaved off by the ATG4 family of cysteine proteases to yield the 116-residue form GABARAP-I, which can subsequently be conjugated to yield the lipidated form GABARAP-II. Lipidation can also be reversed by the ATG4 family of proteases. In the last decade, a plethora of interaction partners have been identified that reveal the essential role of GABARAP especially in vesicle transport and fusion events in autophagy and apoptosis [10, 21, 22]. In order to accomplish this multi-functionality as a well-ordered protein with well-defined secondary and tertiary structure elements, it obviously needs to possess high inherent flexibility and cannot be described by a static tertiary structure. In fact, conformational heterogeneity appears to be a hallmark of the GABARAP/MAP1LC3/Atg8 family and is conserved from yeast to mammals [10, 23, 24]. In this work, we have characterized the picoto nanosecond dynamics of soluble GABARAP and GABARAP-I by an integrated approach using NMR ^{15}N spin relaxation and fluorescent techniques based on time-resolved anisotropy and fluorescence correlation spectroscopy (FCS) in concert with MD simulations.

2. Theory

2.1. Time-resolved fluorescence anisotropy

It has long been recognized that the physical mechanisms underlying fluorescence depolarization on the one hand and NMR relaxation on the other hand are closely related to each other [25]. An understanding of how these different experimental methods complement each other is obtained by revisiting the relevant theory with a particular emphasis on the conceptual similarities. In a nutshell, the anisotropy $r(t+t_c)$ of the fluorescence intensity emitted at time $t+t_c$ after excitation at time t is shown [26] to depend on the orientation of the (electric) transition dipole moment for emission, indicated by the unit vector $\vec{\mu}_e$, according to

$$r(t+t_c) = \frac{F_p(t+t_c) - F_s(t+t_c)}{F_p(t+t_c) + 2F_s(t+t_c)} = \left\langle \left(3\left(\vec{e}_p \bullet \vec{\mu}_e(t+t_c)\right)^2 - 1\right)/2 \right\rangle = \left\langle P_2\left(\vec{e}_p \bullet \vec{\mu}_e(t+t_c)\right)\right\rangle, \qquad (1)$$

where F_p and F_s are the emission intensities with polarization parallel (*p*) and perpendicular (*s*), respectively, to the polarization of the excitation beam indicated by the unit vector $\vec{e}_p$, $P_2(x)=(3x^2-1)/2$ is the second Legendre polynomial, and the angular brackets denote ensemble averaging. Taking into account that the excitation of each molecule depends on the orientation of the (electric) transition dipole moment for absorption, indicated by the unit vector $\vec{\mu}_a$, at time t with respect to the polarization of the incident beam, and making use of the addition theorem for second-order spherical harmonics, eq 1 can be recast [26] into

$$r(t+t_c) = \frac{2}{5}\left\langle P_2\left(\vec{\mu}_a(t) \bullet \vec{\mu}_e(t+t_c)\right)\right\rangle, \qquad (2)$$

which now depends only on the relative orientation of the transition dipole moments for absorption and emission at two different time points. Again making use of the addition theorem for second-order spherical harmonics, this relative orientation can be further separated into the relative orientation of the transition dipole moments of the chromophore itself and the reorientation of the chromophore between absorption and emission,

$$r(t+t_c) = \frac{2}{5}P_2(\cos\delta)\left\langle P_2\left(\vec{\mu}_e(t) \bullet \vec{\mu}_e(t+t_c)\right)\right\rangle = r_0\left\langle P_2\left(\vec{\mu}_e(t) \bullet \vec{\mu}_e(t+t_c)\right)\right\rangle = r_0 C(t_c), \qquad (3)$$

where δ denotes the angle between the transition dipole moments for absorption and emission of the chromophore [25, 26]. In this expression, $r(t_c=0) = r_0 = 0.4\ P_2(\cos\delta)$ is a constant for a given chromophore and the decay of the fluorescence anisotropy is therefore determined by the loss of autocorrelation of the second Legendre polynomial of the orientation of the chromophore,

$$C(t_c) = \left\langle P_2(\vec{\mu}_e(t) \bullet \vec{\mu}_e(t+t_c))\right\rangle = \left\langle (3\cos^2\theta(t+t_c) - 1)/2 \right\rangle, \tag{4}$$

where θ is the angle of the chromophore at time $t+t_c$ with respect to its orientation at time t.

2.2. NMR spin relaxation

NMR relaxation of a particular spin $\vec{I}$ of interest is usually dominated by the (magnetic) dipole-dipole coupling to a spin that is close in space, e. g. to the directly bonded amide ^{1}H in the case of backbone amide ^{15}N spin relaxation. The (magnetic) dipole moment associated with the neighboring spin $\vec{S}$ generates a magnetic dipole field with a secular (i. e., first-order perturbation theory) component of

$$B_{DD}(t+t_c) = \frac{\mu_0}{4\pi}\frac{\gamma_S}{r_{IS}^3(t+t_c)} 2P_2(\vec{e}_z \bullet \vec{\mu}_{IS}(t+t_c))S_z = \frac{\mu_0}{4\pi}\frac{\gamma_S}{r_{IS}^3(t+t_c)}(3\cos^2\beta(t+t_c)-1)S_z, \tag{5}$$

where μ_0 is the permeability of the vacuum, γ_S is the gyromagnetic ratio of spin S, r_{IS} is the distance and $\vec{\mu}_{IS}$ the direction of the internuclear vector connecting I and S, $\vec{e}_z$ is the direction of the static external magnetic field $\vec{B}_0 = B_0\vec{e}_z$, β is the angle of the internuclear vector $\vec{\mu}_{IS}$ with respect to $\vec{B}_0$, and S_z is the component of $\vec{S}$ in the direction of $\vec{B}_0$ [27]. Although the distance r_{IS} (e. g. the ^{1}H-^{15}N bond length) usually varies little, any reorientation of the spin pair gives rise to a fluctuating magnetic field, whose dependence on the orientation relative to the laboratory frame follows a similar mathematical form as the fluorescence anisotropy (1), as expected from the close analogy between (electric) transition dipole moments and (magnetic) nuclear dipole moments. Accordingly, it can be shown [25, 28] that the fluctuation of the magnetic dipole-dipole interaction (or other axially symmetric second-rank tensorial interaction including chemical shift anisotropy (CSA)) is described by the same autocorrelation function as above:

$$C(t_c) = \left\langle P_2(\vec{\mu}_{IS}(t) \bullet \vec{\mu}_{IS}(t+t_c))\right\rangle = \left\langle (3\cos^2\theta(t+t_c) - 1)/2 \right\rangle, \tag{6}$$

where θ is the angle of $\vec{\mu}_{IS}$ at time $t+t_c$ with respect to its orientation at time t.

2.3 Autocorrelation function for rotational diffusion

This orientation θ is modulated by both overall rotational diffusion as well as internal motions, which are assumed to be independent of each other so that the autocorrelation function can be separated into the autocorrelation function for the overall rotational diffusion, C_{global}, and the autocorrelation function for the internal motions, C_{int}:

$$C(t_c) = C_{\text{int}}(t_c)\, C_{global}(t_c)\,. \tag{7}$$

Thus, in order to evaluate the ensemble average in autocorrelation functions (4) and (6) we have to solve the rotational diffusion equation first. Autocorrelation functions of diffusive (more generally, Markovian stochastic) motions can be expressed as a sum of exponentials. In the case of isotropic rotational diffusion, the conditional probability density $p(\vec{\mu}_0(t)\,|\,\vec{\mu}(t+t_c))$ of finding the molecule in orientation $\vec{\mu}(t+t_c)$ at time $t+t_c$ after finding it in orientation $\vec{\mu}_0(t)$ at time t follows the diffusion equation

$$\frac{\partial}{\partial t_c} p(\vec{\mu}_0(t)\,|\,\vec{\mu}(t+t_c)) = D_{rot} L^2 p(\vec{\mu}_0(t)\,|\,\vec{\mu}(t+t_c))\ , \tag{8}$$

where D_{rot} is the rotational diffusion coefficient and $\vec{L}$ the angular momentum operator divided by the Planck constant ħ=h/(2π) [29-31]. Because the spherical harmonics $Y_{lm}(\vec{\mu})$ are a complete set of orthonormal eigenfunctions of the angular momentum operator with

$$L^2 Y_{lm}(\vec{\mu}) = l(l+1) Y_{lm}(\vec{\mu})\,, \tag{9}$$

where *l*=0,1,2,… is the angular momentum quantum number and m=–*l*,–*l*+1,…,0,…,*l*–1,*l* the magnetic quantum number [32], the solution of (8) can be expressed [29-31] as

$$p(\vec{\mu}_0(t)\,|\,\vec{\mu}(t+t_c)) = \sum_{l=0}^{\infty}\sum_{m=-l}^{l} g_{lm} Y_{lm}(\vec{\mu}(t+t_c))\, e^{-l(l+1)D_{rot}t_c}\,. \tag{10}$$

The coefficients g_{lm} have to satisfy the initial condition that $\vec{\mu}_0(t)$ is the only orientation with non-zero probability density in the limit t_c=0:

$$p(\vec{\mu}_0(t)\,|\,\vec{\mu}(t)) = \sum_{l=0}^{\infty}\sum_{m=-l}^{l} g_{lm} Y_{lm}(\vec{\mu}(t)) = \delta(\vec{\mu}(t) - \vec{\mu}_0(t)) \tag{11}$$

Recalling the closure relation of the spherical harmonics [32]

$$\sum_{l=0}^{\infty}\sum_{m=-l}^{l} Y^*_{lm}(\vec{\mu}_0)\, Y_{lm}(\vec{\mu}) = \delta(\vec{\mu} - \vec{\mu}_0) \tag{12}$$

it is straightforward to see that the relevant solution of (8) is:

$$p(\vec{\mu}_0(t) \mid \vec{\mu}(t+t_c)) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} Y_{lm}^{*}(\vec{\mu}_0(t))\, Y_{lm}(\vec{\mu}(t+t_c))\, e^{-l(l+1)D_{rot}t_c} . \tag{13}$$

Note that the Legendre polynomials $P_l(\vec{\mu} \bullet \vec{e}_z) = \sqrt{4\pi/(2l+1)}\, Y_{l0}(\vec{\mu})$ are essentially the subset of the spherical harmonics with m=0. Making use of the addition theorem for spherical harmonics [32]

$$P_l(\vec{\mu}_e(t) \bullet \vec{\mu}_e(t+t_c)) = \frac{4\pi}{2l+1} \sum_{m=-l}^{l} Y_{lm}(\vec{\mu}_e(t))\, Y_{lm}^{*}(\vec{\mu}_e(t+t_c)) \tag{14}$$

the ensemble average in the autocorrelation function of the second Legendre polynomial (4) is now readily evaluated by integrating over all orientations (solid angles):

$$\begin{aligned} C_{global}(t_c) &= \langle P_2(\vec{\mu}_e(t) \bullet \vec{\mu}_e(t+t_c)) \rangle \\ &= \iint P_2(\vec{\mu}_0(t) \bullet \vec{\mu}(t+t_c))\, p(\vec{\mu}_0(t))\; p(\vec{\mu}_0(t) \mid \vec{\mu}(t+t_c))\; d^2\vec{\mu}_0\, d^2\vec{\mu} \end{aligned} \tag{15}$$

In an isotropic sample the probability density of finding the molecule in orientation $\vec{\mu}_0(t)$ at time t is $p(\vec{\mu}_0(t)) = 1/(4\pi)$ for all orientations on the unit sphere. It is obvious from the addition theorem (14) that rotations preserve the order l of the spherical harmonics and it is therefore clear from the orthonormality of the spherical harmonics that only the term with l=2 in the sum of exponentials in (13) contributes to the autocorrelation function of the second (l=2) Legendre polynomial. Thus, we finally obtain the mono-exponential autocorrelation function

$$\begin{aligned} C_{global}(t_c) &= \iint \frac{4\pi}{5} \sum_{m=-2}^{2} Y_{2m}(\vec{\mu}_0)\, Y_{2m}^{*}(\vec{\mu}) \times \frac{1}{4\pi} \sum_{l=0}^{\infty} \sum_{m'=-l}^{l} Y_{lm'}^{*}(\vec{\mu}_0)\, Y_{lm'}(\vec{\mu})\, e^{-l(l+1)D_{rot}t_c}\, d^2\vec{\mu}_0\, d^2\vec{\mu} \\ &= \frac{1}{5} \sum_{m=-2}^{2} \sum_{l=0}^{\infty} \sum_{m'=-l}^{l} \delta_{2l}\, \delta_{mm'}\, \delta_{2l}\, \delta_{mm'}\, e^{-l(l+1)D_{rot}t_c} = \frac{1}{5} \sum_{m=-2}^{2} e^{-6D_{rot}t_c} = e^{-6D_{rot}t_c} = e^{-t_c/\rho_{global}} \end{aligned} \tag{16}$$

with the global rotational autocorrelation time $\rho_{global}=1/(6D_{rot})$.

If there are internal motions on the same time-scale as the global rotational diffusion ρ_{global} or faster, the second Legendre polynomial autocorrelation function will decay more rapidly than in rigid areas of the molecule, which can be described by multi-exponential autocorrelation functions such as

$$\begin{aligned} C(t_c) &= C_{\text{int}}(t_c)\, C_{global}(t_c) = C_{\text{int}}(t_c)\, e^{-t_c/\rho_{\text{global}}} \\ &= \left[(S_{init}^2 - S_{fast}^2)e^{-t_c/\rho_{\text{fast}}} + (S_{fast}^2 - S^2)e^{-t_c/\rho_{slow}} + S^2\right]e^{-t_c/\rho_{\text{global}}} \end{aligned} , \tag{17}$$

where the autocorrelation times ρ_{fast}, ρ_{slow} are the time-scales of two internal motions of a given chromophore or spin pair orientation vector and the order parameters S_{fast}, S describe the motional restriction of these two internal motions on a scale of 0 to 1 [28, 33]. The initial order parameter S_{init} can be used to account for any loss in autocorrelation by motions that are too fast to be resolved experimentally or captured in the MD trajectories, otherwise S_{init}=1. If the overall rotational diffusion is not sufficiently isotropic the rotational diffusion coefficient D_{rot} is described by a symmetric tensor and in this case the global autocorrelation function $C_{global}(t_c)$ must itself be described by a weighted sum of 3 (axially symmetric tensor) or 5 (fully asymmetric tensor) different exponentials with weighting factors that depend on the orientation of the chromophore or spin pair relative to the rotational diffusion tensor [30, 34], which can be determined if the high-resolution structure of the protein is available.

2.4. NMR observables probing pico- to nanosecond dynamics

As a result, internal motions as well as overall rotational diffusion are directly reflected in a multi-exponential decay of the fluorescence anisotropy according to eqs 3 and 17. In NMR spectroscopy, they cause stochastic fluctuations of local electromagnetic fields. This does not affect the positions (frequencies) of the resonances in the NMR spectrum because the secular (first-order perturbation theory) component of the dipole-dipole interaction in eq 5 and of the CSA averages to zero in isotropic solution [27]. However, in second order the fluctuating magnetic fields stimulate transitions between the spin states, which in turn leads to a redistribution of longitudinal magnetization (longitudinal relaxation), loss of phase coherence (transverse relaxation), and longitudinal cross-relaxation effects such as the Nuclear Overhauser Effect (NOE), which are all reflected in the intensities and linewidths of the NMR resonances. The spectral density of the stochastic fluctuations is obtained by Fourier analysis of the autocorrelation function [25, 28]:

$$J(\omega) = \frac{2}{5}\int_0^\infty C(t_c)\cos(\omega t_c)dt_c\ , \tag{18}$$

which is a sum of Lorentzians in the case of the multi-exponential extended Lipari-Szabo-type autocorrelation function $C(t_c)$ given by eq 17:

$$J(\omega) = \frac{2}{5}\sum_{i=-k}^{k} c_i \times \rho_i \left(\frac{S^2}{1+(\omega\rho_i)^2} + \frac{(1-S_{fast}^2)(\rho_{fast}+\rho_i)\rho_{fast}}{(\rho_{fast}+\rho_i)^2 + (\omega\rho_{fast}\rho_i)^2} + \frac{(S_{fast}^2 - S^2)(\rho_{slow}+\rho_i)\rho_{slow}}{(\rho_{slow}+\rho_i)^2 + (\omega\rho_{slow}\rho_i)^2} \right) \tag{19}$$

where ρ_i are the five autocorrelation times that describe the rotational diffusion process in the fully asymmetric case [33, 35]. In second order, it is clear that only the spectral density components J(ω) at the frequencies ω=0 (secular dephasing), $\omega=\omega_I$ (I spin flip), $\omega=\omega_S$ (S spin flip), $\omega=\omega_I-\omega_S$ (flip-flop transition), and $\omega=\omega_I+\omega_S$ (flip-flip transition) fulfill the necessary resonance conditions to stimulate transitions between different spin states, where ω_I and ω_S are the Larmor frequencies of spin I and spin S, respectively. Detailed calculation [36] reveals that the longitudinal relaxation rate, R_1, transverse relaxation rate, R_2, and heteronuclear NOE of a protein backbone amide ^{15}N nucleus are given by the following linear combinations of these spectral density components:

$$R_1 = \frac{1}{T_1} = d\left[J(\omega_H - \omega_N) + 3J(\omega_N) + 6J(\omega_H + \omega_N)\right] + cJ(\omega_N) \tag{20}$$

$$\begin{aligned} R_2 = \frac{1}{T_2} = \frac{d}{2}\left[4J(0) + J(\omega_H - \omega_N) + 3J(\omega_N) + 6J(\omega_H + \omega_N) + 6J(\omega_H)\right] \\ + \frac{c}{6}\left[4J(0) + 3J(\omega_N)\right] + R_{ex} \end{aligned} \tag{21}$$

$$NOE = 1 + \left[\frac{\gamma_H}{\gamma_N} d \frac{6J(\omega_H + \omega_N) - J(\omega_H - \omega_N)}{R_1}\right] \tag{22}$$

Here, the dipolar constant is defined as $d = \frac{1}{4}\left(\frac{\mu_0}{4\pi}\right)^2 \frac{(\gamma_H \gamma_N \hbar)^2}{\left\langle r_{NH}{}^6 \right\rangle}$, where μ_0 is the permeability of the vacuum, γ_H and γ_N are the gyromagnetic ratios of the ^{1}H and ^{15}N spins, ħ is Planck´s constant divided by 2π and r_{NH} is the bond length. The constant $c = (\omega_N \Delta\sigma_N)^2 / 3$ depends on the chemical shift anisotropy $\Delta\sigma_N$ measured in ppm. R_{ex} is the contribution to transverse relation from chemical exchange on the micro- to millisecond time-scale.

Thus, the motional parameters such as autocorrelation times and generalized order parameters are indirectly encoded in the spin relaxation rates via the spectral density at the distinct frequencies 0, ω_H, ω_N, ω_H+ω_N, and ω_H−ω_N (Figure S1). The motional parameters can be extracted as adjustable parameters in an iterative optimization procedure by comparing the spin relaxation rates back-calculated from the so-called “model-free” spectral density (eq 19) inserted into eqs 20 to 22 with the experimentally determined spin relaxation rates. This is usually accomplished using software packages such as relax [35, 37-40].

2.5. Fluorescence correlation spectroscopy

The decay of the anisotropy is not the only fluorescence polarization dependent technique that is sensitive to the global rotational diffusion of the molecule. Polarization-resolved fluorescence correlation spectroscopy (pFCS), in which fluorescence intensity fluctuations $\delta F_{p/s}(t_c) = F_{p/s}(t_c) - \left\langle F_{p/s}(t_c)\right\rangle$ under constant excitation are measured, is also able to resolve molecular rotational motion [29, 31, 41]). It is obvious that two consecutive fluorescence events (absorption followed by emission) of the same chromophore are not independent from each other but correlated. Most importantly, the second absorption process cannot occur immediately after the first absorption process before the molecule has returned to the ground state again, so there is an initial lag phase with an exponential build-up of the photon correlation determined by the excitation rate and the fluorescence lifetime, τ_e, and this build-up is referred to as photon anti-bunching. Accordingly, FCS is primarily sensitive to dynamic processes on the time-scale slower than τ_e, whereas the decay of the fluorescence intensity on the time-scale of τ_e limits the sensitivity of anisotropy measurements to processes on this time-scale or faster. After the initial build-up, photon correlation is slowly lost to a variety of physical processes as the chromophore begins to populate long-lived non-fluorescent states such as triplet states, reorients due to rotational diffusion, and diffuses translationally out of the active volume of the excitation beam. Under certain conditions [29, 31] the normalized photon correlation function can be factorized into:

$$G(t_c) = 1 + \frac{1}{N} G_a(t_c) \times G_b(t_c) \times G_{rot}(t_c) \times G_{diff}(t_c) \tag{23}$$

where $G_a(t_c)$ describes the photon anti-bunching, $G_b(t_c)$ the population of non-fluorescent states, $G_{rot}(t_c)$ the effect of rotational diffusion and $G_{diff}(t_c)$ of translational diffusion, and $1/N$ is a scaling factor to account for the effect of the total number N of fluorophore emitters independently diffusing in the confocal detection volume.

As noted above, the probability density of absorption depends on the orientation of $\vec{\mu}_a$ relative to the polarization of the incident beam, $\vec{e}_p$, and is proportional [29, 31] to $3\left(\vec{e}_p \bullet \vec{\mu}_a\right)^2 = 2P_2\left(\vec{e}_p \bullet \vec{\mu}_a\right) + P_0\left(\vec{e}_p \bullet \vec{\mu}_a\right)$, where P_0=1 and P_2(x)=(3x^2−1)/2 are the Legendre polynomials of order l=0 and l=2, respectively. Similarly, the probability of emitting a photon with polarization $\vec{e}_e$ is proportional to $3(\vec{e}_e \bullet \vec{\mu}_e)^2 = 2P_2(\vec{e}_e \bullet \vec{\mu}_e) + P_0(\vec{e}_e \bullet \vec{\mu}_e)$. The orientation-dependent rotational factor of the correlation function of two photons with polarizations $\vec{e}_1$ and $\vec{e}_2$ emitted at times t and t+t_c, respectively, is hence given by

$$C_{rot}(t_c) = \int_{-\infty}^{t_c} dt'' \int_{-\infty}^{0} dt' \left\langle 3\left(\vec{e}_p \bullet \vec{\mu}_a(t+t')\right)^2 \times 3\left(\vec{e}_1 \bullet \vec{\mu}_e(t)\right)^2 \times 3\left(\vec{e}_p \bullet \vec{\mu}_a(t+t'')\right)^2 \times 3\left(\vec{e}_2 \bullet \vec{\mu}_e(t+t_c)\right)^2 \right\rangle \tag{24}$$

where the integration over the time points of the first (t+t’) and second (t+t”) absorption process accounts for the constant excitation by the incident beam in a typical FCS experiment [29, 31]. Although conceptually similar, this expression is clearly more complex than the second Legendre polynomial autocorrelation function (15) above because it depends on four chromophore orientations at four different time points (absorption, emission, absorption, emission) rather than two (absorption, emission). To make the problem more tractable, it is usually assumed that τ_e << ρ_{global} and the rotational diffusion between absorption and emission is thereby neglected; note that ρ_{global} is typically of the order of several nanoseconds for a small monomeric protein up to hundreds of nanoseconds for a large protein complex, so this assumption is often an approximation for commonly used chromophores with lifetimes τ_e of the order of a few nanoseconds. In this limit (24) simplifies to

$$C_{rot}(t_c) = \left\langle 3\left(\vec{e}_p \bullet \vec{\mu}_e(t)\right)^2 \times 3\left(\vec{e}_1 \bullet \vec{\mu}_e(t)\right)^2 \times 3\left(\vec{e}_p \bullet \vec{\mu}_e(t+t_c)\right)^2 \times 3\left(\vec{e}_2 \bullet \vec{\mu}_e(t+t_c)\right)^2 \right\rangle, \qquad (25)$$

where the additional assumption has been made that the transition dipole moments for absorption, $\vec{\mu}_a$, and emission, $\vec{\mu}_e$, are approximately parallel (δ≈0).

Using the conditional probability for diffusional reorientation (13) this expression can now be calculated for any particular combination of polarizations $\vec{e}_p$, $\vec{e}_1$ and $\vec{e}_2$ in much the same way as (16) above. As we have seen above, the orthonormality of the spherical harmonics guarantees that only the exponential with *l*=2 contributes to the loss in correlation of two second-order spherical harmonics at times t and t+t_c as a consequence of rotational diffusion. By contrast, the rotational factor (25) describes a correlation between a product of two mixed zero-/second-order spherical harmonics of the form $\left(2P_2\left(\vec{e}_p \bullet \vec{\mu}_e\right) + P_0\left(\vec{e}_p \bullet \vec{\mu}_e\right)\right) \times \left(2P_2\left(\vec{e}_1 \bullet \vec{\mu}_e\right) + P_0\left(\vec{e}_1 \bullet \vec{\mu}_e\right)\right)$ at time *t* with another such product at time t+t_c. In the language of quantum mechanics, we are therefore no longer dealing with a single angular momentum wavefunction of *l*=2 at each time point but with products of two angular momentum wave functions, and products of two angular momentum wave functions describe the sum of two angular momenta [32]. In the present case these two angular momenta can add up to a total of *l*=0, *l*=2, or *l*=4, and the rotational factor is therefore bi-exponential:

$$\begin{aligned} C_{rot}(t_c) &= \sum_{l=0,2,4} B_l\left(\vec{e}_p, \vec{e}_1, \vec{e}_2\right) e^{-l(l+1)D_{rot}t_c} \\ &= B_0\left(\vec{e}_p, \vec{e}_1, \vec{e}_2\right) + B_2\left(\vec{e}_p, \vec{e}_1, \vec{e}_2\right) e^{-6D_{rot}t_c} + B_4\left(\vec{e}_p, \vec{e}_1, \vec{e}_2\right) e^{-20D_{rot}t_c} \\ &= B_0\left(\vec{e}_p, \vec{e}_1, \vec{e}_2\right) + B_2\left(\vec{e}_p, \vec{e}_1, \vec{e}_2\right) e^{-t_c/\rho_{global}} + B_4\left(\vec{e}_p, \vec{e}_1, \vec{e}_2\right) e^{-t_c/\left(0.3\rho_{global}\right)} \end{aligned} \qquad (26)$$

with coefficients $B_l\left(\vec{e}_p, \vec{e}_1, \vec{e}_2\right)$ that are related to the Clebsch-Gordan coefficients and have been calculated and tabulated for the most common experimental geometries [29, 31], or, after normalization to the equilibrium value $C_{rot}(\infty) = B_0\left(\vec{e}_p, \vec{e}_1, \vec{e}_2\right)$,

$$\begin{aligned} G_{rot}(t_c) &= \frac{C_{rot}(t_c)}{B_0(\vec{e}_p,\vec{e}_1,\vec{e}_2)} = 1 + \frac{B_2(\vec{e}_p,\vec{e}_1,\vec{e}_2)}{B_0(\vec{e}_p,\vec{e}_1,\vec{e}_2)} e^{-t_c/\rho_{global}} + \frac{B_4(\vec{e}_p,\vec{e}_1,\vec{e}_2)}{B_0(\vec{e}_p,\vec{e}_1,\vec{e}_2)} e^{-t_c/(0.3\rho_{global})} \\ &= 1 + b_{rot}\left(\frac{1}{1+C} e^{-t_c/\rho_{global}} + \frac{C}{1+C} e^{-t_c/(0.3\rho_{global})}\right) \end{aligned} , \qquad (27)$$

where $b_{rot} = \left(B_2(\vec{e}_p,\vec{e}_1,\vec{e}_2) + B_4(\vec{e}_p,\vec{e}_1,\vec{e}_2)\right)/B_0(\vec{e}_p,\vec{e}_1,\vec{e}_2)$ is the amplitude of the correlation due to rotational diffusion and $C = B_4(\vec{e}_p,\vec{e}_1,\vec{e}_2)/B_2(\vec{e}_p,\vec{e}_1,\vec{e}_2)$. In the case of excitation with a laser beam along $\vec{e}_z$ with polarization $\vec{e}_p = \vec{e}_x$ and cross-correlation of two emitted photons with polarization $\vec{e}_1 = \vec{e}_x = \vec{e}_p$ (parallel) and $\vec{e}_2 = \vec{e}_y = \vec{e}_s$ (perpendicular), e. g., which is the experimental geometry used in the present work, the relevant coefficients are $B_0(\vec{e}_x,\vec{e}_x,\vec{e}_y) = 1323$, $B_2(\vec{e}_x,\vec{e}_x,\vec{e}_y) = 540$, $B_4(\vec{e}_x,\vec{e}_x,\vec{e}_y) = -280$ [31], and hence b_{rot}=(540−280)/1323=0.197 and C=−280/540=−0.519. In summary, auto- and cross-correlation functions from pFCS measurements are also sensitive to rotational diffusion, albeit in a somewhat more complex mathematical form than the autocorrelation functions probed by fluorescence anisotropy and NMR relaxation spectroscopy, respectively.

2.6 Hydrodynamic radius

If the molecule is approximately spherical in shape the single rotational diffusion constant D_{rot} and, hence, the global rotational autocorrelation time ρ_{global} are determined by the hydrodynamic radius of the molecule, $R_{h,rot}$, according to the Stokes-Einstein relation for rotational diffusion:

$$D_{rot} = \frac{k_B T}{8\pi\eta R_{h,rot}^3} = \frac{1}{6\rho_{global}}, \qquad (28)$$

Comparison with the Stokes-Einstein relation for translational diffusion (29) reveals that rotational diffusion is obviously an even more sensitive probe of molecular size than translational diffusion.

$$D_{trans} = \frac{k_B T}{6\pi\eta R_{h,trans}} \qquad (29)$$

For a spherical molecule $R_{h,rot}$ and $R_{h,trans}$ are identical. If the particle deviates from spherical geometry the hydrodynamic friction increases compared to the equivalent sphere of equal volume, R_{eq}, by factors of $R_{h,rot}^3 = F_{rot} \times R_{eq}^3$ and $R_{h,trans} = F_{trans} \times R_{eq}$, where F_{rot} and F_{trans} are the Perrin shape factors for rotational and translational diffusion, respectively. After addition of hydrogen atoms, the crystal structure of GABARAP (PDB 1GNU) has approximately the same tensor of inertia as a prolate ellipsoid of revolution with semi-axes of 23.0 Å and 15.2 Å and an equivalent sphere of $R_{eq} = \sqrt[3]{(15.2\ \text{Å})^2 \times 23.0\ \text{Å}} = 17.5\ \text{Å}$. Assuming a hydration layer of 2.8 Å, we obtain an estimate for the equivalent sphere of hydrated GABARAP of $R_{eq} = \sqrt[3]{(18.0\ \text{Å})^2 \times 25.8\ \text{Å}} = 20.3\ \text{Å}$ and for the axial ratio of P = 25.8 Å/18.0 Å = 1.433. Although this deviation from spherical geometry appears significant, the effect of the corresponding shape factors for such an ellipsoid of revolution of $F_{rot} = 1.05$ and $F_{trans} = 1.01$ on D_{rot} and D_{trans}, respectively, is actually smaller than typical experimental uncertainties, and $R_{h,rot}$ is less than 1% larger than $R_{h,trans}$. In other words, the spherical approximation for D_{rot} and D_{trans} is generally justified for proteins with a moderate deviation from spherical geometry such as GABARAP.

2.7 Practical considerations

Although the theories describing overall and internal dynamics in NMR and fluorescence spectroscopy are closely related to each other, the sample conditions and the procedures to extract quantitative motional parameters differ significantly. NMR spectroscopy provides high-resolution three-dimensional structures and insight into global macromolecular diffusion and local intramolecular motions from spin probes abundantly distributed over the entire protein (typically, on a per-residue basis), but requires highly pure and sufficiently stable samples with sample concentrations close to the millimolar range. In contrast, fluorescence spectroscopy is sensitive enough to be performed at low sample concentrations in the pico- to micromolar range but necessitates the protein to be modified by attachment of individual fluorescent probes. In practice, attachment of a small dye with a short linker and no charge such as BodipyFL to the side chain of a cysteine occurring naturally or introduced via site-directed mutagenesis allows quantification of overall and internal dynamics by analyzing the fluorescence anisotropy decay from time-correlated single photon counting (TCSPC) experiments or the correlation curves from pFCS. As explained above, TCSPC is limited by the lifetime of the dye and hence particularly useful for the investigation of fast dynamics ranging from sub-ns to several nanoseconds, whereas pFCS is sensitive to slower motions beyond the upper limit of time-resolved anisotropy [42, 43]. Since the fluorescent dye is attached to the cysteine side-chain, it reports on side chain flexibility instead of the backbone amide bond vector reorientation most commonly probed by NMR relaxation experiments. Backbone and side chain dynamics can differ significantly [44] and are thus complementary to each other. An atomistic visualization and interpretation of the motional parameters measured by NMR and fluorescence spectroscopy can be obtained from MD simulations, thereby revealing the nature and functional relevance of the experimentally detected protein motions.

3. Methods

3.1 Mutagenesis of the plasmid encoding GABARAP

Cysteine variants of GABARAP were obtained using the QuikChange site directed mutagenesis kit according to the protocol provided by Agilent. Primers for the mutagenesis were designed with PrimerX (www.bioinformatics.org/primerx) in order to obtain the following mutations: GABARAP V4C, GABARAP E7C, GABARAP K13C, GABARAP I41C, GABARAP F62C, GABARAP-I G116C.

3.2 Expression, purification and labeling

GABARAP wild type and cysteine variants were expressed from a pET11a vector in *E. coli* BL21(DE3)-T1^{R}. Expression and purification was performed according to the protocol reported by Coyle et al. [45] with minor modifications. Briefly, cells were grown at 37°C in LB (Lysogeny-Broth) medium, or in M9 minimal medium containing $^{15}NH_4Cl$ as the sole nitrogen source for the production of uniformly ^{15}N-enriched ([U-^{15}N]) samples for NMR spectroscopy. Protein expression was induced by 1 mM IPTG (isopropyl β-D-thiogalactopyranoside) at OD_{600} = 0.8 - 1.0. The expression was carried out at 20°C for 15 h. Cells were harvested and resuspended in lysis buffer containing 1 mM EDTA, 10 mg/ml DNAse, complete protease inhibitor cocktail (Roche), 50 mM KCl and 25 mM sodium phosphate at pH 6.5. The cells were sonicated for protein solubilization and insoluble cell debris was removed by centrifugation. For purification the lysate was loaded onto a 20 ml HiLoad 16/10 SP Sepharose column (GE Healthcare) and proteins were eluted with a linear gradient ranging from 0.05 M to 0.60 M KCl. Fractions containing GABARAP were pooled and further purified by gel filtration using a HiLoad 16/600 Superdex 75 pg column (GE Healthcare). Prior to fluorescent dye labeling, proteins were incubated in labeling buffer (25 mM Tris, 300 mM NaCl, 0.5 mM phenylmethane sulfonyl fluoride (PMSF), 0.5 mM ethylene glycol-bis(2-aminoethylether)-N,N,N’,N’-tetraacetate (EGTA), 0.5 mM ascorbic acid, pH 7.5) containing 10 mM dithiothreitol (DTT) at room temperature for 20 min. DTT was removed by desalting on Sephadex G-25 (NAP-5 prepacked columns, GE Healthcare). Proteins were labeled in labeling buffer with an excess of BODIPY FL (BFL) Iodoacetamide (ThermoFisher Scientific) for 2 hours at room temperature. Unreacted dye was removed by desalting on NAP-5 prepacked columns.

3.3 NMR data acquisition and analysis

NMR samples contained 0.5 – 1.0 mM [U-^{15}N] GABARAP or 0.7 mM [U-^{15}N] GABARAP-I G116C in the NMR buffer consisting of 25 mM sodium phosphate, 100 mM KCl, 100 mM NaCl, 0.1 mM EDTA, 0.02% NaN_3 in H_2O/D_2O (9:1). NMR experiments were performed on spectrometers equipped with cryogenically cooled triple or quadruple resonance probes with z axis pulsed field gradient capabilities operating at proton Larmor frequencies of 600 MHz and 900 MHz. The sample temperature was calibrated using methanol-d_4 (99.8%) [46]. At least 615 (96) complex data points were acquired with a spectral width of 16 ppm (29 ppm) in the ^{1}H (^{15}N) dimension. NMR data were processed using NMRPipe [47].

3.4 NMR relaxation experiments

^{15}N spin relaxation data at temperatures of 5°C, 15°C, 25°C, and 35°C were collected on a sample of 1.0 mM [U-^{15}N] GABARAP at 600 MHz and in the case of 25°C also at 900 MHz. In addition, we collected ^{15}N spin relaxation data on a sample of 0.7 mM [U-^{15}N] GABARAP-I G116C at 25°C and 600 MHz. Longitudinal relaxation rates R_1 were determined from ^{15}N inversion recovery experiments [48] with 10 different recovery delays (3 in duplicate for error estimation) between 10 ms and 1200 ms using recycle delays of 2.0 s to 2.5 s. Transverse relaxation rates R_2 were calculated from R_1 and rotating-frame relaxation rates $R_{1\rho}$ determined from ^{15}N spin-lock experiments [49] with 11 different spin-lock periods (3 in duplicate for error estimation) between 2 ms and 100 ms at a field strength of 2.0 kHz using a recycle delay of 3.0 s. All data sets were acquired in an interleaved manner to reduce the effects of any sample or instrument instabilities over the duration of the experiment. R_1 and $R_{1\rho}$ were determined by three-way composition of the pseudo three-dimensional NMR spectra using MUNIN [50, 51]. {^{1}H}^{15}N heteronuclear NOE values were calculated as the ratio of the peak intensities, extracted with NMRViewJ 8.0.3 [52], in two interleaved spectra recorded with and without proton saturation for the final 6 s of the recycle delay of 15 s [48], with uncertainties estimated from the background noise of the spectra. Residues with large internal motions on the sub-nanosecond time-scale as indicated by $\{^1H\}^{15}N$ values below 0.65 or involved in chemical exchange processes as indicated by R_2/R_1 ratios that deviate by more than 10% from the mean were considered to possess significantly increased internal mobility [53]. These mobile residues were excluded from the calculation of the rotational diffusion tensor based on the crystal structure of GABARAP ([54], PDB 1GNU; hydrogen atoms added with the NIH version 1.2.1 [55] of X-PLOR 3.851 [56]) using Tensor 2.0 [57] with the default parameters. Full "model-free" analysis of the ^{15}N relaxation data at 600 MHz and 900 MHz recorded at 25°C was performed using the protocol of d'Auvergne et al. as implemented in relax version 4.0.0 [35, 37-40]. Because of the large number of experimental data points the statistical uncertainties on the extracted model parameters such as ρ_{global} are minute; systematic uncertainties were estimated by repeating the fits with ^{15}N transverse relaxation rates R_2 that are systematically lower or higher than the experimentally determined values by 2%, a typical systematic error described for ^{15}N relaxation experiments in the literature [49, 58].

3.5 NMR translational diffusion experiments

The diffusion coefficient of 0.5 mM [U-^{15}N] GABARAP in the NMR buffer supplemented with 0.1% to 0.5% (v/v) dioxane as a reference was measured using 1D ^1H pulse gradient stimulated echo longitudinal encode-decode (PG-SLED) translational diffusion experiments [59] with individual rectangular-shaped or bipolar [60] sine-shaped encode/decode gradients and suppression of the H_2O resonance by WATERGATE [61] or weak presaturation at 600 MHz, 25°C. The methyl group region in the 1D ^1H spectra was integrated and the resulting intensities as a function of gradient strength fit by a Gaussian decay [62]. The decay constants from these fits were converted into diffusion coefficients [62] based on the absolute strength of the pulsed field gradients, which had been calibrated from a diffusion experiment on D_2O using the known diffusion coefficient of 1.9×10^{-5} cm^2/s at 25.0°C [59]. The resulting diffusion coefficients D_{trans} were in turn converted into hydrodynamic radii $R_{h,trans}$ based on the Stokes-Einstein equation for translational diffusion (eq 29) assuming a viscosity of η = 0.911 mPa s interpolated for 10% D_2O at 25.0°C [63]. Alternatively, the hydrodynamic radii were calculated relative to 0.1% or 0.5% (v/v) internal dioxane assuming a hydrodynamic radius of 2.12 Å [64]. The diffusion coefficient and hydrodynamic radii are reported as mean ± standard deviation over two independent measurements.

3.6 Samples for fluorescence spectroscopy

Fluorescence measurements were performed with approx. 0.5 µM GABARAP (total protein concentrations including both labeled and unlabeled molecules) and up to 1.3 mM GABARAP for crowding experiment, in 25 mM sodium phosphate, 100 mM KCl, 100 mM NaCl, pH 6.9. Due to the different sensitivity of the methods, the fluorescent portion of the molecules was adjusted to approximately 50 nM for the TCSPC measurements, whereas the confocal experiments (pFCS) were performed at concentrations that were at least 10 times lower.

3.7 TCSPC

Note that in fluorescence spectroscopy the time recorded in time correlated single photon counting is usually referred to as *t,* but as outlined in this work it is actually a correlation time t_c, so that we keep this nomenclature for consistency. As the fluorescence decay starts usually at time $t = 0$ (defined by the excitation pulse), this value is omitted for convenience and the time-resolved fluorescence decay is written as $F(t + t_c) = F(t_c)$.

Polarization-resolved ensemble fluorescence decays were recorded using a FluoTime300 fluorescence lifetime spectrometer (PicoQuant, Berlin, Germany) equipped with a pulsed super continuum laser SuperK Extreme (NKT Photonics, Denmark) as a light source running at 19.51 MHz and a wavelength of 485 nm in a temperature-stabilized cell at 20.0°C ± 0.1°C. Typically a total amount of 1.5×10^6 photons in 8 ps bins were collected per sample for both p- and s-polarization, $F_p(t_c)$ and $F_s(t_c)$, respectively, where t_c is the time that has elapsed since the excitation pulse at $t = 0$. The fluorescence and anisotropy decays were recovered by global fitting of the sum and difference curves according to:

$$F_{sum}(t_c) = F_{\mathrm{p}}(t_c) + 2GF_s(t_c) = F(t_c) \tag{30a}$$

$$F_{diff}(t_c) = F_{\mathrm{p}}(t_c) - GF_s(t_c) = F(t_c)r(t_c) \tag{30b}$$

where *G* is the detection efficiency ratio between the parallel and perpendicular channel. The procedure is described in [65]. Due to distinct local environments sensed by the flexibly coupled dye, the fluorescence decays $F(t_c)$ had to be fitted by a tri-exponential decay with species fractions x_i and fluorescence lifetimes τ_i (Tab. S2C):

$$F(t_c) = F_0\left(\sum_{i=1}^{3} x_i\, e^{-t_c/\tau_i}\right);\ \sum_{i=1}^{3} x_i = 1 \tag{31}$$

Fluorescence anisotropy decays $r(t_c)$ were fitted by a weighted sum of exponentials (see Tab. S2 for results) and the parameters obtained were converted into the product of triple-exponential decays with rotational correlation times ρ_{fast}, ρ_{slow} and ρ_{global}, and the related amplitudes r_{fast}, r_{slow} and r_∞ with $\Sigma r_i = r_0$:

$$r(t_c) = \sum_{i=1}^{3} r_i e^{-t_c/\rho_i} = r_0\left(\frac{r_{fast}}{r_0} e^{-t_c/\rho_{fast}} + \frac{r_{slow}}{r_0} e^{-t_c/\rho_{slow}} + \frac{r_\infty}{r_0}\right) e^{-t_c/\rho_{global}} \tag{32}$$

where $r_0 = 0.37$ is the fundamental anisotropy of BodipyFL (BFL) [66]. For reliable calculation of order parameters, ρ_{global} was fitted globally for all six GABARAP variants (Tab. S2B) and the fluorescence order parameter was calculated as $S^2 = r_\infty/r_0$. Using $C(t_c) = r(t_c)/r_0$ (see eq 3), eq 32 becomes equivalent to eq 17.

3.8 pFCS

Polarization-resolved full fluorescence correlation spectroscopy was performed for GABARAP F62C-BFL and I41C-BFL variants with a confocal laser scanning microscope (FV1000, Olympus, Germany) equipped with a single photon counting device with picosecond time-resolution (HydraHarp400, PicoQuant, Berlin, Germany) at 26°C ± 1°C. The sample was excited using the parked beam at 485 nm and the fluorescence F was collected in s- and p- polarized channels, $F_s(t)$ and $F_p(t)$, respectively. Full cross-correlation curves, $G_{s,p}(t_c)$ and $G_{p,s}(t_c)$, were obtained according to ref. [67]

$$G_{s,p}(t_c) = 1 + \frac{\langle \delta F_s(t)\cdot \delta F_p(t+t_c)\rangle}{\langle F_s(t)\rangle \cdot \langle F_p(t)\rangle} \text{ and } G_{p,s}(t_c) = 1 + \frac{\langle \delta F_p(t)\cdot \delta F_s(t+t_c)\rangle}{\langle F_s(t)\rangle \cdot \langle F_p(t)\rangle} \tag{33}$$

with the fluorescence fluctuations, $\delta F(t) = F(t) - \langle F(t)\rangle$. The registered photon events were analyzed employing a custom designed software package for multiparameter fluorescence spectroscopy, full correlation and multiparameter fluorescence imaging [67]. The applied factorized fitting function (23) models translational diffusion in a 3D-Gaussian volume element $G_{diff}(t_c)$, up to three temporary dark states $G_b(t_c)$, rotational diffusion of a spherical rotator $G_{rot}(t_c)$ and photon anti-bunching $G_a(t_c)$:

$$G(t_c) = 1 + \frac{1}{N} G_a(t_c) \times G_b(t_c) \times G_{rot}(t_c) \times G_{diff}(t_c) \text{ with}$$

$$G_{diff}(t_c) = \left(1 + \frac{t_c}{t_{diff}}\right)^{-1} \left(1 + \left(\frac{\omega_0}{z_0}\right)^2 \times \frac{t_c}{t_{diff}}\right)^{-\frac{1}{2}}$$

$$G_b(t_c) = \left(1 - b_1 + b_1 e^{-t_c/t_{b1}} - b_2 + b_2 e^{-t_c/t_{b2}} - b_3 + b_3 e^{-t_c/t_{b3}}\right)$$

$$G_{rot}(t_c) = \left(1 + b_{rot}\left(\frac{1}{1+C} e^{-t_c/\rho_{global}} + \frac{C}{1+C} e^{-t_c/(S\rho_{global})}\right)\right)$$

$$G_a(t_c) = \left(1 - a\ \mathrm{e}^{-t_c/t_a}\right) \tag{34}$$

Here, the observation volume is approximated by a 3D Gaussian volume with 1/e^2 radii in the lateral (ω_0) and axial direction (z_0), t_{diff} is the diffusion time, $b_{1,2,3}$ and $t_{b1,b2,b3}$ are amplitudes and times of the bunching terms, a and t_a are the amplitude and time of the anti-bunching term, S and C characterize the rotation model, b_{rot} and ρ_{global} are the amplitude and correlation time associated with rotational motion. As described above, factorization of the model function (23) is based on the assumption of well-separated time-scales for anti-bunching ($t_a \approx \tau_e$) and rotational correlation (ρ_{global}). With $t_a \approx 5$ ns and $\rho_{global} \approx 7$-10 ns this condition is not sufficiently met in our case. Thus, parameters for an ideal spherical rotator ($S = 0.3$, $C = -0.519$) [31] were not used. Instead, to compensate for distortions by coupling between anti-bunching and rotational terms due to overlapping time-scales, the two parameters $S = 0.65$, $C = -0.97$ were determined in a series of simulations. In addition to the shape of the correlation function its apparent relaxation time is also affected by this coupling. By comparing simulated rotational correlation times with results obtained by fitting the model function to the simulations we generated calibration data to derive correct rotational correlation times from the fit to the measured data (see Figure S2B).

3.9 Prediction of the global rotational diffusion from the molecular shape

Using the HYDROPRO software [68] the global rotational correlation time was calculated as $\rho_{\mathrm{global}} = 8.91$ ns ± 0.11 ns at 20.0°C (7.78 ns ± 0.10 ns at 25.0°C). To this end, several PDB structures were used (PDB IDs: 1GNU [54], 1KOT [69], 1KLV, 1KM7 [70] and 3D32 [71]), the temperature set to 20.0°C (25.0°C), the viscosity to $\eta = 1.002$ mPa s ($\eta = 0.890$ mPa s), and harmonic means of relaxation times were averaged (Table S4).

3.10 Molecular dynamics simulations

To gain a more detailed structural understanding of the nature of GABARAP dynamics we performed multiple MD simulations on the ns and sub-µs timescale using Gromacs version 4.6.5 [72]. The initial coordinates were taken from the PDB structure with ID 1GNU as this structure seems to have well-defined salt bridges and hydrogen bonds. However, the C-terminal region folds back towards the N-terminus in crystal structure 1GNU, thereby forming a conspicuous salt bridge between the terminal carboxyl group and the terminal amino group. A stable salt bridge between the terminal groups is at variance with the NMR structure (PDB ID 1KOT) and dynamics (see below) of GABARAP, suggesting that the salt bridge between the termini featured in the 1GNU crystal structure is not stable in solution. To alleviate any conformational sampling issues that might arise from starting from a potentially artifactual local energy minimum we decided to remove this salt bridge by simulating GABARAP-I instead, which lacks the C-terminal residue L117. The Amber99SB-ILDN [73] force field and the TIP3P [74] water model were used. The protein was centered in a cubic box with a minimum solute-to-wall distance of 1 nm. Water was added to the system as well as sodium and chloride ions to neutralize the system and to achieve a salt concentration of approximately 150 mM. The mass of all protons was increased to 4 u in order to remove the fastest degrees of freedom. After energy minimization and equilibration at a temperature of 27°C and pressure of 1 bar, multiple MD simulations were started in the NVT ensemble with the temperature kept at 27°C via a Nosé-Hoover thermostat [75]. Each simulation started from the same conformation with initial velocities randomly generated from a Maxwell distribution at 27°C. The equations of motions were integrated with a time step of 2 fs and snapshots saved every 5 ps. Electrostatic interactions were calculated with the particle mesh Ewald algorithm [76] using a Fourier grid spacing of 0.12 nm and a cut-off of 0.9 nm for the short-range interactions. The same cut-off was used for the calculation of the van-der-Waals interactions. All bond lengths were constrained with the LINCS algorithm [77]. In total, 40 short MD simulations of 75-79 ns in length and 15 long simulations of 565-580 ns in length were run, amounting to a total of 11.7 µs of collected fast GABARAP dynamics.

3.11 Calculation of S^2 order parameters from MD simulations

To compute S^2 order parameters, the MD trajectories were divided into 1100 subtrajectories of 10 ns length each. This is more than three times longer than the rotational correlation time that GABARAP experiences in the MD simulations. Nevertheless, we tested that the usage of longer subtrajectories of 20 ns length did not affect the results. N–H bond vector autocorrelation functions were computed according to eq 6, where $\vec{\mu}_{IS}(t) = \vec{\mu}_{NH}(t)$ is the normalized bond vector at time *t*. We calculated the S^2 order parameters using both total and internal correlation functions. Internal correlation functions $C_{int}(t_c)$, which only capture internal N–H bond vector motions, were computed from trajectories that had been superimposed on the starting structure of each sub-trajectory by minimizing the C$_\alpha$ root mean square deviation (RMSD), thereby removing overall rotational motion of the protein. S^2 order parameters were then computed for each sub-trajectory as [28, 78]

$$S^2 = \lim_{t_c \to \infty} C_{\text{int}}(t_c) \approx C_{\text{int}}(5\,ns) \tag{35}$$

and were subsequently averaged over all sub-trajectories. Total correlation functions $C(t_c)$ were evaluated from the raw trajectories including protein rotations (see eq 7). They were first averaged over all sub-trajectories and then fitted to multi-exponential decays as given in eq 17. However, $C(t_c)$ was not fitted for all bond vectors using three decays. Instead, the number of exponential decays, ranging from one to three, was chosen for each bond vector individually in order to avoid over-fitting of the data. The best-fit model was selected based on the Aikaike information criterion (AIC) [79] (see the Supporting Information (SI) for more information).

For the calculation of the order parameters for side chains from the MD trajectories, the correlation functions of the side-chain bond vectors between heavy atoms were fitted to multi-exponential decays following the same approach as for the N-H bond vectors above. Except for alanine, the C$_\alpha$-C$_\beta$ bond was omitted from the computation as it is directly correlated to the backbone dihedral angle rotations. Fast-rotating groups like the carboxyl groups of aspartate and glutamate were excluded. Moreover, bonds in aromatic rings, which can change their orientation as a result of overall ring rotation but are not representative of the overall side-chain conformation, were also not considered. The C$_\gamma$-C$_\delta$ bonds in phenylalanine and tyrosine are a good example for such bonds. The resulting S^2 values were averaged over all the bonds considered per side chain to obtain the corresponding side-chain order parameter. For the calculation of S^2 order parameters from the MD trajectories we developed the Python software MOP*S*^{2} (Molecular Order Parameters *S*^{2}), which is available free of charge under the following URL: https://github.com/schilli/MOPS. MOP*S*^{2} allows to calculate internal and global correlation functions and to determine the S^2 order parameters either from the internal correlation functions according to eq 35, by fitting the global correlation functions using the Lipari-Szabo model presented in eq 17 with the possibility to invoke this method together with the AIC, or by using the method presented in [16]. The latter method was not applied in the current work. More information on MOPS^2 is given in the SI (SI section 6).

4. Results

Using an integrated NMR, fluorescence and MD approach we obtained results reporting on the protein mechanics and hydrodynamics of GABARAP. The global diffusion of a protein is mainly described by the overall rotational correlation time ρ_{global} in the approximation of a spherical molecule, whereas the internal motion is described by the order parameter S^2 and the internal correlation times ρ_{slow} and ρ_{fast}. ^{15}N NMR spin relaxation, time-resolved fluorescence anisotropy or fluorescence correlation spectroscopy (FCS), and MD simulations all report on these fast-motional parameters. The order parameter S^2 accounts for the restriction of the motion of one N–H vector in the case of ^{15}N NMR spin relaxation data and of the dye attached to a side chain via a linker in case of the fluorescence-based methods, while from MD simulations S^2 values are obtained for both N–H bonds and side chains.

4.1 NMR Spectroscopy

^{15}N relaxation analysis. Backbone ^{15}N spin relaxation rates R_1, R_2 and the {^{1}H}-^{15}N heteronuclear NOE values of GABARAP at 600 MHz and 900 MHz measured at 25.0°C are shown in Figure 1. The ^{15}N spin relaxation rates are consistent at both field strengths. The overall average of the {^{1}H}-^{15}N heteronuclear NOE values of 0.80 ± 0.06 at 900 MHz and 0.70 ± 0.06 at 600 MHz reveals a stable tertiary fold. However, the C-terminal region from S113 onwards shows continuously decreasing values reaching 0.403 ± 0.003 at 900 MHz and 0.19 ± 0.01 at 600 MHz, suggesting that the backbone of the C-terminal region is largely disordered. Similarly low {^{1}H}-^{15}N NOE values are observed for only one additional residue, I41 in a loop region in spatial proximity to the C-terminal region (Figure 2B). {^{1}H}-^{15}N NOE values below 0.65 indicative of increased backbone mobility [53] are further observed for the N-terminal region and the loop between helix α_3 and strand β_3 (Figure 1). In addition to low {^{1}H}-^{15}N NOE values, the disordered C-terminal region also displays significantly elevated longitudinal ^{15}N relaxation rates, R_1 (Figure 1). The C-terminal residue, L117, and I41 in the loop nearby also show the slowest transverse ^{15}N relaxation rates, R_2 (Figure 1). Several residues in the N- and C-terminal regions as well as N82 in the loop between strand β_3 and helix α_4 exhibit conspicuously elevated ^{15}N R_2 rates at 25.0°C due to large contributions from chemical exchange on the millisecond time-scale to the ^{15}N line width. Unfortunately, the extensive chemical exchange line broadening has a negative influence on the signal to noise ratio of the affected NMR resonances and hence on the experimental uncertainties of the spin relaxation data in the N-terminal region. Importantly, the {^{1}H}-^{15}N heteronuclear NOE values of GABARAP and GABARAP-I G116C are virtually identical within error (Figure 1), indicating that the dynamics of the backbone on the pico- to nanosecond time-scale is fully conserved upon cleavage of L117 by the ATG4 family of proteases.

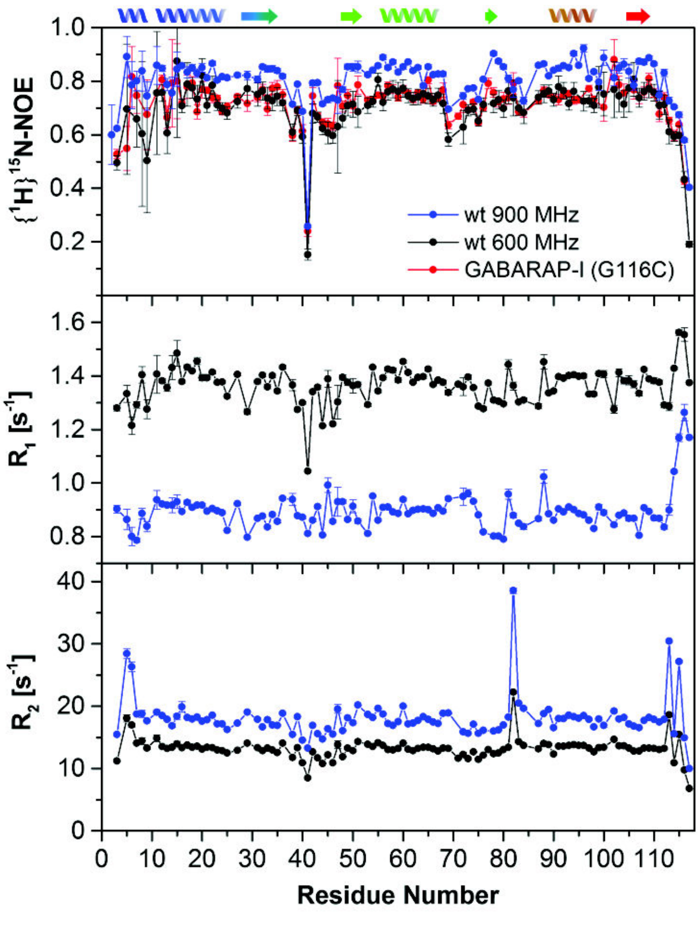


Figure 1: Backbone amide ^{15}N relaxation data of GABARAP as a function of residue number. (A) {^{1}H}-^{15}N NOE values of GABARAP and GABARAP-I G116C (red), (B) longitudinal relaxation rates R_1, and (C) transverse relaxation rates R_2, at 600 MHz (black) and 900 MHz (blue). Regular secondary structure elements are depicted on top of panel (A).

Model-free analysis. Detailed analysis of the ^{15}N relaxation data by the Lipari-Szabo-type "model-free" approach (eq 19) reveals that the rotational diffusion tensor is not completely isotropic but is best described by an ellipsoid model with principal components of D_{xx} = 1.60×10^7 rad/s, D_{yy} = 1.68×10^7 rad/s, D_{zz} = 1.93×10^7 rad/s at 25.0°C (Tab. S1), corresponding to five rotational autocorrelation times of ρ_{-2} = 10.19 ns, ρ_{-1} = 10.00 ns, ρ_0 = 9.76 ns, ρ_{+1} = 9.08 ns, and ρ_{+2} = 9.07 ns. The orientation (principal axes) of the rotational diffusion tensor agrees closely with the overall shape of the protein as represented by the tensor of inertia of the crystal structure of GABARAP (PDB 1GNU) (data not shown). In the spherical approximation this rotational diffusion tensor reduces to an isotropic rotational diffusion constant D_{rot} = $(D_{xx}+D_{yy}+D_{zz})/3$ = 1.74×10^7 rad/s, corresponding to a global rotational correlation time ρ_{global} = $1/(6D_{rot})$ = 9.60 ns. Assuming a viscosity of η = 0.911 mPa s interpolated for 10% D_2O at 25.0°C [63] this corresponds to a hydrodynamic radius (eq 28) of $R_{h,rot}^{(NMR)}$ = 21.8 Å, indicating predominantly monomeric GABARAP molecules. Besides the global diffusion properties, model-free analysis provides generalized order parameters, S^2, which report on the local backbone mobility (Figure 2A, blue squares). The average order parameter of $\langle S^2 \rangle$ = 0.84 ± 0.01 indicates high motional restriction of the orientation of most amide bond vectors, particularly in the regular secondary structure elements (Figure 2A). In contrast, residues at the N-terminus, at the C-terminus, and in the two loops between strands β_1 and β_2 and between helix α_3 and strand β_3 exhibit enhanced flexibility as revealed by low order parameters (Figure 2A). Most notably, I41 in the loop close to the C-terminal region shows a very low S^2 value of 0.37, which is in the same range as the order parameters of the C-terminal residues G116 and L117.

NMR translational diffusion. The hydrodynamic radius of GABARAP was additionally determined by translational diffusion experiments analyzed by two different methods. When referenced to dioxane as a standard molecule with known hydrodynamic radius of 2.12 Å [64] the hydrodynamic radius is determined to be $R_{h,trans}^{(NMR)}$ = 19.2 Å ± 0.3 Å, corresponding to ρ_{global} = 6.6 ns ± 0.4 s under these experimental conditions. The second approach based on the Stokes-Einstein relation (eq 29) requires knowledge of the solvent viscosity, which is estimated to be η = 0.911 mPa s for a mixture of 90% H_2O and 10% D_2O [63]. In this case the hydrodynamic radius is calculated as $R_{h,trans}^{(NMR)}$ = 20.7 Å ± 0.2 Å, corresponding to ρ_{global} = 8.2 ns ± 0.3 ns under these conditions.

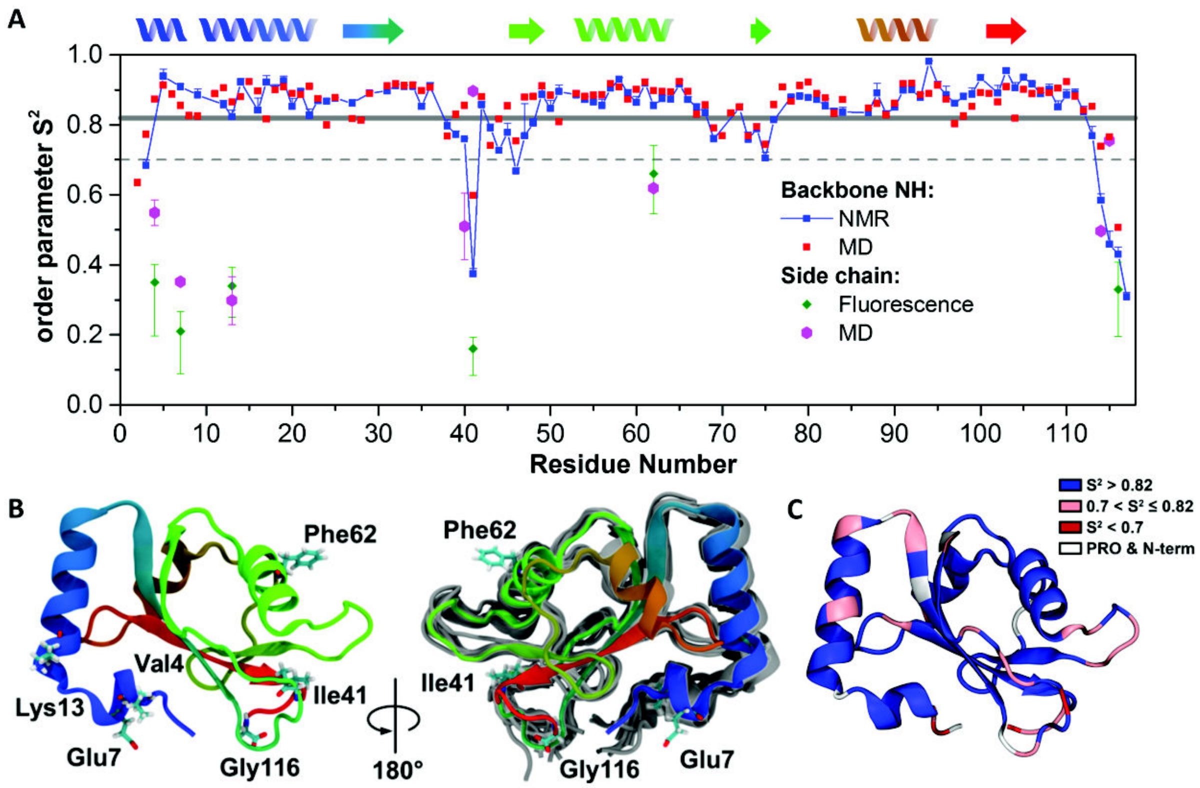


Figure 2: Picosecond-nanosecond dynamics of GABARAP. (A) Comparison of backbone amide bond vector generalized order parameters S^2 determined from ^{15}N NMR spin relaxation (blue squares), order parameters S^2 of the backbone amide bond vector (red squares) as well as of the average side-chain (magenta hexagons) from MD simulations, and order parameters S^2 for BFL attached to cysteine side-chains from fluorescence spectroscopy (green diamonds). The regular secondary structure elements are indicated above the panel. Error bars denote standard error. (B) Structural representation of GABARAP with the coloring changing from blue for the N-terminus via green to red for the C-terminus. The side chains of residues discussed in detail in this work are shown as sticks. In the right figure, several snapshots (in grey) from the MD trajectories are presented, revealing flexible regions. (C) GABARAP is colored based on the backbone S^2 using four discrete categories as indicated by the legend.

4.2 Fluorescence spectroscopy: TCSPC and FCS

We apply two complementary fluorescence techniques [80, 81] to map depolarization dynamics of proteins from picoseconds to milliseconds to unravel their global correlation time and internal dynamics. Here we attached the dye BODIPY FL as extrinsic fluorophore to individual cysteine side-chains for four GABARAP variants with mutations at the residues

V4C, E7C, K13C, I41C, or F62C of full-length GABARAP and G116C of GABARAP-I. While time-resolved fluorescence anisotropy experiments by time-correlated single-photon counting (TCSPC) on the ensemble-level are well suited to resolve rotational relation times faster than the excited state lifetime of the fluorophore, τ_e, polarization-resolved fluorescence correlation spectroscopy (pFCS) on the single-molecule level are especially useful to map depolarization motions on time-scales longer than τ_e. To gain sufficient contrast in polarization-resolved correlation curves for long correlation times, pFCS requires that the fluorophore be attached rather rigidly to the protein so that the fast local motions are restricted and the global motion can be sensed. Figure 3A displays the sterically accessible volume [65] of the dye attached to the variants I41C and F62C using a free diffusion dye model to describe the spatial population density of the dye [82]. The accessible volume is limited by the short linker of BODIPY FL (six spacer atoms including C_β; in the AV simulations the total linker length L(C_β - chromophore center) was set to 10 Å), which is a compromise between minimal disturbance of the protein by the dye and sufficient rigidity of the attachment.

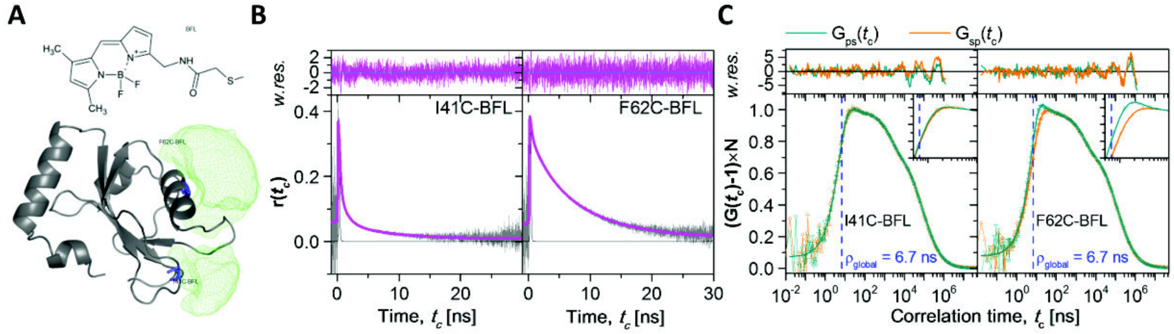


Figure 3: TCSPC and polarization resolved FCS studies of cysteine variants of GABARAP labeled with BODIPY-FL. (A) GABARAP with the volume accessible to BODIPY-FL (structure shown on top) attached to either F62C or I41C (green mesh). (B) Time-resolved anisotropy $r(t_c)$ of the variants I41C-BFL and F62C-BFL from TCSPC. The variant F62C is characterized by a slower anisotropy decay and a larger amplitude r_3b of the global rotational diffusion term than I41C (Tab. S2B). (C) Cross-correlation curves $G(t_c)$ from pFCS were fit by a rotation model (ρ_{global} is marked by the dashed blue line). Note the difference in contrast between $G_{sp}(t_c)$ and $G_{ps}(t_c)$ of variants F62C-BFL and I41C-BFL (see insets for magnification) caused by the significantly higher internal mobility of the dye in the case of I41C-BFL. Top panels in (B) and (C) show the weighted residuals of the fit.

TCSPC. Fluorescence lifetime and anisotropy measurements indicated that our labeling strategy was successful and the dye properties are only weakly perturbed by the coupling. Although we observed multi-exponential fluorescence decays of the tethered BODIPY FL for all positions (see Tables 1A and S2C), which were fitted with three exponential components (eq 31), in most GABARAP variants the fluorescence decay of BODIPY FL has only small fractions of shorter fluorescence lifetimes resulting from quenching. This leads to large species-averaged fluorescence lifetimes $\langle\tau\rangle_x$ ranging between 5.2 and 5.7 ns and by a high fraction of unquenched dye species $x_1 \approx 87\%$, while the lifetime of the free dye is $\tau \approx 5.9$ ns. The variant GABARAP-I G116C with BODIPY FL at the flexible C-terminus has exceptional properties with only 65% of unquenched species and $\langle\tau\rangle_x = 3.8$ ns (see Table S2C).

The time-resolved anisotropy decay curves are more distinct between the variants (Table 1A). The most significant differences are observed between the anisotropy decay curves $r(t_c)$ of the dye attached to positions I41C and F62C (Figure 3B), showing a much faster decay for I41C. The order parameters of the six variants determined from the analysis of $r(t_c)$ using eqs 30-32 and transformation according to eqs 3 and 17 are compiled in Tab. 1A (decays see Figure S3B, fit parameters see Tab. S2) and plotted in Figure 2 (green diamonds). The anisotropy decays exhibit two local relaxation processes with ρ_{fast} ranging from 0.21 ns to 0.32 ns for the different variants and ρ_{slow} ranging from 1.4 ns to 4.1 ns, with a global rotational correlation time $\rho_{global} = 9.0$ ns. The S^2 values at residues V4C, E7C, K13C, I41C, F62C of GABARAP and G116C of GABARAP-I show various extents of rigidity, with F62C-BFL being the most rigid residue ($S^2 = 0.66$) and I41C-BFL being the most flexible ($S^2 = 0.16$). Note that determination of slow global correlation times by fluorescence anisotropy decay analysis is extremely sensitive to instrumental factors such as the G-factor in eq 30. Therefore, special care was taken to calibrate them precisely. In repeat measurements of the variant F62C with the largest S^2 value, the global correlation time was determined to be $\rho_{global} = 9.2$ ns ± 0.4 ns at 20°C; i.e., 8.0 ns ± 0.4 ns recalculated for 25°C, which was converted to a hydrodynamic radius for rotation $R_{h,rot}^{(r(t))} = 20.7$ Å ± 1.0 Å by eq 28.

pFCS. Moreover, we studied the most distinct GABARAP variants labelled at the positions I41C and F62C by analyzing the two polarization resolved full cross correlation curves $G_{sp}(t_c)$ and $G_{ps}(t_c)$ (eqs 33 and 34, Figure 3C). The flexibility of the I41C side-chain results in virtually complete loss of polarization contrast, so that the difference between the cross-correlation curves $G_{sp}(t_c)$ and $G_{ps}(t_c)$ and the corresponding correlation amplitudes b_{rot} (eq 34) becomes very small in Figure 3C, whereas the two curves for F62C are clearly distinct in the time range of 10 ns which is a hallmark for fast rotation motions (Table S3). These findings agree fully the time-resolved anisotropy measurements. A detailed curve analysis including simulations for calibration of the cross-correlation effects between the antibunching and rotational diffusion terms (see section 3.8 and Figure S2A, B)) revealed $\rho_{global} = 6.7^{+2.3}_{-1.3}$ ns at 26°C for the F26C variant ($\rho_{global} = 6.9^{+2.4}_{-1.3}$ ns recalculated for 25°C).

FCS translational diffusion. In addition, we measured also fluorescence correlation curves for determining the translational diffusion coefficient of labeled GABARAP from the FCS diffusion term (see SI, section 4). We measured correlation curves at a range of low irradiances (Figure S4) to exclude saturation effects for determining an average translational diffusion time $\langle t_{diff} \rangle$ = 0.720 ± 0.013 ms, which corresponds to a translational diffusion coefficient D_{trans} = 1.18 ± 0.09 10^{-10} m^2 s^{-1} at 25°C and a hydrodynamic radius of $R^{(FCS)}_{h,trans}$ = 20.9 Å ± 1.9 Å (eq 29).

Sensing molecular flexibility by order parameters. The S^2 values of the various labeling sites covered a wide range between 0.66 and 0.16. To explore the origin of this spread, the dye motion and its connection to the transition dipole moment has to be analyzed. Taking into account that the transition dipole direction of the BODIPY FL chromophore is parallel to the molecular long axis [83] and that the attachment of the linker to the fluorophore is asymmetric (Figure 3A), the wobbling-in-cone model [26, 65, 84] appears to be most appropriate to describe the fluorophore motion, i.e., the depolarization is primarily caused by the motions of the dye weak fluorescence quenching, this reveals that there are only few interactions (10-20% see Tab. S2C) of the dye with the protein surface at position 62. Thus, the spatial population density of the dye is well approximated by the simple free diffusion model in our accessible volume simulations (Figure 3A). Using this model, the trends of the average side chain S^2 order parameters from MD and fluorescence (Figure 2) agree well except I41. Further details will discussed in section 5.2 *Internal Dynamics*.

4.3 MD simulations

Backbone S^2 order parameters. The N–H bond vector order parameters, S^2, computed from multi-exponential fits to the total correlation function $C(t_c)$ derived from the MD simulations are generally in very good agreement with those derived from NMR spectroscopy (Figure 2A). Larger deviations are mainly present at the termini and in other flexible regions. The S^2

values obtained from the internal correlation functions as C_{int}(5 ns) (eq 35) are not shown because they are virtually indistinguishable from the S^2 values from the multi-exponential fits to $C(t_c)$, with a root mean square deviation of 0.014 and a Pearson correlation coefficient of 0.99. Rigid N–H bonds are often but not always overfitted when more than the global exponential decay is used, while more complex models with one or two additional internal decay times fit the flexible N–H bonds best (Figure 4A, Figure S5A-C). The most rigid part of GABARAP can be found in the core of the protein, formed by the β-sheet and parts of α_4, where most of the N–H bonds can be fitted by a global exponential decay only (Figure S5D). Many of the residues located in the various loops, on the other hand, require three decay times due to their inherent flexibility. In three of the four helices there are also residues that had to be fit by three decay times even though the corresponding backbone S^2 values are larger than 0.8, such as residues E17 and K24 in helix α_2. Panels B to D in Figure 4 show histograms of the predicted correlation times. The fastest (i.e., initial) decay times for the N–H bond vector reorientation are shorter than the interval at which the structures sampled during the MD simulations were saved (5 ps). They are a result of fast internal motions like bond angle bending, which lead to fast changes of the N–H bond vector orientation represented in our fits by the order parameter S_{init}^2. The average value of the predicted global rotational correlation times, ρ_{global}, is 2.84 ns with a standard deviation of 0.20 ns and a standard error of 0.02 ns. The values for ρ_{global} cluster narrowly around this average value for the residues fitted with only a global decay, while the range of ρ_{global} for the residues fitted with internal decay times is larger, ranging from about 2 ns up to 3.4 ns (Figure 4B). The simulated global rotational correlation times have to be scaled by a factor of 2.80 ± 0.04 because the viscosity of the TIP3P water model used in our MD simulations is lower than that of real water [85]. Therefore, the predicted global rotational correlation time is ρ_{global} = 7.95 ns ± 0.56 ns at 27°C (8.32 ns ± 0.59 ns at 25°C), corresponding to a hydrodynamic radius of $R_{h,rot}^{(MD)}$ = 20.9 Å ± 0.1 Å. In Figure S6A the calculated ρ_{global} values are color mapped onto the protein structure. Because the rotational diffusion tensor of GABARAP as determined by model-free analysis of the ^{15}N NMR relaxation data is not completely isotropic (see above), ρ_{global} is expected to be longer (slower) or shorter (faster) than the isotropic value depending on whether the corresponding N-H bond vector is aligned parallel or perpendicular, respectively, to the fastest principal axis (D_{zz}) of the rotational diffusion tensor. The strong correlation with the rotational autocorrelation times predicted from the experimentally determined rotational diffusion tensor (Figure 5C) demonstrates that the variance in ρ_{global} extracted from the MD trajectories primarily reflects the anisotropy of rotational diffusion. The single conspicuous outlier in the distribution of ρ_{global} from MD simulations (Figure 4B) is G116 with 1.96 ns, for which the internal motions appear to be sufficiently slow to be entangled with the global rotational motion (i.e., the factorization of eq 7 is not valid in this case) and hence complicate the interpretation of the multi-exponential fit of its N-H bond vector correlation function. For the internal correlation times, we observe the same trend as for ρ_{global}: ρ_{slow} has a larger variance for the residues fitted with two internal decays compared to those fitted with a single internal decay (Figure 4C). For the latter, the ρ_{slow} values are almost all below 120 ps, while for the former the ρ_{slow} values are quite evenly distributed between 120 ps and 1.5 ns. The correlation times ρ_{fast} are rather small with values between 7 ps and 110 ps (Figure 4D).

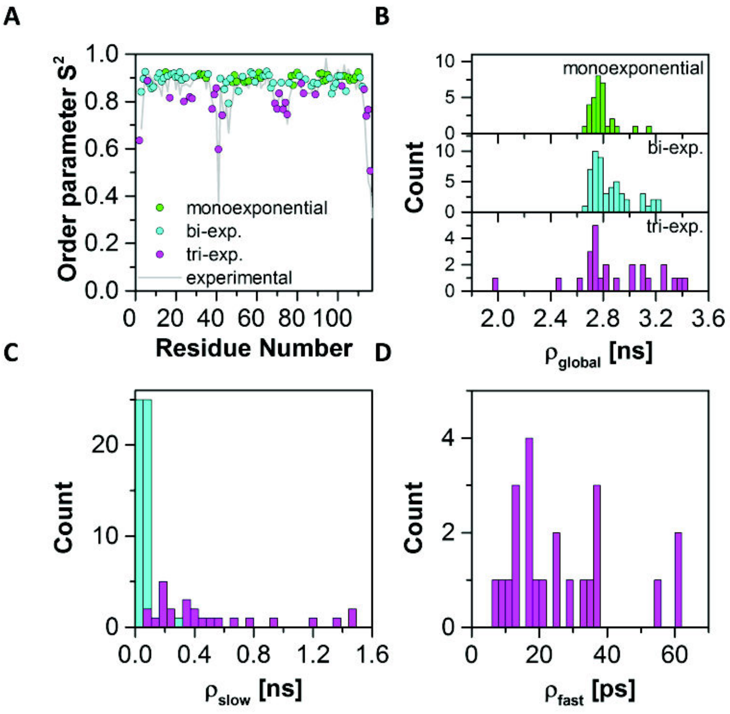


Figure 4: Order parameter S^2 (A) and correlation times (B) to (D) extracted from MD trajectories by fitting the global correlation functions to one (green), two (blue), or three (magenta) exponential decays. The most complex model that did not overfit the bond vector correlation function was selected for each amino acid. This accounts for one global rotational correlation time ρ_{global} and maximum two internal correlation times ρ_{slow} and ρ_{fast}, the distributions of which are shown in panels (B) to (D).

Side-chain S^2 order parameters. The order parameters for the side chains were calculated from MD simulations in the same way as for the N–H bond vectors. In Figure 2A the results are shown for those residues for which S^2 values were also determined using fluorescence, while in Figure S7 the MD results for all side chains can be seen. The MD derived S^2 values were averaged over the individual side-chain bonds. The numbers in Table 1B indicate that the S^2 values for the individual bonds per side chain are generally close to each other so that their average is representative for that side chain. Exceptions are the long side chains of solvent-exposed and charged residues like K13 and R40. Here, some bonds are rather flexible while others do not reorient much. Table 1B further shows that, unlike for the N–H bond vectors, in most cases three exponential decays are required for fitting the total correlation function $C(t_c)$ given in eq 17, which reflects the generally higher flexibility of the side chains compared to that of the protein backbone. The average and standard deviation of the S^2 values for the side chains is 0.69 ± 0.21 while the same values for the backbone amide groups are 0.86 ± 0.07. However, while for most residues the side chains are more flexible than the backbone (see Figure S7), there are cases like I41 for which it is the other way around. In general, the flexibilities of the N–H bond vectors and the side chains are not correlated as the low Pearson correlation coefficient of 0.206 between the S^2 values for the backbone and side chains reveals.

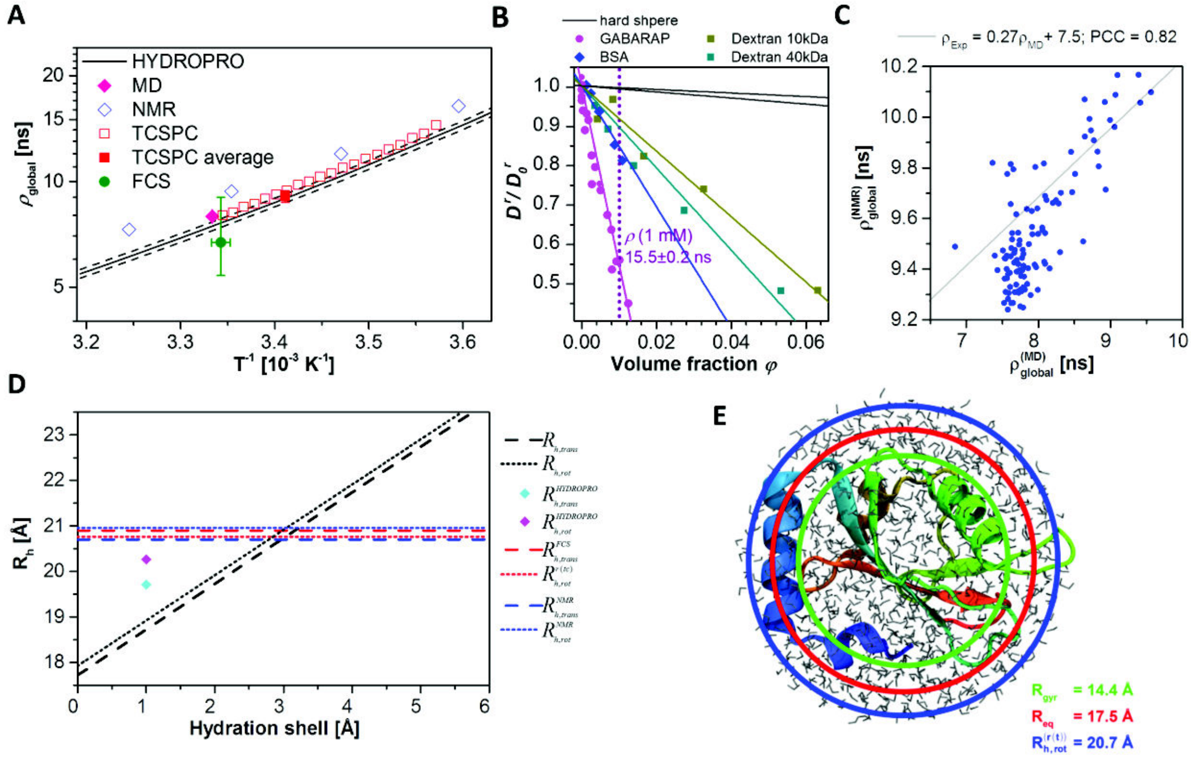


Figure 5: The global rotational correlation time, ρ_{global} as function of temperature. (A) ρ_{global} obtained by NMR, fluorescence spectroscopy and MD shows good agreement with structure-

based prediction tools (HYDROPRO) across various temperatures. (B) Molecular interactions like crowding (BSA, Dextrans) or oligomerization (GABARAP) showing the relation between relative rotational diffusion ($D^r / D_0^r = \rho_0 / \rho$) and concentration (volume fraction, φ) and type of cosolvent as obtained from fluorescence anisotropy decays. The difference in interaction strength is visualized as slope of the solid lines (h_{GABARAP} 41.2±1.6, h_{BSA} 13.9±1.0, $h_{\mathrm{Dextran10}}$ 7.5±0.5, $h_{\mathrm{Dextran40}}$ 9.4±0.6). A large deviation from the theoretical relation for hard spheres (black line; h = 0.41 to 0.7) is clearly visible. (C) The strong correlation between the distribution of residue-specific ρ_{global} calculated from MD trajectories (abscissa) and back-calculated from the fully anisotropic rotational diffusion tensor from NMR relaxation (ordinate) with a Pearson correlation coefficient (PCC) of 0.82 reveals that NMR spectroscopy and MD simulations are both sensitive to the anisotropy of the rotational diffusion of GABARAP. Tyr115 and Gly116 have been omitted from this analysis because their conformation in the PDB structure 1GNU used as a basis for the NMR analysis appears to be a crystallization artifact, as indicated by MD simulations. (D) Hydrodynamic radii $R_{h,rot}$ (dotted black line) and $R_{h,trans}$ (dashed black line) calculated for the prolate ellipsoid of revolution with semi-axes of 23.0 Å and 15.2 Å (equivalent sphere: R_{eq} = 17.5 Å), which has approximately the same tensor of inertia as the crystal structure of GABARAP (PDB 1GNU) after addition of hydrogens, as a function of the thickness of the hydration shell. Diamonds represent the hydrodynamic radii predicted by HYDROPRO, horizontal lines the experimentally determined hydrodynamic radii. (E) Solvated GABARAP is shown (water molecules in grey). The green circle indicates the radius of gyration, R_{gyr}, while the red circle represents the radius R_{eq} of the equivalent sphere of equal volume. The hydrodynamic radius as obtained from fluorescence anisotropy decays, $R_{h,rot}^{(r(t))}$ is shown as a blue circle.

The MD side-chain order parameters match the fluorescence derived S^2 values remarkably well, with the exception of I41 (Figure 2A), whose side-chain is almost completely rigid in the MD simulation (S^2=0.90), while the fluorescence chromophore attached to this residue is almost completely flexible (S^2=0.16). The MD trajectories show that the side chain of I41 is deeply buried inside a hydrophobic pocket pointing towards the protein core, which reduces the solvent accessible surface area (SASA) of I41 by 89% compared to an isolated isoleucine. The stable hydrophobic packing in this position prevents any motions of larger amplitude. The backbone N–H bond of I41, on the other hand, shows very low order parameters in MD

(S^2=0.60) and NMR (S^2=0.38) because it fluctuates between two stable conformations by rotations about the R40 Ψ and I41 Φ backbone dihedral angles (Figure 6A), facilitated by the fact that this NH group is not involved in hydrogen bonding. All other backbone dihedral angles of residues 39 to 42 are restricted to a single stable conformation (Figure. 5B). These restrictions are due to interactions of the side-chain of K35 with the backbone oxygen of I41 as well as hydrogen bonds between the side-chain of D111 and both the NH group of R40 and its side-chain. The most obvious explanation for the discrepancy between MD and fluorescence order parameters for the I41 side-chain is that the mutation of I41 to cysteine with the bulky chromophore BODIPY FL attached to it disrupts the hydrophobic packing as present in the wild-type protein, thereby causing the chromophore to become exposed to the solvent and highly mobile. For comparison, the nearest neighbor R40 is more solvent-exposed with a relative SASA of 52% and therefore more mobile, leading to a simulated side-chain S^2 value of 0.51 (Figure 2A). Figure S8 confirms that the S^2 values of the side-chains are inversely correlated to their SASA. Large side-chain S^2 values exceeding 0.8 are preferentially found in the core of the protein, whereas values below 0.4 occur mostly for solvent-exposed side-chains on the protein surface. In Figure S8D the side-chain S^2 values are plotted versus (1–SASA), where SASA is given for the side-chains relative to their solvent accessibility in the isolated amino acids. The correlation between S^2 and (1–SASA) is confirmed by a Pearson correlation coefficient of 0.71. Another noteworthy residue is F62 in helix α_3, for which we find a medium solvent accessibility of 51% but with the side-chain being oriented towards the solvent (Figure 2B) and an intermediate side-chain mobility of S^2 = 0.62. Together with the highly rigid backbone in this helix, this was our motivation to probe the side-chain flexibility in this position experimentally with a fluorescence chromophore. The almost perfect agreement between simulation and experiment confirmed that our fluorescence approach is able to measure variations in side-chain flexibility as long as their hydrophobic packing is not disrupted by the introduction of the chromophore.

Interpretation of the internal decay times. In principle, the MD data allows to extract the origin of the motions leading to ρ_{fast} and ρ_{slow} for all N–H bond vectors and side-chains. Here, we limit this analysis to a few representative examples. One of them is the aforementioned fluctuation of the backbone of R40/I41 between two stable states of the peptide plane (Figure 6). An analysis of the lifetimes of these two states reveals that the fast correlation time ρ_{fast} = 13 ps results from the fast interconversion between these two (R40 Ψ, I41 Φ) states. The slow correlation time, which is an order of magnitude larger than ρ_{fast} (ρ_{slow} = 173 ps), arises from the motion of the loop to which I41 belongs. A similar situation is evident for the N–H bond vector motion of A75, which was fitted with the two internal decay times ρ_{slow} = 651 ps and ρ_{fast} = 36 ps. Again, the slow and fast correlation times results from the flexibility of this loop region and internal backbone motions, respectively. Here, it is mainly the torsion angle A75 Φ, which fluctuates between -140° and -80°. However, since the carbonyl oxygen of D74 forms one or more H-bonds to the side-chain of R65 80% of the time, this motion is somewhat slower than the backbone motion in I41.

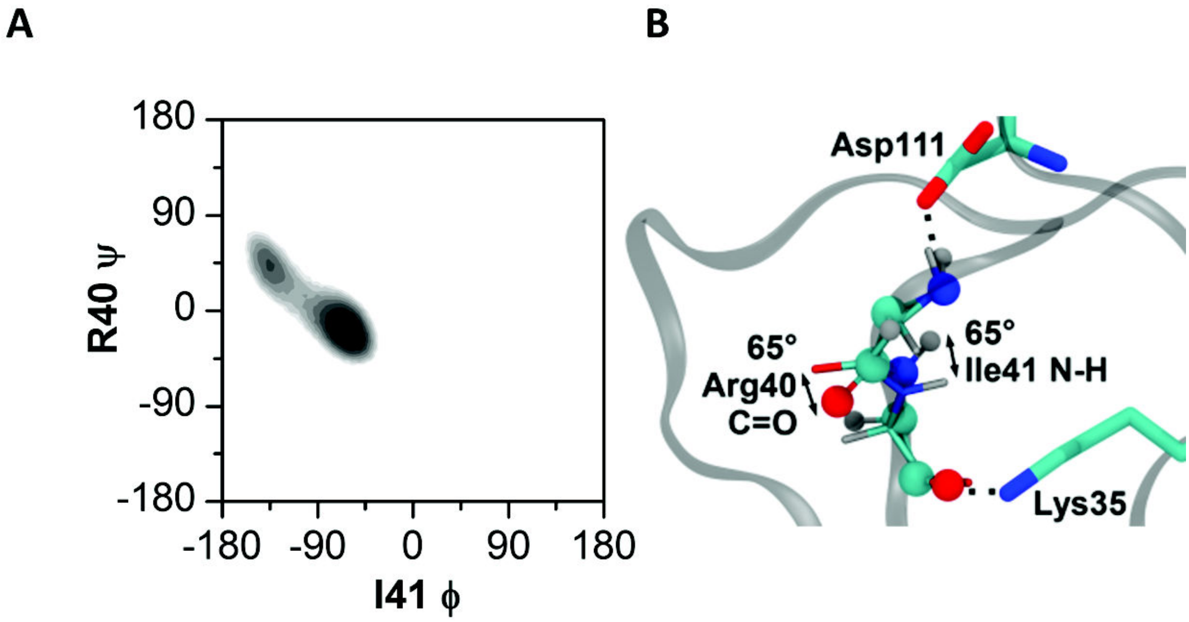


Figure 6: (A) The Ramachandran plot (logarithmic scale) for Φ of I41 and Ψ of R40 reveals that the peptide bond plane between residues 40 and 41 fluctuates between two discrete conformations (with different stabilities, i.e., populations), which are shown in (B). The protein is shown as grey band and the backbone atoms of R40 and I41 and the side chain atoms of K35 and D111 are shown in Licorice and colored by element (carbon, cyan; hydrogen, white; oxygen, red; nitrogen, blue) while the other stable state of the R40-I41 backbone is shown in CPK representation.

5. Discussion

In this work we unraveled the pico- to nanosecond dynamics of GABARAP by an integrated approach using NMR and fluorescence spectroscopy in concert with MD simulations. We provided a common theoretical framework and analysis for the rotational diffusion and local protein dynamics as determined by NMR, fluorescence and MD simulations. A key objective of the present study was to benchmark our integrated approach via extensive comparisons between the different methods to reveal synergies, differences, and limitations of each method. Here, the atomistic picture obtained from MD simulations turned out to be extremely valuable for the interpretation of the experimental results from NMR and fluorescence spectroscopy and the explanation of differences between these techniques.

5.1 Common theoretical framework and benchmarking of hydrodynamics

Remarkably, our benchmark study produced very similar results for the overall rotational correlation times ρ_{global} (Table 2, Figure 5A) and the S^2 order parameters describing the amplitude of internal motions (Figure 2A) obtained with the three approaches under study, demonstrating the reliability of each of these methods. Nonetheless, we also observed several instructive differences for one of the three approaches, which could readily be explained using the other two methods tested here, thus providing us with valuable insight and guidelines for future studies. More specifically, ^{15}N NMR relaxation spectroscopy offers high spatial resolution resulting from an abundance of probes that sample both rotational diffusion (including its anisotropy) as well as internal backbone motions in all parts of the protein, but requires relatively high sample concentrations and long measurement times. Fluorescence spectroscopy can be employed at a wide range of protein concentrations, up to single molecule detection, and primarily samples internal motions of the side-chain to which the chromophore is attached (Figure 2A).

Hydrodynamics: Viscosity, crowding and self-association. The backbone dynamics on the pico- to nanosecond time-scale is dominated by rigid-body overall rotational diffusion of GABARAP. We found a remarkable agreement between rotational correlation times ρ_{global} derived from fluorescence experiments, MD simulations and the theoretical values predicted by HYDROPRO (Table 2, Figure 5A). Moreover, the two hydrodynamic radii R_h from both fluorescence methods are very similar: translational diffusion $R_{h,trans}^{(FCS)}$ = 20.9 Å ± 1.9 Å using FCS and rotational diffusion $R_{h,rot}^{(r(t))}$ = 20.7 Å ± 1.0 Å using fluorescence anisotropy decays.

Notably, the global rotational correlation time of ρ_{global} = 9.4 ns ± 0.2 ns ($R_{h,rot}^{(NMR)}$ = 21.8 Å) determined from ^{15}N NMR relaxation spectroscopy is 18% larger than the value of ρ_{global} =

8.0 ns ± 0.4 ns ($R_{h,rot}^{(r(t))}$ = 20.7 Å ± 1.0 Å) calculated from TCSPC for an aqueous solution of 100% H_2O at 25.0°C (Tab. 2). In fact, this appears to be a very typical case because a literature survey of 17 proteins studied by both fluorescence and NMR spectroscopy showed that NMR relaxation resulted in approximately 18% larger rotational autocorrelation times than fluorescence anisotropy decay on average [86].This observation was attributed to the higher sample concentrations typically used in NMR spectroscopy to obtain a better signal to noise ratio. The authors of this study pointed out that there are three different mechanisms contributing to any concentration dependence of rotational diffusion in a non-ideal (i.e., not infinitely dilute) sample, namely (i) general viscosity effects (also referred to as crowding or microviscosity), (ii) heterogeneous self-association mediated by non-specific interactions, and (iii) specific self-association (dimerization, trimerization, etc.) mediated by specific interactions.

The fact that the hydrodynamic radius of $R_{h,trans}^{(NMR)}$ = 20.7 Å ± 0.2 Å obtained from NMR translational diffusion experiments on 0.5 mM GABARAP agrees much better with the results from fluorescence spectroscopy and MD simulations supports the notion that the elevated rotational correlation time determined from ^{15}N relaxation experiments is indeed caused by the twofold higher sample concentration of 1.0 mM GABARAP. An even smaller hydrodynamic radius of $R_{h,trans}^{(NMR)}$ = 19.2 Å ± 0.3 Å is obtained if any increase in viscosity caused by the protein concentration is eliminated from the data analysis by using dioxane as an internal standard for translational diffusion, although this result should be interpreted with caution because we cannot completely rule out the possibility that GABARAP with its hydrophobic binding pockets for aromatic side-chains might interact with the heterocyclic compound dioxane, which in turn would result in the hydrodynamic radius of GABARAP being artificially underestimated. These observations prompted us to use fluorescence spectroscopy to investigate the effect of sample concentration on the global rotational diffusion of GABARAP in more detail by bridging the gap in the concentration range from the micromolar concentrations typically used in fluorescence experiments to the millimolar concentrations typically employed in NMR spectroscopy. To test for the presence of crowding effects we measured the global rotational autocorrelation time of GABARAP-F62C-BFL by TCSPC in the presence of three crowding agents of different molecular size (10 kD dextran (Dextran10), 40 kD dextran (Dextran40), and bovine serum albumin (BSA, 66 kD)). Following the procedures of Roosen-Runge et al. [87] (see SI, section 2) we plot the relative rotational diffusion coefficient, $D_{rot}(\varphi)/D_{rot}(0) = \rho_{global}(0)/\rho_{global}(\varphi)$ versus the volume fraction φ of the total sample volume that is occupied by the protein. The interaction strength is visualized as the negative slope h of the solid lines (eq S6) in Figure 5B: $h_{\mathrm{Dextran10}} = 7.5 \pm 0.5$, $h_{\mathrm{Dextran40}} = 9.4 \pm 0.6$, $h_{\mathrm{BSA}} = 13.9 \pm 1.0$. The theoretical relation for hard spheres (Figure 5B, black line) results in a considerably smaller interaction strength with h = 0.41 to 0.70 depending on the size ratio [88], which indicates that BSA and dextrans act as strong crowding agents and their effect depends on their molecular size. However, the decrease of the rotational diffusion coefficient of GABARAP with increasing volume fraction corresponding to a concentration range of 0.5 µM to 1.3 mM is even stronger with $h_{\mathrm{GABARAP}} = 41.2 \pm 1.6$, and ρ_{global} increases significantly at GABARAP concentrations above 0.1 mM (see SI, section 3 and Figure S9). The fact that this effect is stronger than observed for commonly used molecular crowding agents suggests that the elevated rotational autocorrelation times observed in NMR relaxation and TCSPC fluorescence experiments at high concentrations probe not only molecular crowding but also self-association of GABARAP molecules. Our experimental data does not readily discriminate between heterogeneous self-association mediated by non-specific interactions and specific self-association (dimerization, trimerization, etc.) mediated by specific interactions. Applying a simple dimerization model that assumes a prolate shape for the dimer with an axial ration of 2:1 we estimate the dissociation constant to be in the low millimolar range (Figure S9). Self-association of GABARAP has been reported in the literature before and implicated in binding to tubulin and promoting its oligomerization into microtubules as well as $\mathrm{GABA_A}$ receptor clustering [45, 89, 90]. More recent studies have shown that upon lipid-conjugation the proteins from the GABARAP/MAP1LC3/Atg8 family oligomerize together with associated phospholipids and other autophagy-related proteins, thereby mediating membrane tethering and hemifusion during the formation of the autophagosomal membrane compartments required for protein degradation by (macro-)autophagy [91, 92]. Note that while the millimolar bulk dissociation constant governing the self-association of free soluble GABARAP is clearly too weak to be physiologically relevant, anchoring of lipidated GABARAP-II molecules to developing autophagosomal membranes can easily result in GABARAP clusters with a local concentration sufficient for effective tethering to other GABARAP-decorated membranes or to the microtubule cytoskeleton for vesicular trafficking. The insight gained from the integrated NMR, fluorescence and MD study presented here constitutes an important stepping stone for a detailed investigation of the structural mechanism and functional role of the self-association of GABARAP-II anchored to suitable membrane mimetics such as nanodiscs [93] or in cells.

The obvious solution to avoid elevated rotational diffusion correlation times caused by concentration-dependent effects in NMR relaxation spectroscopy is to carry out the experiments at very low concentrations or – better yet - at a series of different protein concentrations. This strategy may well be desirable and feasible for some projects. Unfortunately, the typical measurement time required to record a complete set of ^{15}N NMR relaxation data at two different static magnetic fields is about two weeks, and it scales approximately inversely proportional to the square of the sample concentration if crowding and aggregation are neglected. As a consequence, this strategy is not only extremely time-consuming and costly, it is also very vulnerable to even slow changes in sample composition such as sample degradation over time that may no longer allow a combined quantitative analysis of the experimental data. Fortunately, the faster transverse relaxation rates of any stable oligomers are associated with a substantially lower signal in the NMR spectra compared to the monomer, and the experimentally determined relaxation rates are thus expected to predominantly reflect the monomeric form even in the presence of a moderate fraction of oligomers. Indeed, the rotational correlation time of ρ_{global} = 9.4 ns obtained from ^{15}N NMR relaxation experiments on 1.0 mM GABARAP is only 18% larger than the corresponding value of ρ_{global} = 8.0 ns from TCSPC experiments on a dilute sample, which is much smaller than expected on the basis of the concentration-dependent TCSPC experiments (Figure 5B), and such an overestimation by 18% appears to be a very typical case in the literature [86]. Unless an accurate value for ρ_{global} from NMR relaxation spectroscopy is an absolute requirement, our work suggests that in practice it will often be more efficient to record the NMR experiments on samples with high signal/noise ratio to obtain experimental values for the anisotropy of the rotational diffusion and internal motions and to complement the NMR experiments with fluorescence experiments and MD simulations to investigate concentration-dependent effects on ρ_{global}.

Hydrodynamics: Impact of shape anisotropy. Whereas the determination of rotational diffusion parameters by fluorescence spectroscopy with a single chromophore is limited by the assumption of isotropic rotational diffusion, NMR relaxation spectroscopy and MD simulations probe a large number of bond vectors in the molecule simultaneously and depending on its orientation relative to the rotational diffusion tensor each bond vector will experience a slightly different rotational autocorrelation time if the rotational diffusion tensor is anisotropic (Figure 5C). The molecular shape of GABARAP deviates significantly from spherical geometry (Figure 5E). Accordingly, model-free analysis of the ^{15}N NMR relaxation rates revealed a fully anisotropic rotational diffusion tensor, and this anisotropy is also observed in the MD trajectories (Figure 5C). In terms of magnitude, MD simulations appear to be even more sensitive to the anisotropy than ^{15}N NMR relaxation because the variance in ρ_{global} is larger (Figure 5C), which appears to be a property of the TIP3P water model [94]. Although the deviation from a spherical shape causes the rotational diffusion to be anisotropic, it is not large enough to have any significant effect on the isotropic rotational diffusion coefficient, $D_{rot} = 1/(6\rho_{global})$, and the translational diffusion coefficient, D_{trans}, because the Perrin shape factors F_{rot} and F_{trans} deviate from unity by only a few percent. This allows us to compare rotational diffusion and translational diffusion coefficients obtained from different methods by converting D_{rot} and D_{trans} into hydrodynamic radii using eqs (28) and (29), respectively. The experimentally determined hydrodynamic radii cluster between 20.7 Å and 20.9 Å and are compiled in Figure 5D, together with the predictions from HYDROPRO. Because the fluctuation of the molecular shape of GABARAP-I along the MD trajectory is negligible (radius of gyration R_{gyr} = 14.3 Å ± 0.1 Å), we compared the experimental hydrodynamic radii to those calculated for the prolate ellipsoid of revolution with semi-axes of 23.0 Å and 15.2 Å and an equivalent sphere of R_{eq} = 17.5 Å, which has approximately the same tensor of inertia as the crystal structure of GABARAP (PDB 1GNU) after addition of hydrogen atoms (see section 2.6), plus a hydration layer of variable thickness. The experimentally determined hydrodynamic radii correspond to a hydration shell of about 3 Å (Figure 5D), which is similar to the van-der-Waals radius of a single oxygen atom (2.8 Å). They are slightly larger than $R_{h,rot}$ and $R_{h,trans}$ predicted from HYDROPRO, which uses an implicit hydration shell of about 1.1 Å [68]. It has been suggested in the literature that the effect of the hydration layer may be at least partly due to dielectric friction, which depends on the charge distribution of the protein [95]. In fact, such complex electrostatic effects might also play a role in the concentration dependence of GABARAP hydrodynamics discussed above.

5.2 Internal dynamics

Correlation of backbones and side-chains order parameters. Regarding the internal motion, the three methods under study revealed similar trends of flexible regions as indicated by the order parameter S^2. The fluorescence determined order parameters are generally lower compared to those obtained by NMR spectroscopy caused by the fact that fluorescence is sensitive to side chain dynamics, whereas NMR ^{15}N spin relaxation probes amide backbone motion. In particular residues in the N-terminal sub-domain possess lower backbone flexibility compared to side chains. This hypothesis is confirmed by MD simulated backbone as well as side-chain order parameters S^2. The backbone order parameters S^2 reveal internal motions in loop regions and at the termini, whereas regular secondary structure elements show high rigidity. Interestingly, correlated backbone and side chain are not always correlated. While we observe correlated backbone and side chain dynamics for the C-terminus, the N-terminal region shows a low backbone flexibility but a high side chain flexibility.

Although the order parameters reveal similar regions to be flexible, with only few justified exceptions, the fitted internal correlation times, ρ_{fast} and ρ_{slow}, are inaccurate and thus not comparable for all these methods. In case of NMR spectroscopy these inaccuracies are caused by too many parameters when a model with maximum two internal motions is assumed. The global correlation time can be determined by NMR relaxation rates under consideration of all protein residues, while the internal correlation times are resolved from only one residue. MD simulations, on the other hand, suffer from inaccuracies of the protein force fields, which are known to have difficulties to correctly reproduce the timescales of molecular processes as they were not parameterized for this purpose [96].

Selection of labeling sites. These study shows that MD simulations proved to be essential to critically evaluate the experimental results. The trends of the average side chain S2 order parameters from MD and fluorescence (Figure 2) agree well except for I41 and G116. For these cases MD simulations rationalized the disagreement.

For I41 the remarkably low order parameter of I41 located in a loop region of GABARAP suggests high flexibility. However, MD trajectories revealed that the low S^2 value mainly results from a flip of the R40-I41 peptide plane between two distinct states (Figure 4). While fluorescence spectroscopy further suggests high flexibility of the I41C-BFL side chain, the MD trajectories revealed that the sidechain is stably trapped into a hydrophobic pocket of GABARAP and is therefore almost not mobile. In this particular case, the dye attached to cysteine at position 41 might sterically hinder this interaction with the hydrophobic pocket. Furthermore, the cysteine lacks the affinity for hydrophobic interaction, such that the side chain dynamics probed by fluorescence at position 41 is dominated by the protein backbone. For G116 the order parameter determined by fluorescence spectroscopy reveals a higher rigidity than it is expected for a residue in an unstructured C-terminus and with a low NMR order parameter. This is likely to be caused by interactions of the dye molecule (BFL) with the protein surface, indicated also by a significant reduction of the average fluorescence lifetime (see Table S2C).

In section 4.2 we concluded the dye motion is approximated by the wobbling-in-cone model, which has several practical implications for fluorescence spectroscopic applications. Our integrated order analysis allowed us to interpret fluorescence anisotropy measurements with respect to protein backbone and side-chain flexibility for monitoring its function. Moreover, MD simulations are very helpful for selecting informative and conflict free amino acid residues for labeling by analyzing backbone and side chain motions and sidechain packing. Finally, this methodology has the potential for being also very useful for accurate Förster resonance energy transfer (FRET) measurements, because the presented characterization of the dye labeling sites via order parameters represents an independent validation tool for the choice of an appropriate dye model. Note that the dye description is an essential basis for the accurate analysis of high-precision FRET measurements [82, 97-99].

Biological relevance. Most notably, the C-terminus is highly mobile on the pico- to nanosecond time-scale, indicated by low backbone and side-chain order parameters, which is contrary to the presence of a salt bridge between M1 and L117 connecting the N- and the C-terminus in the GABARAP structure 1GNU. Therefore, this salt bridge is most likely transiently formed allowing an open and closed state of the N- and C-terminus, which was also observed in the GABARAP homologue GATE-16 [24]. This assumption is supported by our MD simulations of GABARAP-I, which revealed that the salt bridge between the terminal charges NH_3^+ and COO^- at M1 and G116, respectively, can form and break.

The C-terminal flexibility is conserved among GABARAP homologues and likely to be a prerequisite for enzyme processing required for lipidation and subsequent membrane anchoring [10]. GABARAP and its homologues are C-terminally processed by proteases of the Atg4 family cleaving L117 in order to expose the C-terminal G116. The heteronuclear NOE values of GABARAP-I G116C reveal similar results compared to the wild-type GABARAP concluding that the mobility of the C-terminus on the pico- to nanosecond time-scale remains unaffected by Atg4B cleavage of L117. This is contrary to Atg8, the yeast homologue of GABARAP, which reveals higher rigidity upon exposure of the C-terminal glycine [23]. GABARAP and Atg8 share a strong structural similarity except of the N-terminus, which is unstructured in Atg8. Thus, the C-terminal dynamics might be affected by the N-terminus as they tend to interact with each other via a salt bridge. However, the propensity for this salt bridge may be reduced in Atg8 due to the positive charge of the C-terminal arginine residue and increases after cleavage of this residue, rigidifying the C-terminus.

Contrary to the correlated backbone (NMR) and side chain (fluorescence) dynamics observed for the C-terminus, anti-correlated behavior is revealed in the N-terminal region showing low backbone flexibility but high side chain flexibility. The N-terminal sidechain motion is likely to be relevant for ligand binding, since GABARAP is known to interact with tubulin and microtubules via its N-terminus [18]. Moreover, the N-terminal dynamics may also play a role for self-association of GABARAP, because crystallization under high salt conditions resulted in an alternate conformation in which the N-terminal region is associated with the hydrophobic binding pockets of a neighboring molecule [45].

5.3 Protein mechanics: detection of potential sites for hinge motions

Structural plasticity and conformational transitions are essential for a multi-functional protein like GABARAP to interact with a multitude of different binding partners. Hinge motions of relatively rigid subdomains about flexible joints [100] can result in large relative rotations, for example by 154° in calmodulin [101]. Such hinges often involve only a small number of flexible residues because even a single backbone torsion angle can potentially provide the required rotational freedom. Therefore we inspected our set of S^2 backbone and sidechain order parameters from MD simulations whether they can be used to characterize protein mechanics and identify potential hinges. For that we computed the average and standard deviation for the backbone amide groups $\langle S_{NH}^2 \rangle = 0.87 \pm 0.05$, excluding the highly mobile residues K2, I41, and G116. For the identification of mobile segments, we found the backbone order parameter minus one standard deviation (i.e., $\langle S_{NH}^2 \rangle < 0.82$) as a suitable threshold criterion, revealing four mobile segments in GABARAP around residues K24, D27, R28, V51, E97, and F104 (Figure S7), all of which contribute to the interface between helix α_2 in the NHD and the central β-sheet in the ULD (Figure 2C). The backbone dihedral angle mobility of the short loop before residues D27 and R28, which connects helix α_2 and strand β_1 and acts as a hinge for the hydrophobic ligand binding pocket between the NHD and the

ULD, is also detected in dihedral angle principal component analysis applied to the MD trajectories (Figure S10). Although the NHD is connected to the ULD via several long-range interactions, most of the side-chains surrounding the hinge itself show only intermediate rigidity (Figure S8B). Plasticity of the hinge between the NHD and ULD is functionally important for GABARAP to be able to accommodate a variety of different ligands in this hydrophobic binding pocket [91]. [45]P26 in this hinge was also found to be an important determinant of the conformational mobility of the NHD in yeast Atg8 [102]. Moreover, additional NHD motions (helix α_1) were suggested by to occur for tubulin binding [45].

Unfortunately, the NMR-derived backbone order parameters in the hinge region of the residues 27 and 28 are inconclusive because no relaxation data is available for R28 due to resonance overlap. However, preliminary single-molecule FRET measurements between NHD and ULD suggest that this hinge acts indeed as a pivot for a large-scale conformational change on the microsecond time-scale (to be published). We therefore propose such a detailed MD-based order parameter analysis as a more general tool to identify short, up to three residue-long loops that might be primed to act as hinges for functionally relevant conformational changes, even if these changes are too slow to be sampled during the length of the MD trajectory.

6. Conclusion and Outlook

In this work we presented an integrated approach using NMR, fluorescence, and MD to study protein dynamics on the pico- to nanosecond timescale. The combination of these three methodologies proved to be tremendously beneficial, because each methodology entails unique strengths and shortcomings: NMR spectroscopy provides insight at atomic resolution on a per residue basis yet requiring highly concentrated and pure protein samples, whereas fluorescence spectroscopy involves residual mutations and dye attachments, but will be the method of choice to study protein dynamics and interactions in complex systems, crowded environments or in the cell. While NMR and fluorescence spectroscopy provide information on the amplitude of dynamics provide movies of protein motions at atomic resolution, enabling insight into protein mechanics and a critical evaluation of experimental data. However, current MD simulations of proteins in explicit solvent are generally limited to a few microseconds, restricting the processes that can be studied by MD to this timescale. Yet the complementarity and synergy of these three techniques allow a remarkably detailed analysis of the mechanics and hydrodynamics of proteins as demonstrated in this work for

GABARAP. To this end, we developed a strategy to compare the different methods by means of the global correlation time and fast local protein dynamics. In particular, our work revealed that MD is a relevant tool to determine which solvent accessible amino acids serve best as probes for fluorescence anisotropy experiments. Thereby, the number of cysteine variants can be minimized and possible errors due to mutation are reduced. Additionally, we showed that pFCS is a well-suited technique in order to complement the results obtained by time-resolved anisotropy due to its sensitivity for slower processes. In future, we want to apply the here established approach integrating NMR, fluorescence spectroscopy, and MD simulations to study protein dynamics on multiple timescales ranging from nano- to milliseconds. Using enhanced simulation techniques in combination with high performance computing, the exploration of protein dynamics on the sub-millisecond timescale is in principle possible. Moreover, we aim to extend our studies of protein dynamics from *in vitro* to live cells. For all three methods pioneering work has demonstrated that they are also applicable in cells [103-106], extending molecular biology to cell biology or mimicking cellular conditions in case of MD simulations [107].

Acknowledgements

Judith Fabig did the QuikChange site directed mutagenesis of GABARAP variant I41C. The authors acknowledge access to the Jülich-Düsseldorf Biomolecular NMR Center. C.M. and O.S. thank the graduate school iGRASP seed and C.M. also the NRW Research School BioStruct for funding. This work was supported by the Deutsche Forschungsgemeinschaft through the Collaborative Research Center SFB 1208 ("Identity and Dynamics of Membrane Systems - from Molecules to Cellular Functions", Düsseldorf). G.S. and B.S. gratefully acknowledge the computing time granted by the JARA-HPC Vergabegremium and VSR commission on the supercomputer JURECA at Forschungszentrum Jülich.

Tables

Table 1A: Species-averaged fluorescence lifetime $\langle\tau\rangle_x$ and fluorescence anisotropy decay parameters according to eqs (17) and (29-31).

Residue	**$\langle\tau\rangle_x$ [ns]**	**S^2_{fast}**	**ρ_{fast} [ns]**	**S^2_{slow}**	**ρ_{slow} [ns]**	**S^2**	**ρ_{global} [ns]**
V4C	5.20	1	0.21	0.63	2.58	**0.35**	9.0
E7C	5.53	1	0.32	0.60	3.31	**0.21**	9.0
K13C	5.34	1	0.23	0.66	1.95	**0.34**	9.0
I41C	5.17	1	0.22	0.47	1.37	**0.16**	9.0
F62C	5.69	1	0.25	0.86	4.05	**0.66**	9.0
G116C	3.77	1	0.31	0.69	4.12	**0.33**	9.0

All fit parameters of the detailed time-resolved fluorescence decay analysis are compiled in Table S2. The decays are plotted in Figure S3B.

Table 1B: MD decay parameters for side-chain bonds of residues shown in Figure 2 according to eq 17. Decay times need to be scaled by 2.8 for comparison with experimental data to correct for viscosity effects of the TIP3P water model.

Residue	**Bond**[a]	**# decays**	$\mathbf{S^2_{fast}}$	$\boldsymbol{\rho_{fast}}$ **[ns]**	$\mathbf{S^2_{slow}}$	$\boldsymbol{\rho_{slow}}$ **[ns]**	$\mathbf{S^2}$	$\boldsymbol{\rho_{global}}$ **[ns]**
V4	CB-CG1	3	0.89	0.031	0.87	0.59	**0.60**	2.4
	CB-CG2	3	0.89	0.095	0.84	0.90	**0.50**	2.8
	average						**0.55**	2.6
E7	CB-CG	3	0.88	0.067	0.69	0.57	**0.34**	2.0
	CG-CD	3	0.88	0.024	0.76	0.37	**0.36**	2.4
	average						**0.35**	2.2
K13	CB-CG	2	-	-	0.89	0.29	**0.49**	2.2
	CG-CD	3	0.86	0.211	0.54	2.70	**0.12**	6.2
	CD-CE	3	0.82	0.057	0.62	0.33	**0.35**	2.4
	CE-NZ	3	0.77	0.037	0.40	0.29	**0.22**	2.6
	average						**0.30**	3.4
R40	CB-CG	3	0.93	0.025	0.88	0.35	**0.75**	2.5
	CG-CD	3	0.90	0.038	0.82	0.47	**0.61**	2.1
	CD-NE	3	0.87	0.124	0.65	1.67	**0.26**	2.8
	NE-CZ	3	0.85	0.028	0.59	0.37	**0.42**	1.8
	average						**0.51**	2.3
I41	CB-CG1	2	-	-	0.93	0.12	**0.90**	2.9
	CB-CG2	3	0.93	0.016	0.92	0.19	**0.89**	3.1
	average						**0.90**	3.0
F62	CB-CG	3	0.90	0.047	0.81	0.52	**0.63**	2.7
	CD1-CE1	3	0.89	0.022	0.81	0.38	**0.61**	2.6
	CD2-CE2	3	0.89	0.014	0.81	0.34	**0.62**	2.6
	average						**0.62**	2.6
V114	CB-CG1	2	-	-	0.86	0.51	**0.51**	1.9
	CB-CG2	2	-	-	0.84	0.54	**0.48**	2.0
	average						**0.50**	2.0
Y115	CB-CG	3	0.95	0.028	0.89	0.36	**0.77**	2.5
	CD1-CE1	3	0.93	0.027	0.87	0.36	**0.76**	2.5
	CZ-OH	3	0.92	0.027	0.85	0.36	**0.74**	2.5
	CB-CG1	3	0.89	0.031	0.87	0.59	**0.60**	2.4
	average						**0.72**	2.5

[a] The atom name notion corresponds to that used in PDB files.

Table 1C: NMR decay parameters for selected residues.

Residue	S^2_{fast}	ρ_{fast} [ns]	S^2_{slow}	ρ_{slow} [ns]	S^2	τ_e [ns]
E7	-	-	-	-	**0.91**	0.03
K13	0.91	-	0.90	0.85	**0.82**	
R40	0.88	-	0.87	0.62	**0.76**	
I41	0.52	0.04	0.72	2.22	**0.37**	
F62	0.91	-	0.94	1.18	**0.86**	
V114	0.78	0.03	0.75	2.50	**0.59**	
Y115	0.74	0.03	0.62	3.14	**0.46**	
G116	0.75	0.05	0.58	1.90	**0.43**	

Table 2: Summary of global rotational correlation times, ρ_{global} [ns] [a]

Method	**Temperature [°C]**							
	5	**15**	**20**	**25**	**26**	**27**	**30**	**35**
$r(t_c)$ [ns] [b]			9.2±0.4	8.0±0.4				
FCS [ns] [c]				$6.9^{+2.4}_{-1.3}$	$6.7^{+2.3}_{-1.3}$			
NMR [ns] [d]	16.9±0.2	12.3±0.2		9.6±0.1				7.5±0.1
NMR [ns] [e]	16.4±0.2	12.0±0.2		9.4±0.2				7.3±0.1
NMR [ns] [f]	14.5	10.6		8.3				6.4
MD [ns] [g]				8.32±0.59		7.95±0.56		
HYDROPRO [ns] [h]	14.43 ±0.41	10.03 ±0.29	8.95 ±0.25	7.82 ±0.22	7.61 ±0.22	7.42 ±0.21	6.88 ±0.20	6.11 ±0.17

[a] Gray filling indicates values extrapolated with the Stokes-Einstein law. The viscosity of the buffer was approximated by using data for water at different temperatures [108], scaled for the effect of dissolved salt (η = 0.9000 mPa·s for 248 mM NaCl:KCl (ratio 1:1) aqueous solutions at 25 °C [109]. [b] GABARAP F62C-BFL approx. 50 nM and 0.5 µM unlabeled GABARAP. The error was determined by repeated measurements. [c] The correction is outlined in Figure S2A. [d] Fit to an ellipsoid diffusion model (the parameters are compiled in Tab. S1). [e] Values extrapolated assuming NMR buffer with 100 % H_2O. [f] Values extrapolated to low concentration (scaling factor 7.5/8.5 as measured by TCSPC). [g] Mean and standard error from data in Figure 4B, scaled by a factor of 2.80 ± 0.04 to account for too low TIP3P-water viscosity (from [85], interpolated to 300 K). [h] Average value obtained by HydroPRO for the PDB IDs 1GNU, 3D32, 1KLV, 1KOT (for individual values see Tab. S4, here scaled for viscosity of the buffer by a factor 0.90/0.89).

Abbreviations

NMR, NOE, GABARAP, TCSPC, pFCS, MD,

References

1. Palmer, A.G., *NMR characterization of the dynamics of biomacromolecules.* Chem. Rev., 2004. **104**(8): p. 3623-40.
2. Cantor, C.R. and P.R. Schimmel, *Biophysical Chemistry: Part I: The Conformation of Biological Macromolecules*. 1980: W. H. Freeman.
3. Yang, D. and L.E. Kay, *Contributions to Conformational Entropy Arising from Bond Vector Fluctuations Measured from NMR-Derived Order Parameters: Application to Protein Folding.* Journal of Molecular Biology, 1996. **263**(2): p. 369-382.
4. Tzeng, S.-R. and C.G. Kalodimos, *Protein activity regulation by conformational entropy.* Nature, 2012. **488**(7410): p. 236-240.
5. Sharp, K.A., et al., *On the relationship between NMR-derived amide order parameters and protein backbone entropy changes.* Proteins, 2015. **83**(5): p. 922-30.
6. Neudecker, P., et al., *Solution Structure, Dynamics, and Hydrodynamics of the Calcium-bound Cross-reactive Birch Pollen Allergen Bet v 4 Reveal a Canonical Monomeric Two EF-Hand Assembly with a Regulatory Function.* Journal of Molecular Biology, 2004. **336**(5): p. 1141-1157.
7. Stahl, Y., et al., *Moderation of Arabidopsis Root Sternness by CLAVATA1 and ARABIDOPSIS CRINKLY4 Receptor Kinase Complexes.* Current Biology, 2013. **23**(5): p. 362-371.
8. Peng, T., et al., *Sequence-specific dynamics modulate recognition specificity in WW domains.* Nature structural & molecular biology, 2007. **14**(4): p. 325-31.
9. Bozoky, Z., et al., *Regulatory R region of the CFTR chloride channel is a dynamic integrator of phospho-dependent intra- and intermolecular interactions.* Proceedings of the National Academy of Sciences of the United States of America, 2013. **110**(47): p. E4427-E4436.
10. Mohrlüder, J., M. Schwarten, and D. Willbold, *Structure and potential function of γ-aminobutyrate type A receptor-associated protein.* FEBS Journal, 2009. **276**(18): p. 4989-5005.
11. Chen, W., M.W. van der Kamp, and V. Daggett, *Structural and dynamic properties of the human prion protein.* Biophysical Journal, 2014. **106**(5): p. 1152-63.
12. Sisamakis, E., et al., *Accurate Single-Molecule Fret Studies Using Multiparameter Fluorescence Detection.* Methods in Enzymology, Vol 475: Single Molecule Tools, Pt B, 2010. **475**: p. 455-514.
13. Karplus, M. and J. Kuriyan, *Molecular dynamics and protein function.* Proc Natl Acad Sci, U. S. A., 2005. **102**(0027-8424 (Print)): p. 6679-6685.
14. Shaw, D.E., et al., *Millisecond-scale molecular dynamics simulations on Anton*, in *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis (SC '09)*. 2009, ACM: Portland, Oregon. p. 1-11.
15. Clore, G.M., et al., *Analysis of the Backbone Dynamics of Interleukin-1-Beta Using 2-Dimensional Inverse Detected Heteronuclear N-15-H-1 Nmr-Spectroscopy.* Biochemistry, 1990. **29**(32): p. 7387-7401.

16. Trbovic, N., et al., *Structural analysis of protein dynamics by MD simulations and NMR spin-relaxation.* Proteins, 2008. **71**(2): p. 684-694.
17. Wang, H., et al., *GABAA-receptor-associated protein links GABAA receptors and the cytoskeleton.* Nature, 1999. **397**(6714): p. 69-72.
18. Wang, H. and R.W. Olsen, *Binding of the GABAA Receptor-Associated Protein (GABARAP) to Microtubules and Microfilaments Suggests Involvement of the Cytoskeleton in GABARAPGABAA Receptor Interaction.* Journal of Neurochemistry, 2000. **75**(2): p. 644-655.
19. Leil, T.A., et al., *GABAA Receptor-Associated Protein Traffics GABAA Receptors to the Plasma Membrane in Neurons.* The Journal of Neuroscience, 2004. **24**(50): p. 11429-11438.
20. Kittler, J.T., et al., *The Subcellular Distribution of GABARAP and Its Ability to Interact with NSF Suggest a Role for This Protein in the Intracellular Transport of GABAA Receptors.* Molecular and Cellular Neuroscience, 2001. **18**(1): p. 13-25.
21. Ma, P., et al., *Interaction of Bcl-2 with the Autophagy-related GABA(A) Receptor-associated Protein (GABARAP): BIOPHYSICAL CHARACTERIZATION AND FUNCTIONAL IMPLICATIONS.* The Journal of Biological Chemistry, 2013. **288**(52): p. 37204-37215.
22. Schwarten, M., et al., *Nix directly binds to GABARAP: A possible crosstalk between apoptosis and autophagy.* Autophagy, 2009. **5**(5): p. 690-698.
23. Schwarten, M., et al., *Solution structure of Atg8 reveals conformational polymorphism of the N-terminal domain.* Biochemical and Biophysical Research Communications, 2010. **395**(3): p. 426-431.
24. Ma, P., et al., *Conformational Polymorphism in Autophagy-Related Protein GATE-16.* Biochemistry, 2015.
25. Lipari, G. and A. Szabo, *Effect of librational motion on fluorescence depolarization and nuclear magnetic resonance relaxation in macromolecules and membranes.* Biophysical journal, 1980. **30**(3): p. 489-506.
26. Kinosita, K., Jr., S. Kawato, and A. Ikegami, *A Theory of Fluorescence Polarization Decay in Membranes.* Biophysical Journal, 1977. **20**: p. 289-305.
27. Levitt, M.H., *Spin dynamics: basics of nuclear magnetic resonance*. 2008: John Wiley & Sons.
28. Lipari, G. and A. Szabo, *Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. Theory and range of validity.* Journal of the American Chemical Society, 1982. **104**(17): p. 45464559.
29. Aragon, S. and R. Pecora, *Fluorescence correlation spectroscopy and Brownian rotational diffusion.* Biopolymers, 1975. **14**(1): p. 119-137.
30. Werbelow, L.G. and D.M. Grant, *Determination of motional asymmetry of methyl rotators from 13C spin dynamics.* Canadian Journal of Chemistry, 1977. **55**(9): p. 1558-1563.
31. Kask, P., et al., *Separation of the Rotational Contribution in Fluorescence Correlation Experiments.* Biophysical Journal, 1989. **55**(2): p. 213-220.
32. Merzbacher, E., *Quantum Mechanics*. 1998, Wiley, New York.
33. Clore, G.M., et al., *Deviations from the simple two-parameter model-free approach to the interpretation of nitrogen-15 nuclear magnetic relaxation of proteins.* Journal of the American Chemical Society, 1990. **112**(12): p. 4989-4991.
34. Woessner, D.E., *Nuclear Spin Relaxation in Ellipsoids Undergoing Rotational Brownian Motion.* The Journal of Chemical Physics, 1962. **37**(3): p. 647-654.

35. d'Auvergne , E.J. and P.R. Gooley, *Model-free model elimination: A new step in the model-free dynamic analysis of NMR relaxation data.* J. Biomol. NMR, 2006. **35(2)**: p. 117-135.

36. Cavanagh, J., et al., *Protein NMR Spectroscopy*. 2nd ed. 2007, Burlington: Elsevier Academic Press.

37. d'Auvergne , E.J. and P.R. Gooley, *The use of model selection in the model-free analysis of protein dynamics.* J. Biomol. NMR, 2003. **25(1)**: p. 25-39.

38. d'Auvergne , E.J. and P.R. Gooley, *Set theory formulation of the model-free problem and the diffusion seeded model-free paradigm.* J. Biomol. NMR, 2007. **3(7)**: p. 483-393.

39. d'Auvergne , E.J. and P.R. Gooley, *Optimisation of NMR dynamic models I. Minimisation algorithms and their performance within the model-free and Brownian rotational diffusion spaces.* J. Biomol. NMR, 2008. **40(2)**: p. 107-119.

40. d'Auvergne , E.J. and P.R. Gooley, *Optimisation of NMR dynamic models II. A new methodology for the dual optimisation of the model-free parameters and the Brownian rotational diffusion tensor.* J. Biomol. NMR, 2008. **40(2)**: p. 121-133.

41. Ehrenberg, M. and R. Rigler, *Rotational Brownian motion and fluorescence intensify fluctuations.* Chemical Physics, 1974. **4**(3): p. 390-401.

42. Shi, J.X., et al., *Nanosecond dynamics of the mouse acetylcholinesterase Cys(69)-Cys(96) omega loop.* Journal of Biological Chemistry, 2003. **278**(33): p. 30905-30911.

43. Alexiev, U., I. Rimke, and T. Pohlmann, *Elucidation of the nature of the conformational changes of the EF-interhelical loop in bacteriorhodopsin and of the helix VIII on the cytoplasmic surface of bovine rhodopsin: A time-resolved fluorescence depolarization study.* Journal of Molecular Biology, 2003. **328**(3): p. 705-719.

44. Lee, A.L., S.A. Kinnear, and A.J. Wand, *Redistribution and loss of side chain entropy upon formation of a calmodulin-peptide complex.* Nat Struct Mol Biol, 2000. **7**(1): p. 72-77.

45. Coyle, J.E., et al., *Structure of GABARAP in Two Conformations: Implications for GABA(A) Receptor Localization and Tubulin Binding.* Neuron, 2002. **33**(0896-6273 (Print)): p. 63-74.

46. Findeisen, M., T. Brand, and S. Berger, *A 1H-NMR thermometer suitable for cryoprobes.* Magn. Reson. Chem., 2007. **45**: p. 175-178.

47. Delaglio, F., et al., *NMRPipe: A multidimensional spectral processing system based on UNIX pipes.* Journal of Biomolecular NMR, 1995. **6**: p. 277-293.

48. Farrow, N.A., et al., *Backbone Dynamics of a Free and a Phosphopeptide-Complexed Src Homology 2 Domain Studied by 15N NMR Relaxation.* Biochemistry, 1994. **33**(19): p. 5984-6003.

49. Korzhnev, D.M., et al., *An NMR Experiment for the Accurate Measurement of Heteronuclear Spin-Lock Relaxation Rates.* Journal of the American Chemical Society, 2002. **124**(36): p. 10743-10753.

50. Orekhov, V.Y., I.V. Ibraghimov, and M. Billeter, *MUNIN: a new approach to multi-dimensional NMR spectra interpretation.* J Biomol NMR., 2001. **20**(1): p. 49-60.

51. Korzhnev, D.M., et al., *MUNIN: application of three-way decomposition to the analysis of heteronuclear NMR relaxation data.* J Biomol NMR, 2001. **21**(3): p. 263-268.

52. Johnson, B.A. and R.A. Blevins, *NMRView: A computer program for the visualization and analysis of NMR data.* J. Biomol. NMR, 1994. **4**: p. 603-614.

53. Pawley, N.H., et al., *An improved method for distinguishing between anisotropic tumbling and chemical exchange in analysis of ^{15}N relaxation parameters.* J. Biomol. NMR, 2001. **20**: p. 149-165.

54. Knight, D., et al., *The X-ray crystal structure and putative ligand-derived peptide binding properties of gamma-aminobutyric acid receptor type A receptor-associated protein.* J. Biol. Chem., 2002. **277**(7): p. 5556-61.

55. Schwieters, C.D., et al., *The Xplor-NIH NMR molecular structure determination package.* J. Magn. Reson., 2003. **160**: p. 65-73.

56. Brünger, A.T., *X-PLOR Version 3.1. A System for X-ray Crystallography and NMR*. 1992, New Haven: Yale University Press.

57. Dosset, P., et al., *Efficient analysis of macromolecular rotational diffusion from heteronuclear relaxation data.* Journal of Biomolecular NMR, 2000. **16**(1): p. 23-28.

58. Lakomek, N.-A., J. Ying, and A. Bax, *Measurement of 15N relaxation rates in perdeuterated proteins by TROSY-based methods.* Journal of Biomolecular NMR, 2012. **53**(3): p. 209-221.

59. Altieri, A.S., D.P. Hinton, and R.A. Byrd, *Association of Biomolecular Systems via Pulsed Field Gradient NMR Self-Diffusion Measurements.* Journal of the American Chemical Society, 1995. **117**(28): p. 7566-7567.

60. Wu, D.H., A.D. Chen, and C.S. Johnson, *An Improved Diffusion-Ordered Spectroscopy Experiment Incorporating Bipolar-Gradient Pulses.* Journal of Magnetic Resonance, Series A, 1995. **115**(2): p. 260-264.

61. Piotto, M., V. Saudek, and V. Sklenár, *Gradient-tailored excitation for single-quantum NMR spectroscopy of aqueous solutions.* J. Biomol. NMR, 1992. **2**: p. 661-665.

62. Stejskal, E.O. and J.E. Tanner, *Spin Diffusion Measurements: Spin Echoes in the Presence of a Time-Dependent Field Gradient.* J. Chem. Phys., 1965. **42**: p. 288-292.

63. Cho, C.H., et al., *Thermal Offset Viscosities of Liquid H2O, D2O, and T2O.* Journal of Physical Chemistry B, 1999. **103**(11): p. 1991-1994.

64. Wilkins, D.K., et al., *Hydrodynamic radii of native and denatured proteins measured by pulse field gradient NMR techniques.* Biochemistry, 1999. **38**(0006-2960 (Print)): p. 16424-31.

65. Sindbert, S., et al., *Accurate Distance Determination of Nucleic Acids via Forster Resonance Energy Transfer: Implications of Dye Linker Length and Rigidity.* Journal of the American Chemical Society, 2011. **133**(8): p. 2463-2480.

66. Karolin, J., et al., *Fluorescence and Absorption Spectroscopic Properties of Dipyrrometheneboron Difluoride (BODIPY) Derivatives in Liquids, Lipid Membranes, and Proteins.* Journal of the American Chemical Society, 1994. **116**(17): p. 7801-7806.

67. Felekyan, S., et al., *Full correlation from picoseconds to seconds by time-resolved and time-correlated single photon detection.* Review of Scientific Instruments, 2005. **76**(8).

68. Ortega, A., D. Amorós, and J. García de la Torre, *Prediction of Hydrodynamic and Other Solution Properties of Rigid Proteins from Atomic- and Residue-Level Models.* Biophysical Journal, 2011. **101**(4): p. 892-898.

69. Stangler, T., L.M. Mayr, and D. Willbold, *Solution Structure of Human GABAA Receptor-associated Protein GABARAP: Implications for Biological Function and its Regulation.* Journal of Biological Chemistry, 2002. **277**: p. 13363-13366.

70. Kuono, T., et al., *1H, 13C and 15N resonance assignments of GABARAP, GABAA receptor associated protein.* Journal of Biomolecular NMR, 2002. **22**(1): p. 97-98.

71. Weiergräber, O.H., et al., *Ligand Binding Mode of GABAA Receptor-Associated Protein.* Journal of Molecular Biology, 2008. **381**(5): p. 1320-1331.

72. Sander, P., et al., *GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit.* Bioinformatics (Oxford, England), 2013. **29**(7): p. 845-854.

73. Kresten, L.-L., et al., *Improved side-chain torsion potentials for the Amber ff99SB protein force field.* Proteins, 2010. **78**(8): p. 1950-1958.

74. Jorgensen, W.L., et al., *Comparison of simple potential functions for simulating liquid water.* The Journal of chemical physics, 1983. **79**(2): p. 926-935.

75. Hoover, W.G., *Canonical dynamics: equilibrium phase-space distributions.* Physical Review A, 1985.

76. Darden, T., et al., *New tricks for modelers from the crystallography toolkit: the particle mesh Ewald algorithm and its use in nucleic acid simulations.* Structure (London, England : 1993), 1999. **7**(3): p. 60.

77. Hess, B., *P-LINCS: A parallel linear constraint solver for molecular simulation.* Journal of Chemical Theory and Computation, 2008.

78. Maragakis, P., et al., *Microsecond molecular dynamics simulation shows effect of slow loop dynamics on backbone amide order parameters of proteins.* The Journal of Physical Chemistry B, 2008. **112**(19): p. 6155-6158.

79. Burnham, K.P. and D.R. Anderson, *Multimodel inference understanding AIC and BIC in model selection.* Sociological methods & research, 2004. **33**(2): p. 261-304.

80. Kühnemuth, R. and C.A.M. Seidel, *Principles of Single Molecule Multiparameter Fluorescence Spectroscopy.* Single Molecules, 2001. **2**(4): p. 251-254.

81. Borst, J.W., et al., *Structural Changes of Yellow Cameleon Domains Observed by Quantitative FRET Analysis and Polarized Fluorescence Correlation Spectroscopy.* Biophysical Journal, 2008. **95**(11): p. 5399-5411.

82. Dimura, M., et al., *Quantitative FRET studies and integrative modeling unravel the structure and dynamics of biomolecular systems.* Current Opinion in Structural Biology, 2016. **40**: p. 163-185.

83. Bergstrom, F., et al., *Dimers of dipyrrometheneboron difluoride (BODIPY) with light spectroscopic applications in chemistry and biology.* Journal of the American Chemical Society, 2002. **124**(2): p. 196-204.

84. Kinosita, K., A. Ikegami, and S. Kawato, *On the wobbling-in-cone analysis of fluorescence anisotropy decay.* Biophysical Journal, 1982. **37**(2): p. 461-464.

85. Mao, Y. and Y. Zhang, *Thermal conductivity, shear viscosity and specific heat of rigid water models.* Chemical Physics Letters, 2012. **542**: p. 37-41.

86. Damberg, P., et al., *C-13-H-1 NMR relaxation and fluorescence anisotropy decay study of tyrosine dynamics in motilin.* Biophysical Journal, 2002. **83**(5): p. 2812-2825.

87. Roosen-Runge, F., et al., *Protein self-diffusion in crowded solutions.* Proc Natl Acad Sci U S A, 2011. **108**(29): p. 11815-20.

88. Bernado, P., J. Garcia de la Torre, and M. Pons, *Macromolecular crowding in biological systems: hydrodynamics and NMR methods.* J Mol Recognit, 2004. **17**(5): p. 397-407.

89. Nymann-Andersen, J., H. Wang, and R.W. Olsen, *Biochemical identification of the binding domain in the GABAA receptor-associated protein (GABARAP) mediating dimer formation.* Neuropharmacology, 2002. **43**(4): p. 476-481.

90. Pacheco, V., et al., *Assessment of GABARAP self-association by its diffusion properties.* Journal of Biomolecular NMR, 2010. **48**(1): p. 49-58.
91. Weiergräber, O.H., D. Willbold, and J. Mohrlüder, *Atg8 Family Proteins--Autophagy and Beyond*. 2013: INTECH Open Access Publisher.
92. Kaufmann, A., et al., *Molecular Mechanism of Autophagic Membrane-Scaffold Assembly and Disassembly.* Cell, 2014. **156**(3): p. 469-481.
93. Ma, P., et al., *Preparation of a Functional GABARAP–Lipid Conjugate in Nanodiscs and its Investigation by Solution NMR Spectroscopy.* ChemBioChem, 2010. **11**(14): p. 1967-1970.
94. Wong, V. and D.A. Case, *Evaluating Rotational Diffusion from Protein MD Simulations.* The Journal of Physical Chemistry B, 2008. **112**(19): p. 6013-6024.
95. Mukherjee, A. and B. Bagchi, *Solvent frictional forces in the rotational diffusion of proteins in water.* Current Science, 2006. **91**(9): p. 1208-1216.
96. Vitalini, F., et al., *Dynamic properties of force fields.* The Journal of Chemical Physics, 2015. **142**(8): p. 084101.
97. Kalinin, S., et al., *A toolkit and benchmark study for FRET-restrained high-precision structural modeling.* Nature Methods, 2012. **9**(12): p. 1218-1227.
98. Muschielok, A., et al., *A nano-positioning system for macromolecular structural analysis.* Nature Methods, 2008. **5**(11): p. 965-971.
99. Beckers, M., et al., *Quantitative structural information from single-molecule FRET.* Faraday Discussions, 2015. **184**: p. 117-129.
100. Wriggers, W. and K. Schulten, *Protein Domain Movements: Detection of Rigid Domains and Visualization of Hinges in Comparisons of Atomic Coordinates.* Proteins: Structure, Function, and Bioinformatics, 1997. **29**(1): p. 1-14.
101. Hayward, S., *Structural Principles Governing Domain Motions in Proteins.* Proteins: Structure, Function, and Bioinformatics, 1999. **36**(4): p. 425-435.
102. Kumeta, H., et al., *The NMR structure of the autophagy-related protein Atg8.* Journal of Biomolecular NMR, 2010. **47**(3): p. 237-241.
103. Freedberg, D.I. and P. Selenko, *Live Cell NMR*, in *Annual Review of Biophysics, Vol 43*, K.A. Dill, Editor. 2014. p. 171-192.
104. Dror, R.O., et al., *Biomolecular Simulation: A Computational Microscope for Molecular Biology.* Annual Review of Biophysics, Vol 41, 2012. **41**: p. 429-452.
105. Sakon, J.J. and K.R. Weninger, *Detecting the conformation of individual proteins in live cells.* Nature Methods, 2010. **7**(3): p. 203-U56.
106. Konig, I., et al., *Single-molecule spectroscopy of protein conformational dynamics in live eukaryotic cells.* Nature Methods, 2015. **12**(8): p. 773-U129.
107. Yu, I., et al., *Biomolecular interactions modulate macromolecular structure and dynamics in atomistic model of a bacterial cytoplasm.* eLife, 2016. **5**: p. e19274.
108. Huber, M.L., et al., *New International Formulation for the Viscosity of H2O.* Journal of Physical and Chemical Reference Data, 2009. **38**(2): p. 101-125.
109. Zhang, H.L. and S.J. Han, *Viscosity and density of water plus sodium chloride plus potassium chloride solutions at 298.15 K.* Journal of Chemical and Engineering Data, 1996. **41**(3): p. 516-520.

Chapter 4

Essential Achievements

The research presented in this dissertation exemplifies the benefit of MD simulations for the interpretation of experimental data obtained with diverse methods. It is demonstrated, how computational and experimental methods can be combined to build models of macromolecules, complement each other in the description of their local and global dynamics, and enable the generation of functional models and assembly mechanisms of proteins.

In particular, several key achievements provide new insight in the application potential of molecular dynamics and other computational biochemistry methods. With the aid of computational biochemistry, in particular MD simulations and SAXS curve reconstruction, we were able to generate an assembly model for a large fusion protein dimer from sub-domains of known structure that is in agreement with available low resolution experimental data. MD simulations and SAXS data have been proven to valuably complement each other in other projects as well [100].

In the GATE-16 project, multiple X-ray and solution NMR structures could only indicate a general flexibility of the C-terminus. However, due to crystal packing artifacts it was not possible to asses the conformational polymorphism of GATE-16's C-terminus by these methods alone. Substantial sampling in the MD simulations was needed to identify all possible C-terminal states. Finally, the MD data was crucial for the generation of a functional model of the C-terminal dynamics.

With the aid of MD simulations it was possible to find evidence for a theory of hitherto unexplained biochemical assembly mechanisms of the IL-6 receptor complex and provide detailed structural and dynamical evidence.

Finally, the potential of the combination of multiple experimental and computational models to assess protein dynamics was demonstrated through the description of the dynamics of GABARAP. The combined theoretical framework and approach can be used by many other researchers in the future. Moreover, my analysis tools for the calculation of S^2 order parameters for both side chains and N–H bond vectors have been made available to the public in the software **MOP$\boldsymbol{S^2}$** [43].

While MD simulations have become a standard tool to study protein dynamics, I demonstrated how important modern sampling methods are for the correct assessment of all conformational states of a protein. The crucial step in the application of MD simulations is not only the simulation itself but also the data analysis, which is often complicated and not straight forward. Knowledge of the protein under study, in particular of its functional mechanisms, is very helpful to extract the relevant information from MD trajectories, which finally enables the explanation of functional mechanisms from protein structures and their dynamical behaviour, a demonstrated in this work.

Chapter 5

Outlook

At present, all-atom MD simulations are ubiquitously employed to study various types of molecular systems: from mixtures of small molecules, simulating their properties under different conditions, to biological macromolecules like proteins, DNA and RNA in solution and at membrane interfaces. The two major limitations MD simulations face are the upper bound on sampling due to the high computational demand of simulations with up to several million atoms and the deficiencies of modern force fields.

The size of simulated systems has grown over the past 40 years from a few 100 atoms simulated for less than 10 ps [101], finally enabling the study of systems comprising more than 2 million atoms for multiple microseconds, such as the translocation of RNA in the ribosome [102] or even 100 million atoms for short periods of approximately 100 ns to study molecular crowding effects in the cytosol [103]. Apart from the ability to study larger systems, it is also crucial to simulate systems for ever longer periods, as many important biological phenomena happen on a timescale of microseconds and beyond. Furthermore, sufficient sampling is also important for the collection of sufficient statistical data to obtain converged results that do not change with extended simulation time. To this end, enhanced sampling methods are under development for which it was statistically proven that convergence is achievable [104]. The metadynamics approach speeds up sampling through a biasing potential applied to carefully selected collective variables or reaction coordinates, penalizing the system when visiting already sufficiently sampled states [105]. Methods are developed to guide the difficult and potentially risky definition of collective variables [106]. Replica exchange methods, as discussed at length in this thesis (section 2.2.3), increase sampling efficiencies through replicated systems, which are heated or whose Hamiltonians are changed in order to overcome energy barriers more easily. Markov State Models (MSMs) [107] are a method to characterize all conformational states of a molecule and the transition rates between them. Their main advantage is that MSMs can be constructed from many short simulations, thereby using the power of distributed computing. The independent simulations can be preferably started from regions of the conformational space that have not been sufficiently sampled. Statistical methods can be used to decide which states still lack sampling for reliable results and to judge convergence of the simulations [104].

The second major problem of MD simulations is the quality of the employed force fields for biomolecules. The AMBER and CHARMM force field families have seen continuously updated parameter sets. While these yield increasingly better results for folded proteins, yet they often bias secondary structure elements [108, 109]. Intrinsically disordered proteins are difficult to study with these force fields, as they overstabilize compact states [110]. Especially for protein folding simulations it is highly desirable to use well balanced force fields, that predict the properties of the unfolded and folded states equally accurate, in order to reliably predict intermediate states of the folding process. Additional limitations are inherent to the classical force fields used. Electrostatic interactions are only accounted for by fixed partial charges on each atom. However, the electron distributions of real atoms are polarizable and influenced by the local electrostatic environment. This effect is accounted for in special polarizable force fields, like AMOEBA [111]. However, the improved physical description comes at the cost of higher computational complexity and cost and hence polarizable force fields are not used much at the moment. Finally, different protonation states, especially of histidine residues, are only accounted for implicitly through multiple simulations of different fixed protonation states, or often not at all. Force fields with variable protonation states to enable simulations at constant pH have been suggested [112], but are not in widespread use at present.

The investigation and prediction of protein function from their structure and dynamics is an active field or research. While individual cases have been studied in great detail, as exemplified in this thesis, a general relationship between protein structure, dynamics and function is yet to be discovered. Interesting work on the dynasome, the collective space of dynamic behaviour of proteins, shows correlations between dynamical properties of proteins and their function [113, 114]. However, no simple interpretations of this relationship have been proposed.

The holy grail of molecular simulation remains in far reach: the computational descriptions of a whole cell at atomistic level. While unquestionably a goal only achievable in a distant future, a virtual cell would enable the investigation of cellular processes in their entirety.

Bibliography

[1] Jan Drenth. *Principles of protein X-ray crystallography.* Springer Science & Business Media, 2007.

[2] Wikimedia Commons, Hydrargyrum. *Bragg diffraction from a cubic crystal lattice.* Creative Commons Attribution-Share Alike 3.0 Unported license. 2001. URL: https://en.wikipedia.org/wiki/File:Bragg_diffraction_2.svg (visited on 04/05/2017).

[3] Wayne A Hendrickson. "Analysis of protein structure from diffraction measurement at multiple wavelengths". In: *Trans. Am. Crystallogr. Assoc* 21.11 (1985).

[4] Wayne A Hendrickson and Craig M Ogata. "Phase determination from multiwavelength anomalous diffraction measurements". In: *Methods in Enzymology* 276 (1997), pp. 494–523.

[5] Eric de La Fortelle and Gérard Bricogne. "Maximum-likelihood heavy-atom parameter refinement for multiple isomorphous replacement and multiwavelength anomalous diffraction methods". In: *Methods in enzymology* 276 (1997), pp. 472–494.

[6] Wikimedia Commons, Jeff Dahl. *X-ray diffraction pattern of crystallized 3Clpro, a SARS protease. (2.1 Angstrom resolution).* This file is licensed under the Creative Commons Attribution-Share Alike 3.0 Unported license. 2006. URL: https://en.wikipedia.org/wiki/File:Bragg_diffraction_2.svg (visited on 04/05/2017).

[7] Naishadh Shah et al. "Magnetic resonance spectroscopy as an imaging tool for cancer: a review of the literature". In: *The Journal of the American Osteopathic Association* 106.1 (2006), pp. 23–27.

[8] Giovanni Lipari and Attila Szabo. "Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. Theory and range of validity". In: *J. Am. Chem. Soc* 104.17 (1982), pp. 4546–4559.

[9] Nikola Trbovic et al. “Structural analysis of protein dynamics by MD simulations and NMR spin-relaxation”. In: *Proteins: Structure, Function, and Bioinformatics* 71.2 (2008), pp. 684–694.

[10] Po-chia Chen and Jochen S Hub. “Validating solution ensembles from molecular dynamics simulation by wide-angle X-ray scattering data”. In: *Biophysical journal* 107.2 (2014), pp. 435–447.

[11] Maxim V Petoukhov and Dmitri I Svergun. “Global rigid body modeling of macromolecular complexes against small-angle scattering data”. In: *Biophysical journal* 89.2 (2005), pp. 1237–1250.

[12] Dmitri I Svergun and Michel HJ Koch. “Small-angle scattering studies of biological macromolecules in solution”. In: *Reports on Progress in Physics* 66.10 (2003), p. 1735.

[13] Christopher D Putnam et al. “X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution”. In: *Quarterly reviews of biophysics* 40.03 (2007), pp. 191–285.

[14] Wikipedia. *X-ray solution scattering curves.* This file is licensed under the Creative Commons Attribution-ShareAlike license version 3.0. 2007. URL: `https://en.wikipedia.org/wiki/File:Sax_curve.png` (visited on 05/15/2017).

[15] D Svergun, Claudio Barberato, and Michel HJ Koch. “CRYSOL–a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates”. In: *Journal of applied crystallography* 28.6 (1995), pp. 768–773.

[16] K_ Kinosita, Suguru Kawato, and Akira Ikegami. “A theory of fluorescence polarization decay in membranes”. In: *Biophysical journal* 20.3 (1977), pp. 289–305.

[17] S. A. Adcock and J. A. McCammon. “Molecular Dynamics: Survey of Methods for Simulating the Activity of Proteins”. In: *Chem. Rev.* 106.5 (2006), 1589–1615. ISSN: 0009-2665. DOI: `10.1021/cr040426m`.

[18] S. Pronk et al. “GROMACS 4.5: A High-Throughput and Highly Parallel Open Source Molecular Simulation Toolkit”. In: *Bioinformatics* 29.7 (2013), 845–854. ISSN: 1367-4803. DOI: `10.1093/bioinformatics/btt055`.

[19] J. C. Butcher. *Numerical Methods for Ordinary Differential Equations.* 2nd ed. ISBN 978-0470723357. Wiley, 2008.

[20] D. Frenkel and B. Smit. *Understanding Molecular Simulation – From Algorithms to Applications*. 2nd ed. Vol. 1. Computational Science Series. ISBN 0-12-267351-4. 84 Theobald's Road, London: Academic Press, 2002.

[21] L. Verlet. "Computer Experiments on Classical Fluids .I. Thermodynamical Properties of Lennard-Jones Molecules". In: *Phys. Rev.* 159.1 (1967), 98–&. ISSN: 0031-899X. DOI: 10.1103/PhysRev.159.98.

[22] P. Atkins and J. de Paula. *Physical Chemistry*. 9th ed. ISBN 978-1429218122. W. H. Freeman, 2009.

[23] H. J. C. Berendsen. "Transport-Properties Computed by Linear Response Through Weak-Coupling to a Bath". In: *Computer Simulation in Materials Science: Interatomc Potentials, Simulation Techniques and Applications*. Ed. by Meyer, M. and Pontikis, V. Vol. 205. NATO Advanced Science Institutes Series, Series E, Applied Sciences. 1991, 139–155. ISBN: 0-7923-1455-7.

[24] S. Nosé. "A Unified Formulation Of The Constant Temperature Molecular-Dynamics Mehtod". In: *J. Chem. Phys.* 81.1 (1984), 511–519. ISSN: 0021-9606.

[25] W. G. Hoover. "Canonical Dynamics - Equilibrium Phase-space Distributions". In: *Phys. Rev.* 31.3 (1985), 1695–1697. ISSN: 1050-2947. DOI: 10.1103/PhysRevA.31.1695.

[26] G. Bussi, D. Donadio, and M. Parrinello. "Canonical Sampling Through Velocity Rescaling". In: *J. Chem. Phys.* 126.1 (2007). ISSN: 0021-9606. DOI: 10.1063/1.2408420.

[27] M. Parrinello and A. Rahman. "Polymorphic Transitions In Single-Crystals - A New Molecular-Dynamics Method". In: *J. Appl. Phys.* 52.12 (1981), 7182–7190. ISSN: 0021-8979. DOI: 10.1063/1.328693.

[28] G. J. Martyna et al. "Explicit Reversible Integrators for Extended Systems Dynamics". In: *Mol. Phys.* 87.5 (1996), 1117–1157. ISSN: 0026-8976. DOI: 10.1080/00268979600100761.

[29] T. Darden, D. York, and L. Pedersen. "Particle Mesh Ewald - An N.log(n) Method for Ewald Sums in Large Systems". In: *J. Chem. Phys.* 98.12 (1993), 10089–10092. ISSN: 0021-9606. DOI: 10.1063/1.464397.

[30] K. A. Beauchamp et al. "Are Protein Force Fields Getting Better? A Systematic Benchmark on 524 Diverse NMR Measurements". In: *J. Chem. Theory Comput.* 8.4 (2012), 1409–1414. ISSN: 1549-9618. DOI: 10.1021/ct2007814.

[31] D. W. Li and R. Brueschweiler. “NMR-Based Protein Potentials”. In: *Angew. Chem. Int. Edit.* 49.38 (2010), 6778–6780. ISSN: 1433-7851. DOI: 10.1002/anie.201001898.

[32] David J Earl and Michael W Deem. “Parallel tempering: Theory, applications, and new perspectives”. In: *Physical Chemistry Chemical Physics* 7.23 (2005), pp. 3910–3916.

[33] Nitin Rathore, Manan Chopra, and Juan J de Pablo. “Optimal allocation of replicas in parallel tempering simulations”. In: *The Journal of chemical physics* 122.2 (2005), p. 024111.

[34] Lingle Wang, Richard A Friesner, and BJ Berne. “Replica exchange with solute scaling: a more efficient version of replica exchange with solute tempering (REST2)”. In: *The journal of physical chemistry. B* 115.30 (2011), p. 9431.

[35] Giovanni Bussi. “Hamiltonian replica exchange in GROMACS: a flexible implementation”. In: *Molecular Physics* 112.3-4 (2014), pp. 379–384.

[36] Gareth A Tribello et al. “PLUMED 2: New feathers for an old bird”. In: *Computer Physics Communications* 185.2 (2014), pp. 604–613.

[37] Wolfgang Kabsch. “A solution for the best rotation to relate two sets of vectors”. In: *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography* 32.5 (1976), pp. 922–923.

[38] Pu Liu, Dimitris K Agrafiotis, and Douglas L Theobald. “Fast determination of the optimal rotational matrix for macromolecular superpositions”. In: *Journal of computational chemistry* 31.7 (2010), pp. 1561–1563.

[39] Alexandros Altis et al. “Dihedral angle principal component analysis of molecular dynamics simulations”. In: *The Journal of chemical physics* 126.24 (2007), p. 244111.

[40] Indira Chandrasekhar et al. “A 500 ps molecular dynamics simulation study of interleukin-1β in water: correlation with nuclear magnetic resonance spectroscopy and crystallography”. In: *Journal of molecular biology* 226.1 (1992), pp. 239–250.

[41] Kenneth P Burnham and David R Anderson. “Multimodel inference understanding AIC and BIC in model selection”. In: *Sociological methods & research* 33.2 (2004), pp. 261–304.

[42] David A Case. “Calculations of NMR dipolar coupling strengths in model peptides”. In: *Journal of biomolecular NMR* 15.2 (1999), pp. 95–102.

[43] Oliver Schillinger. ***MOPS*^{2}** *– Molecular Order Parameters S^2; Compute bond vector S^2 order parameters from MD trajectories.* 2017. URL: https://github.com/schilli/MOPS (visited on 05/22/2017).

[44] David S Wishart. "Interpreting protein chemical shift data". In: *Progress in nuclear magnetic resonance spectroscopy* 58.1 (2011), pp. 62–87.

[45] Shen Yang and Ad Bax. "SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network". In: *Journal of Biomolecular NMR* 48.1 (2010), pp. 13–22.

[46] Beomsoo Han et al. "SHIFTX2: significantly improved protein chemical shift prediction". In: *Journal of biomolecular NMR* 50.1 (2011), p. 43.

[47] Anna Katharina Dehof et al. "NightShift: NMR shift inference by general hybrid model training-a framework for NMR chemical shift prediction". In: *BMC bioinformatics* 14.1 (2013), p. 98.

[48] Jakob T Nielsen, Hamid R Eghbalnia, and Niels Chr Nielsen. "Chemical shift prediction for protein structure calculation and quality assessment using an optimally parameterized force field". In: *Progress in nuclear magnetic resonance spectroscopy* 60 (2012), pp. 1–28.

[49] V. Dartois et al. "Cloning, Nucleotide-Sequence and Expression in Escherichia-Coli of a Lipase Gene From Bacillus-Subtilis". In: *Biochim. Biophys. Acta* 1131.3 (1992), 253–260. ISSN: 0006-3002. DOI: 10.1016/0167-4781(92)90023-S.

[50] H. Tjalsma et al. "Signal Peptide-Dependent Protein Transport in Bacillus Subtilis: a Genome-Based Survey of the Secretome". In: *Microbiol. Mol. Biol. R.* 64.3 (2000), 515+. ISSN: 1092-2172. DOI: 10.1128/MMBR.64.3.515-547.2000.

[51] C. R. Harwood. "Bacillus-Subtilis and its Relatives - Molecular Biological and Industrial Workhorses". In: *Trends Biotechnol.* 10.7 (1992), 247–256. ISSN: 0167-7799. DOI: 10.1016/0167-7799(92)90233-L.

[52] G. van Pouderoyen et al. "The Crystal Structure of Bacillus Subtilis Lipase: A Minimal α/β Hydrolase Fold Enzyme". In: *J. Mol. Biol.* 309.1 (2001), 215–226.

[53] K. E. Jaeger and M. T. Reetz. "Microbial Lipases form Versatile Tools for Biotechnology". In: *Trends Biotechnol.* 16.9 (1998), 396–403. ISSN: 0167-7799. DOI: 10.1016/S0167-7799(98)01195-0.

[54] M. Nardini and B. W. Dijkstra. "α/β Hydrolase Fold Enzymes: the Family Keeps Growing". In: *Curr. Opin. Struc. Biol.* 9.6 (1999), 732–737. ISSN: 0959-440X. DOI: 10.1016/S0959-440X(99)00037-8.

[55] J. Uppenberg et al. “Sequence, Crystal-Structure Determination And Refinement Of 2 Crystal Forms of Lipase-B from Canadia-Antarctica”. In: *Structure* 2.4 (1994), 293–308. ISSN: 0969-2126. DOI: 10.1016/S0969-2126(00)00031-9.

[56] M. Bocola et al. “Learning from Directed Evolution: Theoretical Investigations into Cooperative Mutations in Lipase Enantioselectivity”. In: *Chembiochem.* 5.2 (2004), 214–223.

[57] S. Kaspar et al. “The periplasmic Domain of the Histidine Autokinase CitA Functions as a Highly Specific Citrate Receptor”. In: *Mol. Microbiol.* 33.4 (1999), 858–872. ISSN: 0950-382X. DOI: 10.1046/j.1365-2958.1999.01536.x.

[58] J. B. Stock et al. “Two-Component Signal Transduction Systems: Structure-Function Relationships and Mechanisms of Catalysis”. In: *Two-Component Signal Transduction.* Ed. by J. A. Hoch and T. J. Silhavy. American Society for Microbiology Press, 1995, 25–51.

[59] S. Reinelt et al. “The Structure of the Periplasmic Ligand-Binding Domain of the Sensor Kinase CitA Reveals the First Extracellular PAS Domain”. In: *J. Biol. Chem.* 278.40 (2003), 39189–39196. ISSN: 0021-9258. DOI: 10.1074/jbc.M305864200.

[60] M. Sevvana et al. “A Ligand-Induced Switch in the Periplasmic Domain of Sensor Histidine Kinase CitA”. In: *J. Mol. Biol.* 377.2 (2008), 512–523.

[61] Ml Bott, M. Meyer, and P. Dimroth. “Regulation of Anaerobic Citrate Metabolism in Klebsiella-Pneumoniae”. In: *Mol. Microbiol.* 18.3 (1995), 533–546. ISSN: 0950-382X. DOI: 10.1111/j.1365-2958.1995.mmi_18030533.x.

[62] Joshua D Rabinowitz and Eileen White. “Autophagy and metabolism”. In: *Science* 330.6009 (2010), pp. 1344–1348.

[63] Daniel J Klionsky. “Autophagy revisited: a conversation with Christian de Duve”. In: *Autophagy* 4.6 (2008), pp. 740–743.

[64] Noboru Mizushima and Masaaki Komatsu. “Autophagy: renovation of cells and tissues”. In: *Cell* 147.4 (2011), pp. 728–741.

[65] Noboru Mizushima, Tamotsu Yoshimori, and Yoshinori Ohsumi. “The role of Atg proteins in autophagosome formation”. In: *Annual review of cell and developmental biology* 27 (2011), pp. 107–132.

[66] Carmine Settembre et al. “Signals from the lysosome: a control centre for cellular clearance and energy metabolism”. In: *Nature reviews Molecular cell biology* 14.5 (2013), pp. 283–296.

[67] Hilla Weidberg et al. “LC3 and GATE-16/GABARAP subfamilies are both essential yet act differently in autophagosome biogenesis”. In: *The EMBO journal* 29.11 (2010), pp. 1792–1802.

[68] Yuval Sagiv et al. “GATE-16, a membrane transport modulator, interacts with NSF and the Golgi v-SNARE GOS-28”. In: *The EMBO Journal* 19.7 (2000), pp. 1494–1504.

[69] David Knight et al. “The X-ray crystal structure and putative ligand-derived peptide binding properties of γ-aminobutyric acid receptor type A receptor-associated protein”. In: *Journal of Biological Chemistry* 277.7 (2002), pp. 5556–5561.

[70] Yurong Xin et al. “Cloning, expression patterns, and chromosome localization of three human and two mouse homologues of GABA A receptor-associated protein”. In: *Genomics* 74.3 (2001), pp. 408–413.

[71] Kenji Sugawara et al. “The crystal structure of microtubule-associated protein light chain 3, a mammalian homologue of Saccharomyces cerevisiae Atg8”. In: *Genes to Cells* 9.7 (2004), pp. 611–618.

[72] Isei Tanida, Takashi Ueno, and Eiki Kominami. “LC3 conjugation system in mammalian autophagy”. In: *The international journal of biochemistry & cell biology* 36.12 (2004), pp. 2503–2518.

[73] Endalkachew Ashenafi Alemu et al. “ATG8 family proteins act as scaffolds for assembly of the ULK complex sequence requirements for LC3-interacting region (LIR) motifs”. In: *Journal of Biological Chemistry* 287.47 (2012), pp. 39275–39290.

[74] Joseph E Coyle et al. “Structure of GABARAP in two conformations: implications for GABAA receptor localization and tubulin binding”. In: *Neuron* 33.1 (2002), pp. 63–74.

[75] Aster Legesse-Miller et al. “Isolation and characterization of a novel low molecular weight protein involved in intra-Golgi traffic”. In: *Journal of Biological Chemistry* 273.5 (1998), pp. 3105–3109.

[76] Marco BE Schaaf et al. “LC3/GABARAP family proteins: autophagy-(un) related functions”. In: *The FASEB Journal* 30.12 (2016), pp. 3961–3978.

[77] A Mackiewicz et al. “Complex of soluble human IL-6-receptor/IL-6 upregulates expression of acute-phase proteins.” In: *The Journal of Immunology* 149.6 (1992), pp. 2021–2027. ISSN: 2022-1767. eprint: http://www.jimmunol.org/content/149/6/2021.full.pdf. URL: http://www.jimmunol.org/content/149/6/2021.

[78] Björn Rabe et al. “Transgenic blockade of interleukin 6 transsignaling abrogates inflammation”. In: *Blood* 111.3 (2008), pp. 1021–1028.

[79] Peter C Heinrich et al. “Principles of interleukin (IL)-6-type cytokine signalling and its regulation”. In: *Biochemical journal* 374.1 (2003), pp. 1–20.

[80] Jürgen Scheller, Joachim Grötzinger, and Stefan Rose-John. “Updating interleukin-6 classic-and trans-signaling”. In: *Signal Transduction* 6.4 (2006), pp. 240–259.

[81] Stephanie Tenhumberg et al. “gp130 dimerization in the absence of ligand: preformed cytokine receptor complexes”. In: *Biochemical and biophysical research communications* 346.3 (2006), pp. 649–657.

[82] Björn Schuster et al. “The human interleukin-6 (IL-6) receptor exists as a preformed dimer in the plasma membrane”. In: *FEBS letters* 538.1-3 (2003), pp. 113–116.

[83] Jürgen Mülberg et al. “The soluble interleukin-6 receptor is generated by shedding”. In: *European journal of immunology* 23.2 (1993), pp. 473–480.

[84] Stefan Rose-John et al. “The IL-6/sIL-6R complex as a novel target for therapeutic approaches”. In: *Expert opinion on therapeutic targets* 11.5 (2007), pp. 613–624.

[85] Athena Chalaris et al. “Apoptosis is a natural stimulus of IL6R shedding and contributes to the proinflammatory trans-signaling function of neutrophils”. In: *Blood* 110.6 (2007), pp. 1748–1755.

[86] Christoph Becker et al. “TGF-β suppresses tumor progression in colon cancer by inhibition of IL-6 trans-signaling”. In: *Immunity* 21.4 (2004), pp. 491–501.

[87] Guang-Yi Xu et al. “Solution structure of recombinant human interleukin-6”. In: *Journal of molecular biology* 268.2 (1997), pp. 468–481.

[88] William Somers, Mark Stahl, and Jasbir S Seehra. “1.9 Å crystal structure of interleukin 6: implications for a novel mode of receptor dimerization and signaling”. In: *The EMBO journal* 16.5 (1997), pp. 989–997.

[89] Martin J Boulanger et al. “Hexameric structure and assembly of the interleukin-6/IL-6 α-receptor/gp130 complex”. In: *Science* 300.5628 (2003), pp. 2101–2104.

[90] Joachim Grötzinger et al. “IL-6 type cytokine receptor complexes: hexamer, tetramer or both?” In: *Biological chemistry* 380.7-8 (1999), pp. 803–813.

[91] Jürgen Müllberg et al. "IL-6 receptor independent stimulation of human gp130 by viral IL-6". In: *The Journal of Immunology* 164.9 (2000), pp. 4672–4677.

[92] Dar-chone Chow et al. "Structure of an extracellular gp130 cytokine receptor signaling complex". In: *Science* 291.5511 (2001), pp. 2150–2155.

[93] Nina Adam et al. "Unraveling viral interleukin-6 binding to gp130 and activation of STAT-signaling pathways independently of the interleukin-6 receptor". In: *Journal of virology* 83.10 (2009), pp. 5117–5126.

[94] Jürgen Scheller et al. "The pro-and anti-inflammatory properties of the cytokine interleukin-6". In: *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research* 1813.5 (2011), pp. 878–888.

[95] Christopher A Hunter and Simon A Jones. "IL-6 as a keystone cytokine in health and disease". In: *Nature immunology* 16.5 (2015), pp. 448–457.

[96] Marco Kaschner et al. "A combination of mutational and computational scanning guides the design of an artificial ligand-binding controlled lipase". In: *Scientific Reports* 7 (2017).

[97] Peixiang Ma et al. "Conformational Polymorphism in Autophagy-Related Protein GATE-16". In: *Biochemistry* 54.35 (2015), pp. 5469–5479.

[98] Oliver Schillinger et al. "Molecular Dynamics Simulations Reveal Key Roles of the Interleukin-6 Alpha Receptor in the Assembly of the Human Interleukin-6 Receptor Complex". In: *The Journal of Physical Chemistry B* (2017).

[99] Marcus B Kubitzki, Bert L de Groot, and Daniel Seeliger. "Protein DynamicsProtein Dynamics: From Structure to Function". In: *From Protein Structure to Function with Bioinformatics*. Springer, 2017, pp. 393–425.

[100] Lauren Boldon, Fallon Laliberte, and Li Liu. "Review of the fundamental theories behind small angle X-ray scattering, molecular dynamics simulations, and relevant integrated application". In: *Nano reviews* 6 (2015).

[101] J Andrew McCammon, Bruce R Gelin, and Martin Karplus. "Dynamics of folded proteins." In: *Nature* 267.5612 (1977), p. 585.

[102] Lars V Bock et al. "Energy barriers and driving forces in tRNA translocation through the ribosome". In: *Nature structural & molecular biology* 20.12 (2013), pp. 1390–1396.

[103] Isseki Yu et al. "Biomolecular interactions modulate macromolecular structure and dynamics in atomistic model of a bacterial cytoplasm". In: *eLife* 5 (2016), e19274.

[104] Morgan Lawrenz, Diwakar Shukla, and Vijay S Pande. “Cloud computing approaches for prediction of ligand binding poses and pathways”. In: *Scientific reports* 5 (2015).

[105] Alessandro Barducci, Giovanni Bussi, and Michele Parrinello. “Well-tempered metadynamics: a smoothly converging and tunable free-energy method”. In: *Physical review letters* 100.2 (2008), p. 020603.

[106] James McCarty and Michele Parrinello. “A variational conformational dynamics approach to the selection of collective variables in metadynamics”. In: *arXiv preprint arXiv:1703.08777* (2017).

[107] Vijay S Pande, Kyle Beauchamp, and Gregory R Bowman. “Everything you wanted to know about Markov State Models but were afraid to ask”. In: *Methods* 52.1 (2010), pp. 99–105.

[108] Kresten Lindorff-Larsen et al. “Systematic validation of protein force fields against experimental data”. In: *PloS one* 7.2 (2012), e32131.

[109] Jeetain Mittal and Robert B Best. “Tackling force-field bias in protein folding simulations: folding of Villin HP35 and Pin WW domains in explicit water”. In: *Biophysical journal* 99.3 (2010), pp. L26–L28.

[110] João Henriques, Carolina Cragnell, and Marie Skepö. “Molecular dynamics simulations of intrinsically disordered proteins: force field evaluation and comparison with experiment”. In: *Journal of chemical theory and computation* 11.7 (2015), pp. 3420–3431.

[111] Yue Shi et al. “The polarizable atomic multipole-based AMOEBA force field for proteins”. In: *Journal of chemical theory and computation* 9.9 (2013), p. 4046.

[112] Harry A Stern. “Molecular simulation with variable protonation states at constant pH”. In: *The Journal of chemical physics* 126.16 (2007), 04B627.

[113] Herman JC Berendsen and Steven Hayward. “Collective protein dynamics in relation to function”. In: *Current opinion in structural biology* 10.2 (2000), pp. 165–169.

[114] Ulf Hensen et al. “Exploring protein dynamics space: the dynasome as the missing link between protein structure and function”. In: *PloS one* 7.5 (2012), e33931.

Chapter 6

Appendices

6.1 Appendix A: Supporting Information – A combination of mutational and computational scanning guides the design of an artificial ligand-binding controlled lipase

Supplementary Materials

A combination of mutational and computational scanning guides the design of an artificial ligand-binding controlled lipase

Marco Kaschner, Oliver Schillinger, Timo Fettweiss, Christina Nutschel, Frank Krause, Alexander Fulton, Birgit Strodel, Andreas Stadler, Karl-Erich Jaeger and Ulrich Krauss*

Supplementary Methods

Chemicals. If not mentioned otherwise, all chemicals were purchased from Carl Roth GmbH (Karlsruhe Germany) or Sigma-Aldrich (St. Louis, MO, USA) in highest purity available.

General Molecular Biological techniques. Isolation of recombinant plasmids, gel extraction of DNA fragments, DNA ligation, and transformation into *E. coli* strains were carried out according to standard laboratory protocols [1].

Whole-protein site-saturation scanning mutagenesis. A full library containing every possible single-point amino-acid substitution of BsLA was generated using a statistically optimized two-stage procedure [2]. The results of this mutagenesis and a detailed description of the experimental set up have been described previously [3].

Construction of gene fusions. All plasmids used in this study are shown in Supplementary Table 1. The CitAP-BsLA gene fusion was constructed by subcloning a synthetic gene fragment coding for the isolated CitAP domain into a BsLA containing expression vector. The periplasmic CitA PAS domain (CitAP) DNA fragment was obtained as a synthetic gene fragment (Eurofins Genomics, Ebersberg, Germany) (pCR2.1-CitAP). The pCR2.1-CitAP vector was hydrolyzed using *Nde*I and *Hind*III restriction endonucleases. The resulting CitAP gene fragment was subsequently cloned into a similarly hydrolyzed pET28a-based BsLA expression vector (pET28a-nLOV-BsLA) which already contained a gene fusion consisting of the YtvA-LOV gene region and the full-length BsLA encoding gene [4]. Please note that, by hydrolyzing pET28a-YLOV-BsLA with *Nde*I and *Hind*III, the YtvA-LOV domain encoding segment is released enabling substitution of the LOV domain encoding gene fragment with CitAP PAS. Using this strategy a tripartite gene fusion consisting of the CitAP PAS domain (residues 45 to 177 of full-length CitA), an YtvA Jα-linker segment (residues 132 to 147 of full-length YtvA) and the BsLA encoding gene (residues 1 to 181 of full-length BsLA) was constructed. Additionally, a cleavage site for the Factor Xa protease was introduced between the YtvA Jα linker and the first amino acid of the BsLA domain (amino acid sequence: IEGRE). Before heterologous expression, all constructs were verified by sequencing (SeqLab GmbH, Göttingen, Germany). Additionally, all gene fusions contained an expression vector derived N-terminal His6-Tag (amino acid sequence: MGSSHHHHHHSSGLVPRGSH) to enable the easy purification of the recombinant protein from *E. coli* by immobilized metal affinity chromatography (IMAC).

Bacterial strains and plasmids. All strains and plasmids used in this study are listed in Supplementary Table 1. All strains were grown either in Luria-Bertani (LB) broth (Carl Roth®, Arlesheim, Switzerland) or in autoinduction (AI) media (adapted from Studier et al. 2005 [5] for heterologous production of recombinant proteins. In brief, AI medium contained 15 g/L casein hydrolysate, 0.5 g/L glucose, 30 g/L yeast extract and 6,25 g/L glycerol in 100 mM potassium phosphate buffer pH 7 and 2 g/L lactose added for induction. Glucose and lactose solutions were autoclaved separately as stock solutions (50 g/L glucose, 20 g/L lactose) and mixed with the rest of the medium immediately before use. For maintenance of plasmids 50 µg/ml kanamycin was added to all media. CitAP-BsLA and wild-type BsLA were expressed in *E. coli* BL21(DE3) by using an overnight seed culture grown in LB medium (37 °C, 120 rpm) to inoculate 1 L of AI medium in 5 L Erlenmeyer flasks to an OD_{600nm} of 0.05. The cultures were grown at 37 °C for 2 hours at constant agitation (120 rpm) and then shifted to 15 °C and grown for another 72 hours.

Protein purification. CitAP-BsLA and wild-type BsLA were produced in *E. coli* BL21(DE3) as described above. The cell pellet (5 g cells, wet weight) was dissolved in 25 ml lysis buffer

(50 mM sodium phosphate buffer pH 8, supplemented with 300 mM NaCl and 10 mM imidazole). Cells were lysed by passing the cell-suspension 3-times through a chilled French Pressure Cell (ThermoScientific, Waltham, USA) at a constant pressure of 1100 bar. The soluble fraction was separated from cell debris by centrifugation for 20 minutes at 38465 x g at 4 °C. The lysate was purified by immobilized metal affinity chromatography using a Superflow Ni-NTA resin (QIAGEN, Hilden, Germany). All purification steps were carried out at room-temperature employing an ÄKTApurifier FPLC system (GE Healthcare, Buckinghamshire, UK) fitted with a XK16/20 column. The column was equilibrated with lysis buffer. The crude cell-free extracts were loaded using a 50 ml Superloop^{TM} (GE Healthcare, Buckinghamshire, UK) and unspecifically bound proteins were removed by washing the column with 50 mM sodium phosphate buffer pH 8, supplemented with 300 mM NaCl and 20 mM imidazole. Elution of the His6-tagged target proteins was performed in 50 mM sodium phosphate buffer pH 8, supplemented with 300 mM NaCl and 250 mM imidazole. The purity of the eluted fractions was evaluated densitometrically by SDS-PAGE. Pure fractions were pooled and the buffer was exchanged to 10 mM Glycin pH 10 supplemented with 10 mM NaCl using a Sephadex^{TM} G25 (GE Healthcare, Buckinghamshire, UK) column. Desalted protein samples were concentrated using either Vivaspin20 (Sartorius, Göttingen, Germany) centrifugal concentrators or Vivacell (Sartorius, Göttingen, Germany) pressure cells employing a molecular weight cutoff of 10.000 Da. Concentrated samples were frozen in liquid nitrogen and stored at -20 °C until further use. For analytical ultracentrifugation (AUC) and small angle X-ray scattering (SAXS) analyses, CitAP-BsLA was further purified by preparative size-exclusion chromatography (SEC) by employing a Superdex^{TM} HR 10/30 column (GE Healthcare, Buckimhamshire, UK) and an ÄKTApurifier FPLC system at a constant flow-rate of 0.75 ml/min in 10 mM glycine buffer pH 10 supplemented with 10 mM NaCl.

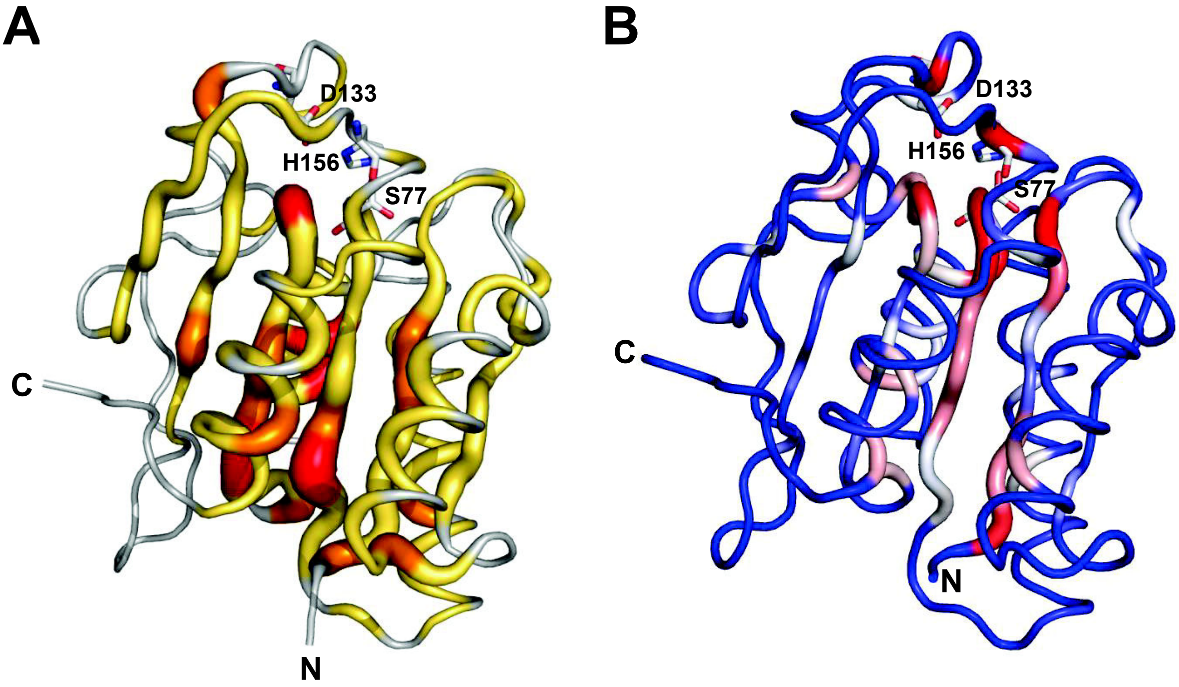
Molecular dynamics (MD) simulations. All starting models (generated as described in the main manuscript) were energy minimized, equilibrated and subjected to 100 ns molecular dynamic (MD) simulations using Gromacs 5.0 [6]. Simulations were performed in the NTP ensemble at a temperature of 298 K and a pressure of 1 bar. The amber99sb-ildn-NMR force field [7] was used with the TIP3P water model [8] and sodium and chloride ions were added to neutralize the system and mimic a salt concentration of 10 mM. Force field parameters for citrate were taken from the general amber force field (GAFF) [9] using the program acpype [10]. Long-range electrostatics were treated with the PME algorithm [11]. Short range electrostatic and van der Waals interactions were computed up to a cut-off of 1.2 nm. To increase the time step length to 5 fs, virtual sites were introduced for all hydrogens [12]. As this cannot be done automatically for citrate, the simulations containing citrate did not use virtual sites and a time step of 2 fs was used instead. χ values were computed every 200 ps.

Supplementary Table 1. Bacterial strains and plasmids used in this study.

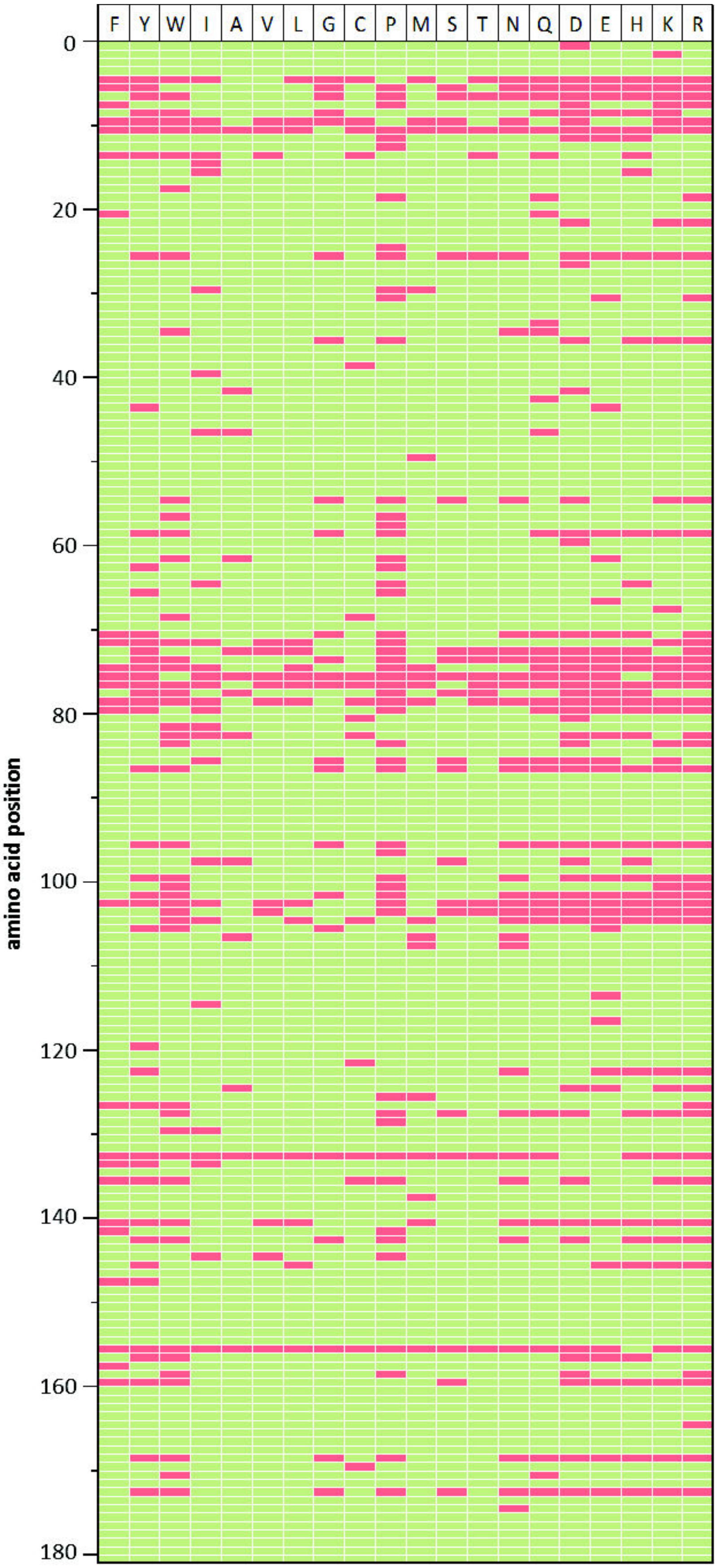
Bacterial strain	**Relevant genotype**	**Reference**
Escherichia coli DH5α	F– Φ80*lac*ZΔM15 Δ(*lac*ZYA-*arg*F) U169 *rec*A1 *end*A1 *hsd*R17 (r_K^-, m_K^+) *gal*$^-$ *pho*A *sup*E44 λ$^-$ *thi*$^-$1 *gyr*A96 *rel*A1	Invitrogen
Escherichia coli BL21(DE3)	F– *ompT gal dcm lon hsd* S_B(r_B^-, m_B^-) λ(DE3[*lac*I *lac*UV5-T7 gene 1 *ind*1 *sam*7 *nin*5])	Invitrogen
Plasmid name	**Relevant features**	**Reference**
pET28a	Km^R, P_{T7}, N-terminal His-tag, *lacI*	Novagen
pET28a-CitAP-BsLA	pET28a derivative, P_{T7}>CitAP-BsLA	this work
pET28a-BsLA	pET28a derivative, P_{T7}>BsLA	this work
pET22-BsLA	pET22b derivative, P_{T7}>BsLA	[13]

Evolutionary coupling analysis of BsLA

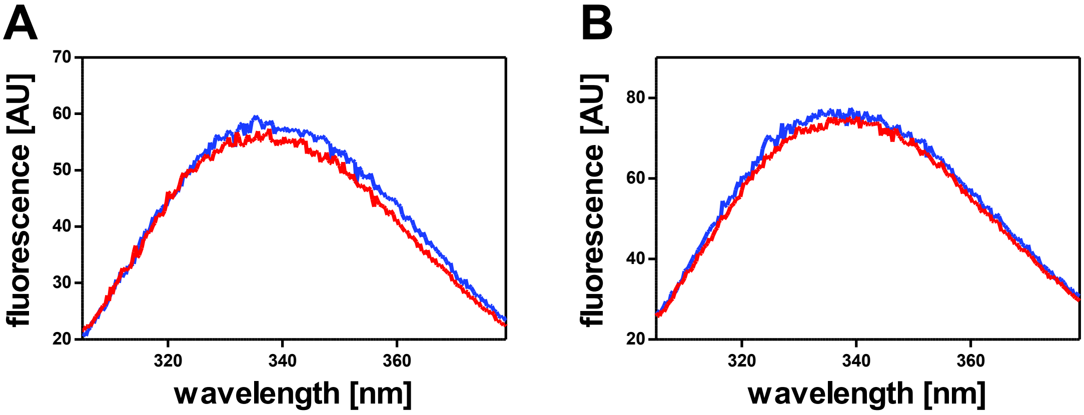
To further validate the evolutionary-coupling analyses presented in the main manuscript, a more restricted search was conducted, restricting the number of aligned sequences to 20.000, which marginally increased the alignment coverage from 168 (unrestrained run) to 176 out of 181 residues of BsLA. While the magnitude of the observed EC values differs between the two computations, the overall distribution of coupled residues remained largely the same (Supplementary Figure 1).



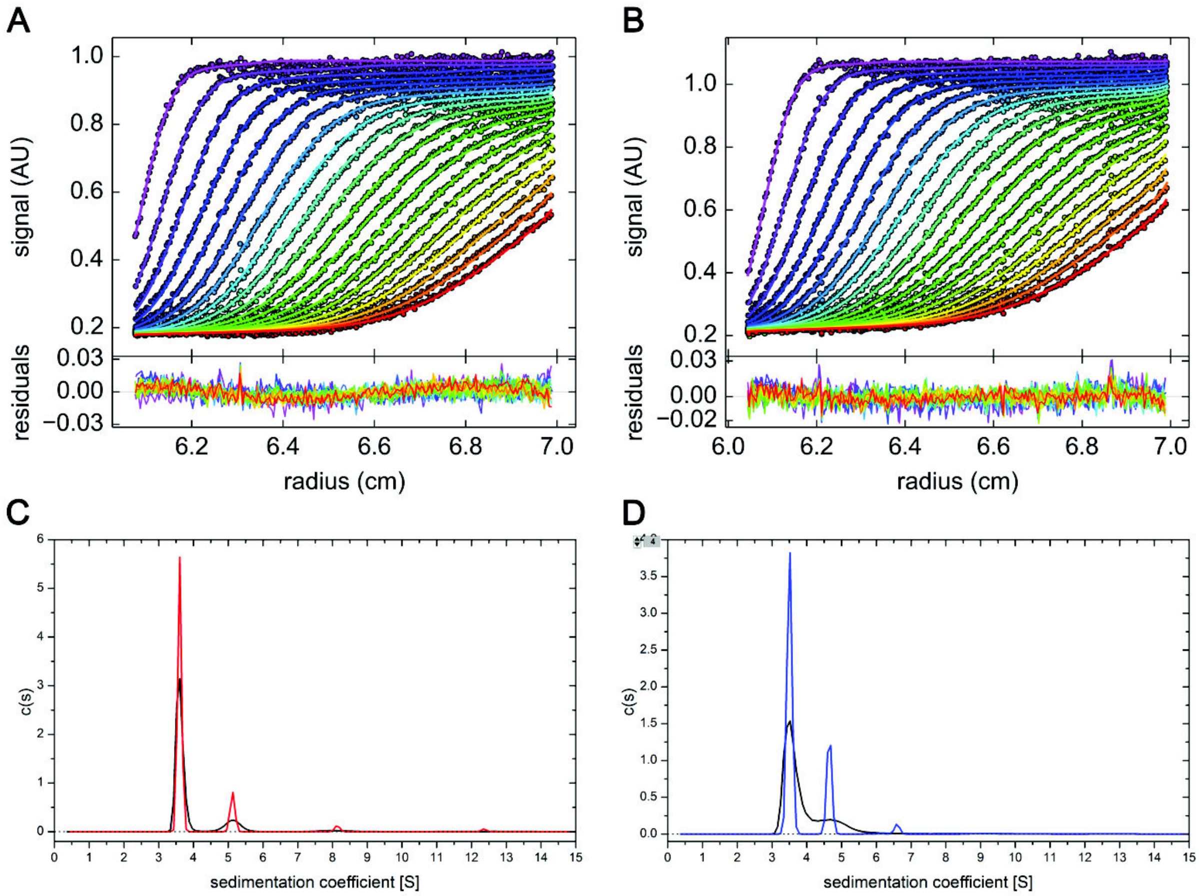
Supplementary Figure 1: Comparison of evolutionary-coupling analyses for a more restricted run (A) and site-saturation scanning mutagenesis data (B) mapped onto the X-ray structure of BsLA. Evolutionary coupled residues were inferred from a multiple sequence alignment containing 20.000 sequences (E-value cut-off 10E-3) using the EVcoupling webserver (www.evfold.org). The obtained EC scores were mapped onto the X-ray structure of BsLA (PDB Entry: 1I6W) [14]. The magnitude of the obtained EC scores is color coded (low values in yellow; high values in red). Additionally EC scores are encoded by sausage thickness representing the magnitude of the EC score. The number of inactive BsLA variants per residue was obtained from comprehensive whole protein site-saturation mutagenesis data (B) and mapped onto the BsLA X-ray structure. The fraction of inactive variants is encoded by color (blue: low values; red high values) and sausage thickness. The N- and C-terminus of BsLA are marked accordingly. The residues of the catalytic triad, Ser77, Asp133 and His156 are shown as sticks with oxygen in red, carbon in grey and nitrogen atoms in blue.



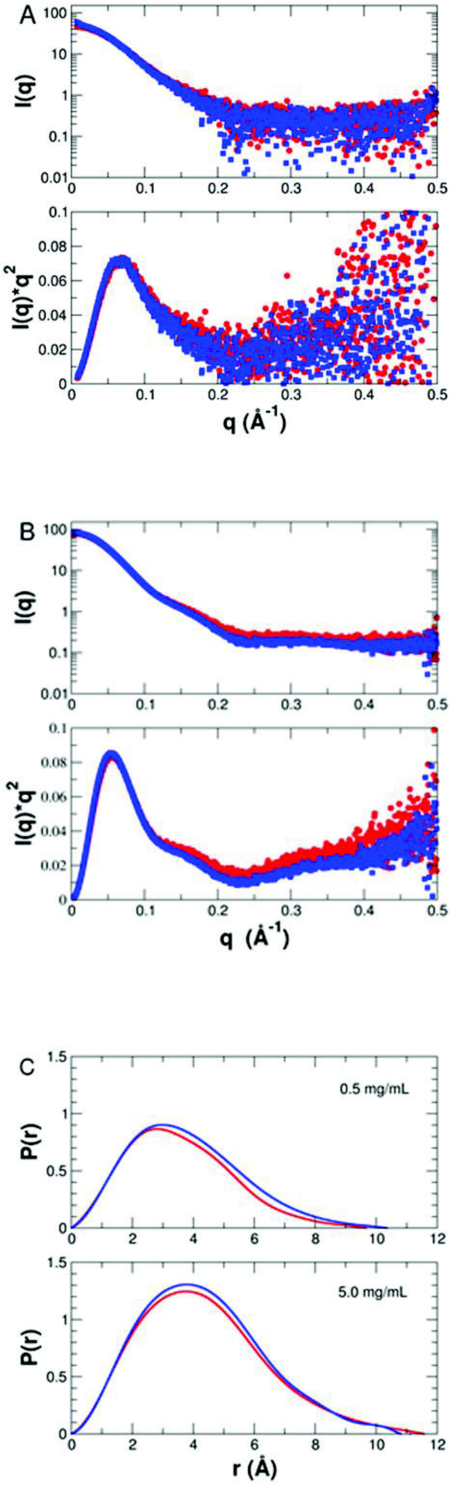
Supplementary Figure 2: Heatmap of inactive BsLA variants (red = inactive, green =active), Positions are indicated numerically and the substitutions by their amino acid abbreviation.



Supplementary Figure 3: Tryptophan fluorescence of CitAP-BsLA (A) and wild-type BsLA (B). Protein samples, diluted to 3 µM in 10 mM glycine buffer pH 10 supplemented with 10 mM NaCl, were excited at 295 nm. All fluorescence emission spectra were recorded in the presence (red line) and absence (blue line) of 1 mM sodium citrate.



Supplementary Figure 4: Exemplary sedimentation velocity analysis of 0.5 mg/mL CitAP-BsLA solutions with (A and C) and without citrate (B and D), respectively, using absorbance detection. (A and B) The top panels show raw data (circles) and best fits (lines). For clarity, only every second scan of the data set is shown. The bottom panels display best fit residuals of the plotted scans. (C and D) The conventional sedimentation coefficient distributions c(s) are shown as black solid lines. In a secondary analysis, the $c^{(P\delta)}(s)$ distributions based on the prior expectation that each protein sample exclusively features monodisperse species are shown as red (with citrate) and blue lines (without citrate), respectively. The relative abundances of monomers and dimers are derived from the $c^{(P\delta)}(s)$ distributions.



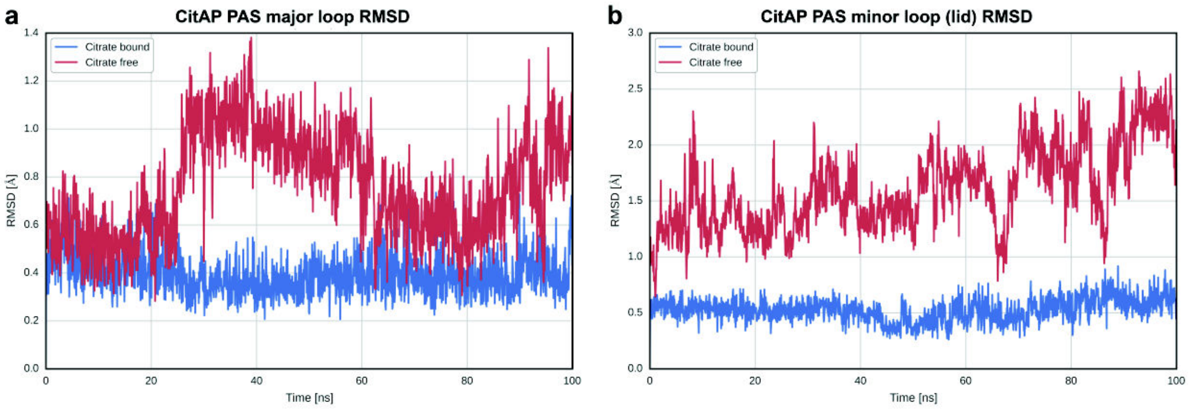
Supplementary Figure 5: SAXS data obtained for CitAP-BsLA. A) Scattering curves at low concentration (mainly monomer) and at B) high concentration (mainly dimer) with citrate (red line) and without citrate (blue line). (C) Distance distribution function P(r) derived from scattering data of CitAP-BsLA at 0.5 mg/ml and at 5.0 mg/ml. The P(r) function was derived from scattering data of samples of the indicated concentration with citrate (red line) and without citrate (blue line) by using the program DATGNOM of the ATSAS package [15].

Computational modelling and molecular dynamics simulations of the CitAP-BsLA complex

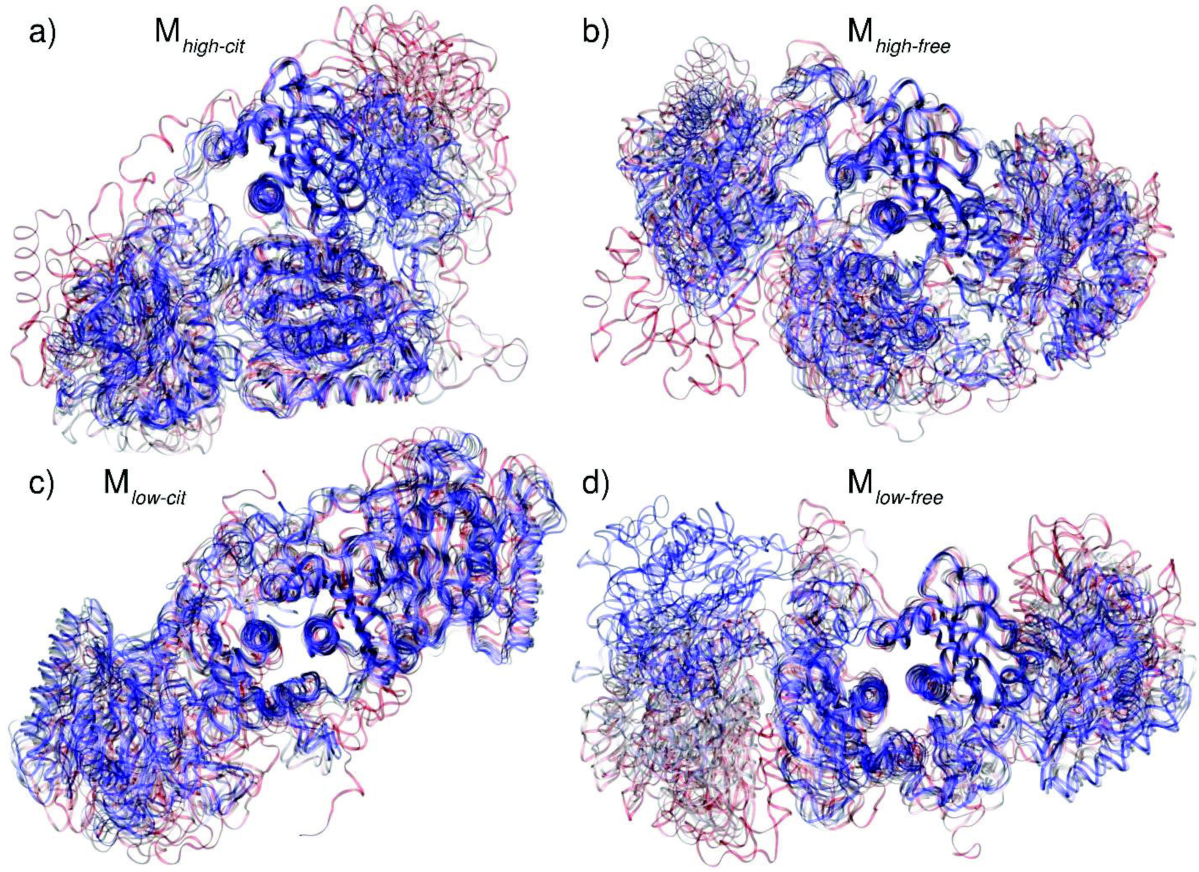
Initial monomer models were obtained from SAXS data at low concentration (0.5 mg/ml) where CitAP-BsLA is predominately monomeric. Monomer models were built using the program BUNCH of the ATSAS package [15] employing chain A of the citrate-bound CitAP-PAS structure (PDB-ID: 2J80) [16] and the monomeric BsLA structure (PDB-ID: 1I6W) [14] in the q-range from 0.01 up to 0.15 Å^{-1} for the citrate saturated and 0.017 up to 0.15 Å^{-1} for the citrate free solution. The constructed BUNCH models gave a good fit to the SAXS data with χ=1.06 for the citrate saturated and χ=1.12 for the citrate free protein solution. The R_g and D_{max} of the citrate-saturated model are 2.80 nm and 9.49 nm, and for the citrate free model 3.03 nm and 10.5 nm, respectively, which is in agreement with the measured experimental values. A dimeric assembly was subsequently generated by superimposing the BUNCH-derived monomer model onto the dimeric citrate-bound CitAP-PAS structure (PDB-ID 2J80) resulting in the model M$_{\text{low-cit}}$. To derive a model with a citrate-free CitAP-PAS conformation, we had to employ a CitAP-PAS structure which shows an open citrate-binding site. Unfortunately, the available citrate-free structure of CitAP-PAS (PDB-ID: 2V9A) [16], lacks electron density in the surface exposed major loop (residues 68 to 89 of CitAP-PAS) of the citrate-binding site [16], which was previously interpreted as increased flexibility of the respective loops, which are directly involved, and hence stabilized by binding of citrate [16]. Therefore, to obtain an all-atom model of citrate-free CitAP-PAS we used the closed (citrate-bound) CitAP-PAS structure (PDB-ID: 2J80), removed the citrate and performed a 100 ns MD simulation. During the simulation, the major loop of the CitAP-PAS domain shows increased mobility i.e. manifested as an increased RMSD over the trajectory (Supplementary Figure 6a). Moreover, the CitAP-PAS domain opens up and reaches a stable conformation with an open lid (minor loop movement) (see Figure 6a and Supplementary Figure 6), thus corroborating previous NMR and X-ray crystallographic studies of CitAP-PAS [16]. In contrast, the citrate-bound structure did not undergo any major structural rearrangements (Supplementary Figure 6). The lid of the corresponding open (citrate-free) CitAP-PAS model was subsequently transferred to the M$_{\text{low-cit}}$ model resulting in M$_{\text{low-free}}$. In addition, a set of two dimer models was generated using the corresponding CitAP-BsLA monomer models, assembling them using the program SASREF [15], which models oligomer complexes by optimizing the subunit orientation against experimental SAXS data, and then generating the open form of the CitAP-PAS domain as outlined above. This resulted in the models M$_{\text{high-cit}}$ and M$_{\text{high-free}}$, respectively.

Supplementary Table 2: Summary of the model-building strategy

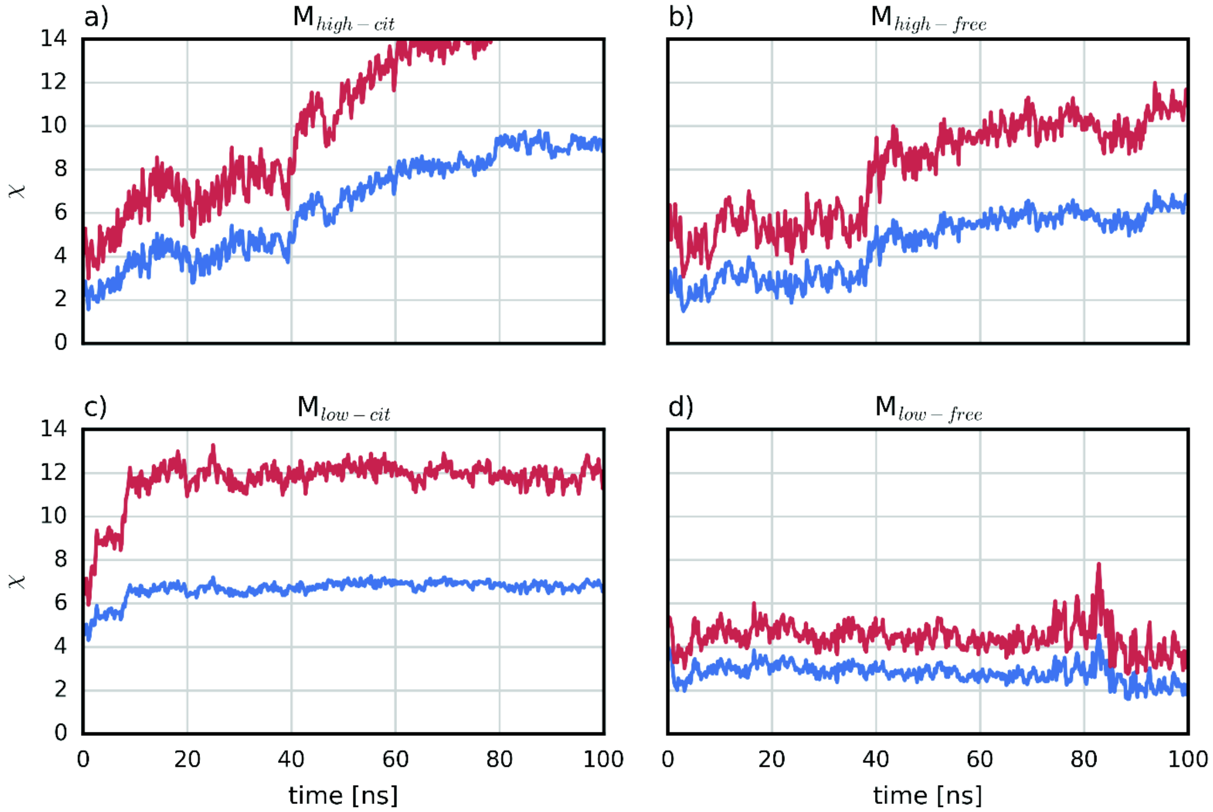
Model name	**Protein concentration**	**Dimer/Monomer ratio (SAXS)**	**Citrate concentration**	**Citrate binding pocket**	**Rigid body fit program**
$M_{\text{high-cit}}$	high (5 mg/ml)	95:5	1 mM	bound/closed	SASREF + BUNCH
$M_{\text{high-free}}$	high (5 mg/ml)	98:2	0	free/open	SASREF + BUNCH
$M_{\text{low-cit}}$	low (0.5 mg/ml)	14:86	1 mM	bound/closed	BUNCH
$M_{\text{low-free}}$	low (0.5 mg/ml)	27:73	0	free/open	BUNCH



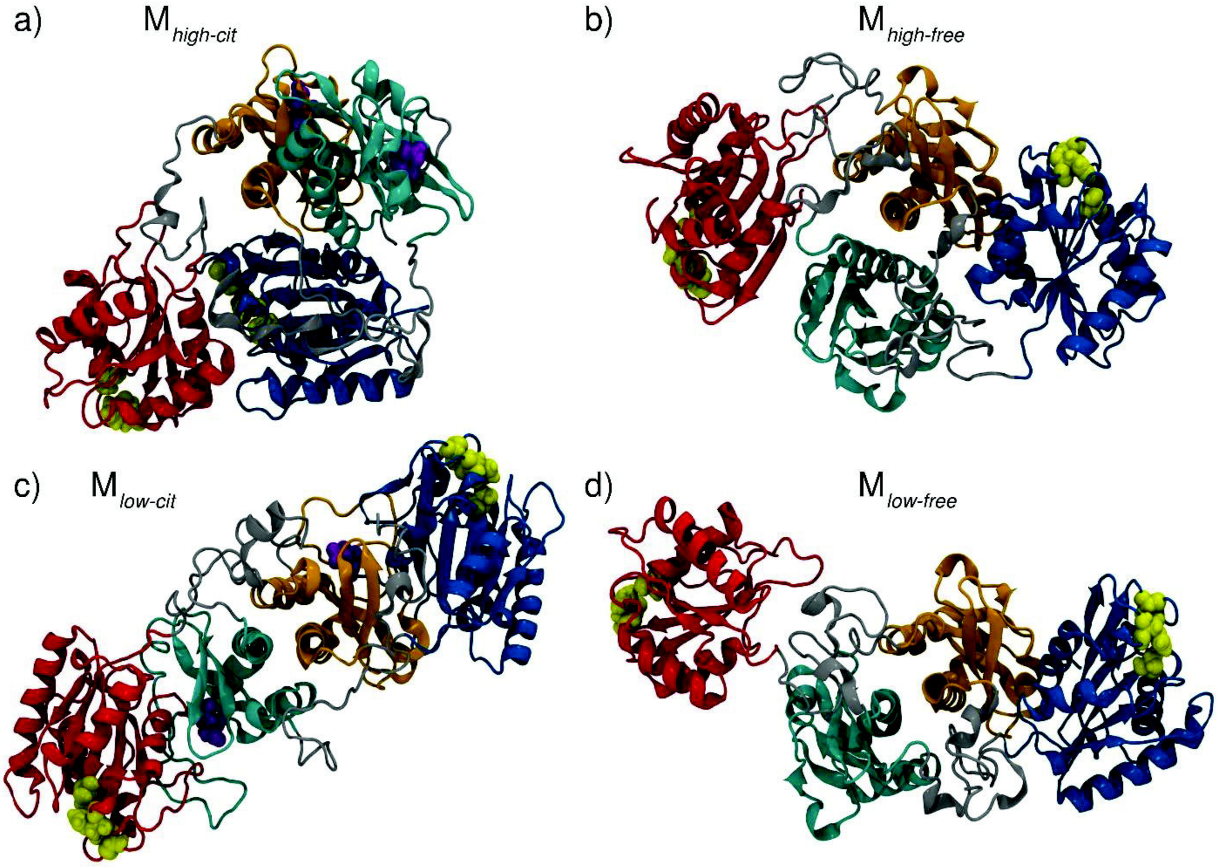
Supplementary Figure 6: (a) C$_\alpha$ RMSD of the CitAP PAS major loop (residues 68 to 89) and **(b)** minor loop (lid, residues 99-104) with respect to the mean positions during the citrate-bound (i.e. closed conformation) MD simulation. The trajectories were fitted to the closed crystal structure by minimizing the C$_\alpha$ RMSD of residues 12 to 94 and 109 to 128 (i.e. omitting residues in the opening lid and at the termini).



Supplementary Figure 7: Structural changes during the MD simulations of the models of the fusion protein dimer. The proteins are shown as ribbons and the colors represent structures at different times, changing from red at t = 0 ns to blue at t= 100 ns.

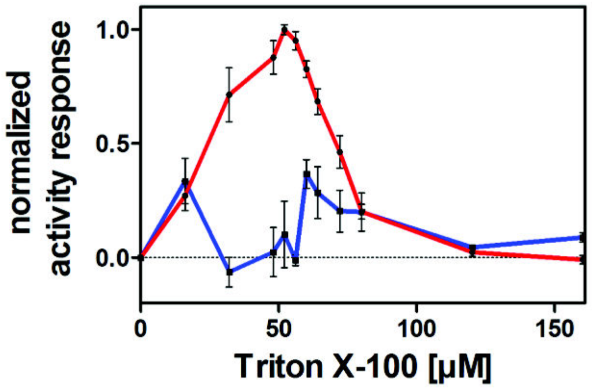


Supplementary Figure 8 Time evolution of χ during the MD simulations of the four fusion protein models. Deviation between the models and experimental data measured at high citrate concentration (blue) and without citrate (red) are shown. It should be noted that χ and not χ2 is presented as for the models in (a) – (c) χ2 can become large (> 100), making it difficult to present the results for all four models on the same scale.



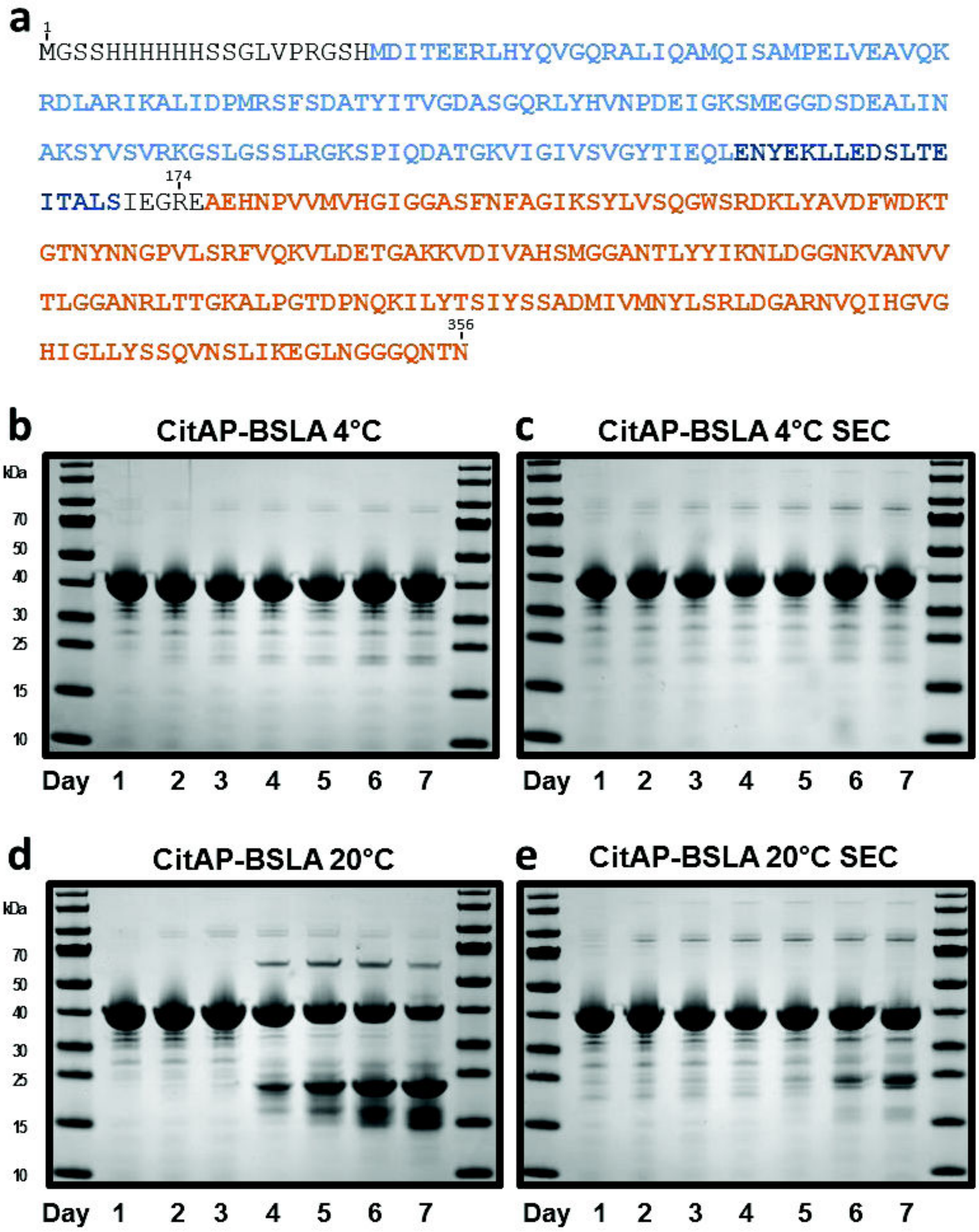
Supplementary Figure 9: Models of the fusion protein dimer at the end of the MD simulations. The proteins are represented in cartoon style with chain A being colored red (BsLA) and orange (CitAP), chain B is shown in blue (BsLA) and cyan (CitAP), and the linker and His-tag are colored gray in both chains. The BsLA residues forming the catalytic triad residues are represented as yellow van der Waals (vdW) surface, while in the models with citrate the ligand is shown as purple vdW surface.

Triton-X100 dependency of the citrate-dependent activity response of CitAP-BsLA and wild type BsLA



Supplementary Figure 10: Triton-X100 dependency of the citrate-dependent activity response of CitAP-BsLA (red line) and wild type BsLA (blue line). The data was normalized to the maximal acitivity response of CitAP-BsLA. Lipolytic activity was measured using *p*-nitrophenylbutyrate as substrate. Error bars depict the standard deviation of the mean derived from three independent measurements.

Proteolytic stability of purified CitAP-BsLA during long-term storage and citrate-dependent lipolytic activity of long-term stored CitAP-BsLA samples

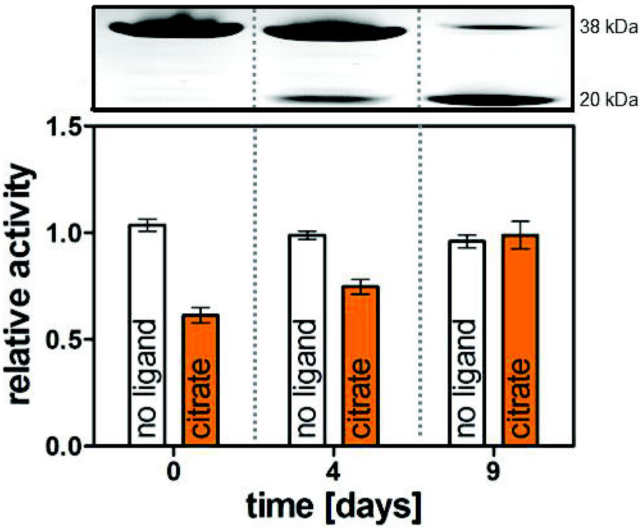

a
1
MGSSHHHHHHSSGLVPRGSHMDITEERLHYQVGQRALIQAMQISAMPELVEAVQK
RDLARIKALIDPMRSFSDATYITVGDASGQRLYHVNPDEIGKSMEGGDSDEALIN
AKSYVSVRKGSLGSSLRGKSPIQDATGKVIGIVSVGYTIEQLENYEKLLEDSLTE
174
ITALSIEGREAEHNPVVMVHGIGGASFNFAGIKSYLVSQGWSRDKLYAVDFWDKT
GTNYNNGPVLSRFVQKVLDETGAKKVDIVAHSMGGANTLYYIKNLDGGNKVANVV
TLGGANRLTTGKALPGTDPNQKILYTSIYSSADMIVMNYLSRLDGARNVQIHGVG
356
HIGLLYSSQVNSLIKEGLNGGGQNTN
b
CitAP-BSLA 4°C
c
CitAP-BSLA 4°C SEC
kDa
70
50
40
30
25
15
10
Day 1 2 3 4 5 6 7
Day 1 2 3 4 5 6 7
d
CitAP-BSLA 20°C
e
CitAP-BSLA 20°C SEC
kDa
70
50
40
30
25
15
10
Day 1 2 3 4 5 6 7
Day 1 2 3 4 5 6 7

Supplementary Figure 11: Sequence of the fusion protein (a) and SDS-PAGE analysis of the proteolytic stability of purified CitAP-BsLA during long-term storage. In (a) the amino acid sequence of the CitAP-BsLA fusion protein is shown. The CitAP-PAS domain is marked in blue and the BsLA domain is highlighted in orange. The first and last amino acids are labelled, as is the position of the Factor Xa protease cleavage site (recognition sequence IEGRE). The panels (b-e) illustrate the proteolytic stability of the purified fusion protein under different storage conditions. Samples were either purified by immobilized metal ion affinity chromatography (IMAC) or IMAC followed by preparative size exclusion chromatography SEC) (b) IMAC purified CitAP-BsLA stored at 4 °C. (c) IMAC and SEC purified CitAP-BsLA stored at 4 °C. (d) IMAC purified CitAP-BsLA stored at 20 °C. (e) IMAC and SEC purified CitAP-BsLA stored at 20 °C.

In order to address the storage stability of purified CitAP-BsLA, we stored purified samples for 7 days under different conditions. Samples purified by immobilized metal ion affinity chromatography (IMAC) or samples purified by IMAC and preparative size-exclusion chromatography (SEC) were stored at either 20 °C or 4 °C. At defined time intervals a sample was withdrawn and analysed by SDS-PAGE (Supplementary Figure 11, b-e).

When the purified protein is stored at 20 °C, an additional lower molecular weight band (approx. 20 kDa) appears in SDS-PAGE analyses after about 4 days, which is accompanied by the reduction of the band corresponding to the fusion protein (approx. 38 kDa) (Supplementary Figure 11d). The effect is largely absent when the protein is stored at 4 °C (Supplementary Figure 11a and b) and is less pronounced when the IMAC purified protein preparation is further purified by preparative size-exclusion chromatography (Supplementary Figure 11c and e). This suggests that during prolonged storage the purified CitAP-BsLA is most likely proteolytically cleavaged by a co-purified protease of the host used for heterologous expression. The resulting 20 kDa band can be assigned to a mixture of the CitAP-PAS domain and the BsLA domain, proteolytically cleaved in the linker region that contains a Factor Xa protease recognition site (Supplementary Figure 11a, sequence position approx. 174, sequence: IEGRE). A sample stored at 20 °C for up to 9 days still shows lipolytic activity indicating that the BsLA domain is structurally intact.

Thus, in a preparation, e.g. stored for an extended period at 20 °C, the BsLA and CitAP-PAS domain will to a large degree not be covalently linked anymore, which should abolish the signal-relay between the CitAP-PAS domain and the BsLA domain thereby abolishing the functional citrate-dependent response. To address this issue, the lipolytic activity in the presence and absence of 1 mM citrate of a sample stored for 0, 4 and 9 days at 20 °C was determined (Supplementary Figure 12). While the initial sample shows a pronounced citrate-dependent response and only one band in the corresponding SDS-PAGE analysis (see upper panel of Supplementary Figure 12), the sample stored for 4 days at 20 °C shows a defined band at a lower molecular weight (20 kDa) and correspondingly attenuated citrate-dependency. After 9 days, the band corresponding to the fusion protein (38 kDa) has nearly disappeared and a well pronounced band at around 20 kDa is visible in the corresponding SDS-PAGE analysis. The corresponding sample does not show any detectable citrate-dependency.



Supplementary Figure 12: Citrate-dependent activity response of an IMAC purified CitAP-BsLA sample stored for up to 9 days at 20 °C. The upper panel shows the SDS-PAGE analysis carried out using CitAP-BSLA samples stored for 0, 4 and 9 days. For clarity only the region showing the bands of the fusion protein (38 kDa) and of the proteolysis product (20 kDa) are shown. The lower panel shows the lipolytic activity of the same samples determined in either the presence (orange bars) or absence (white bars) of citrate. Lipolytic activity was measured using *p*-nitrophenylbutyrate as substrate. Error bars depict the standard deviation of the mean derived from three independent measurements.

Deconvolution of Far-UV CD spectra of CitAP-BsLA

The results of the deconvolution of CD spectra using the Convex-Constraint-Analysis Tool CCA+ [17] employing five pure secondary structural components (α-helix, random coil, parallel β-sheet, anti-parallel β-sheet and β-turn)[18,19] are presented in Supplementary Table 3. A direct comparison to the secondary structure content derived from the X-ray structures of the constituting parts (CitAP and BsLA) of the fusion protein is shown in Supplementary Table 4. The addition of Triton X100 (TX100) does not have a significant impact on the folding and secondary structure of CitAP-BsLA (Supplementary Table 3). Moreover, the secondary structural composition derived from the CD spectrum of CitAP-BsLA agrees very well with the theoretical secondary structure content derived from the X-ray structures of the components, which suggests that all structural domains of the fusion protein are well folded and functional.

Supplementary Table 3. Convex Constraint Analysis (CCA) Deconvolution of CitAP-BSLA far-UV CD data.

CitA-BSLA	**α-helix**	**β-sheet parallel**	**β-sheet antiparallel**	**Turn and other**	**random coil**
	% secondary structure (number of amino acids)				
- citrate	36.3 (129)	11.3 (40)	16.1 (57)	23.8 (85)	12.5 (45)
+ citrate	35.5 (126)	11.3 (40)	17.4 (62)	23.2 (83)	12.6 (45)
+ citrate + TX100	35.9 (128)	13.4 (47)	15.7 (56)	24.6 (88)	10.4 (37)
associated error (%)	0.021	0.002	0.004	0.009	0.002

Supplementary Table 4. Comparison of the CD predicted secondary structure content and the secondary structure content derived from DSSP analysis of CitAP-BSLA component structures.

	PDB ID	**Number amino acids**	**α-helix**	**β-sheet**	**Turn, random coil and other$^{\$}$**
			number of amino acids with given secondary structure		
His$_6$-tag§		20	-	-	20
DSSP CitA	2J80	129	51	47	31
Linker$^{\&}$	2J80; 2PR5	21	21	-	-
DSSP BSLA	1ISP	179	56	38	85
Sum:		349	128	85	136
CD CitA-BSLA		356	129	97	130

$^{\$:}$ includes all potential other secondary structure elements such as 3-10 helices, β-hairpins etc.$^{§:}$ theoretical secondary structure of the N-terminal 20 amino acid His6-tag. $^{\&:}$ The theoretical secondary structure composition was derived for the corresponding structural segments of the CitAP X-ray structure (2J80) and the corresponding YtvA X-ray structure (2PR5).

Supplementary References

1 Green, M. R., Sambrook, J. & Sambrook, J. *Molecular cloning : a laboratory manual*. 4th edn, (Cold Spring Harbor Laboratory Press, 2012).

2 Nov, Y., Fulton, A. & Jaeger, K. E. Optimal Scanning of All Single-Point Mutants of a Protein. *J Comput Biol* **20**, 990-997, doi:10.1089/cmb.2013.0026 (2013).

3 Fulton, A. *et al.* Exploring the Protein Stability Landscape: Bacillus subtilis Lipase A as a Model for Detergent Tolerance. *Chembiochem* **16**, 930-936, doi:10.1002/cbic.201402664 (2015).

4 Rahmen, N. *et al.* Exchange of single amino acids at different positions of a recombinant protein affects metabolic burden in Escherichia coli. *Microb Cell Fact* **14**, doi:ARTN 1010.1186/s12934-015-0191-y (2015).

5 Studier, F. W. Protein production by auto-induction in high-density shaking cultures. *Protein Expres Purif* **41**, 207-234, doi:10.1016/j.pep.2005.01.016 (2005).

6 Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1-2**, 19-25 (2015).

7 Li, D. W. & Bruschweiler, R. NMR-based protein potentials. *Angewandte Chemie* **49**, 6778-6780, doi:10.1002/anie.201001898 (2010).

8 Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J Chem Phys* **79**, 926-935, doi:Doi 10.1063/1.445869 (1983).

9 Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general amber force field. *Journal of computational chemistry* **25**, 1157-1174, doi:10.1002/jcc.20035 (2004).

10 Sousa da Silva, A. W. & Vranken, W. F. ACPYPE - AnteChamber PYthon Parser interfacE. *BMC research notes* **5**, 367, doi:10.1186/1756-0500-5-367 (2012).

11 Darden, T., York, D. & Pedersen, L. Particle Mesh Ewald - an N.Log(N) Method for Ewald Sums in Large Systems. *J Chem Phys* **98**, 10089-10092, doi:Doi 10.1063/1.464397 (1993).

12 Feenstra, K. A., Hess, B. & Berendsen, H. J. C. Improving efficiency of large time-scale molecular dynamics simulations of hydrogen-rich systems. *Journal of computational chemistry* **20**, 786-798 (1999).

13 Eggert, T. *Die lipolytischen Enzyme LipA und LipB von Bacillus subtilis: Charakterisierung und Optimierung mit gerichteter Evolution.* , Universität Bochum, (2001).

14 van Pouderoyen, G., Eggert, T., Jaeger, K. E. & Dijkstra, B. W. The crystal structure of Bacillus subtilis lipase: a minimal alpha/beta hydrolase fold enzyme. *J Mol Biol* **309**, 215-226 (2001).

15 Petoukhov, M. V. *et al.* New developments in the ATSAS program package for small-angle scattering data analysis. *J Appl Crystallogr* **45**, 342-350, doi:10.1107/S0021889812007662 (2012).

16 Sevvana, M. *et al.* A ligand-induced switch in the periplasmic domain of sensor histidine kinase CitA. *J Mol Biol* **377**, 512-523, doi:10.1016/j.jmb.2008.01.024 (2008).

17 Perczel, A., Hollosi, M., Tusnady, G. & Fasman, G. D. Convex constraint analysis: a natural deconvolution of circular dichroism curves of proteins. *Protein engineering* **4**, 669-679 (1991).

18 Buttani, V. *et al.* Conformational analysis of the blue-light sensing protein YtvA reveals a competitive interface for LOV-LOV dimerization and interdomain interactions. *Photoch Photobio Sci* **6**, 41-49, doi:10.1039/b610375h (2007).

19 Rani, R. *et al.* Conservation of dark recovery kinetic parameters and structural features in the pseudomonadaceae "short" light, oxygen, voltage (LOV) protein family: implications for the design of LOV-based optogenetic tools. *Biochemistry-Us* **52**, 4460-4473, doi:10.1021/bi400311r (2013).

6.2 Appendix B: Supporting Information – Conformational Polymorphism in Autophagy-Related Protein GATE-16

Ma et al.,

Conformational polymorphism in the autophagy-related protein GATE-16

Supporting Information

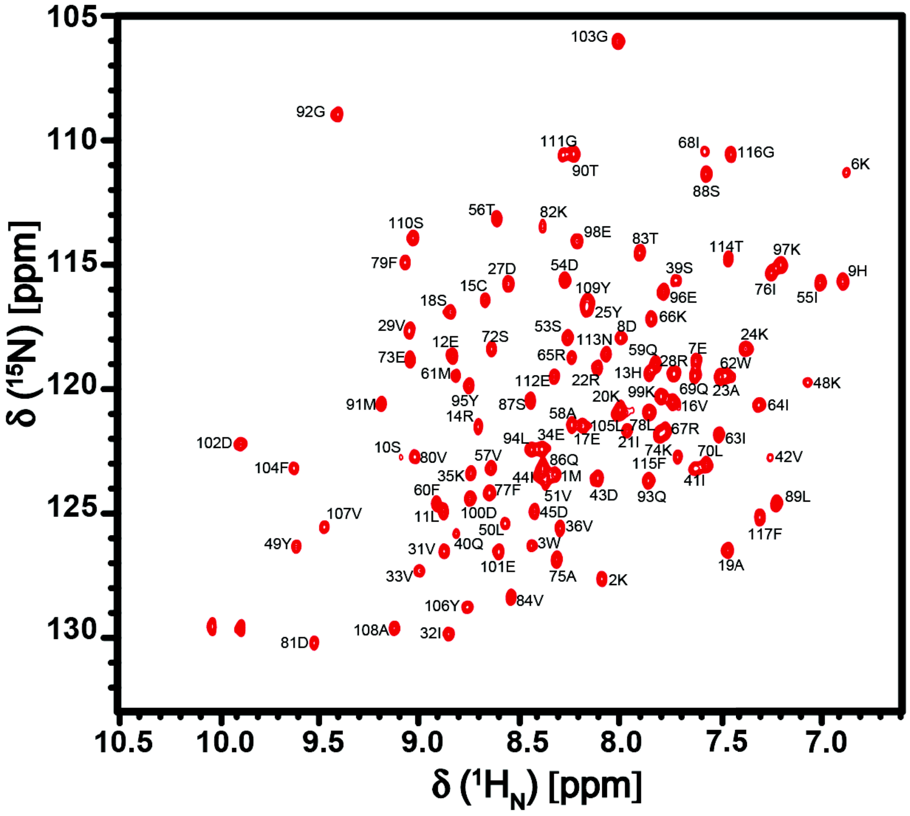


Figure S1. Assigned ^{1}H-^{15}N-TROSY HSQC spectrum of GATE-16. Assignment details are available in the Biological Magnetic Resonance Data Bank (entry 18827).

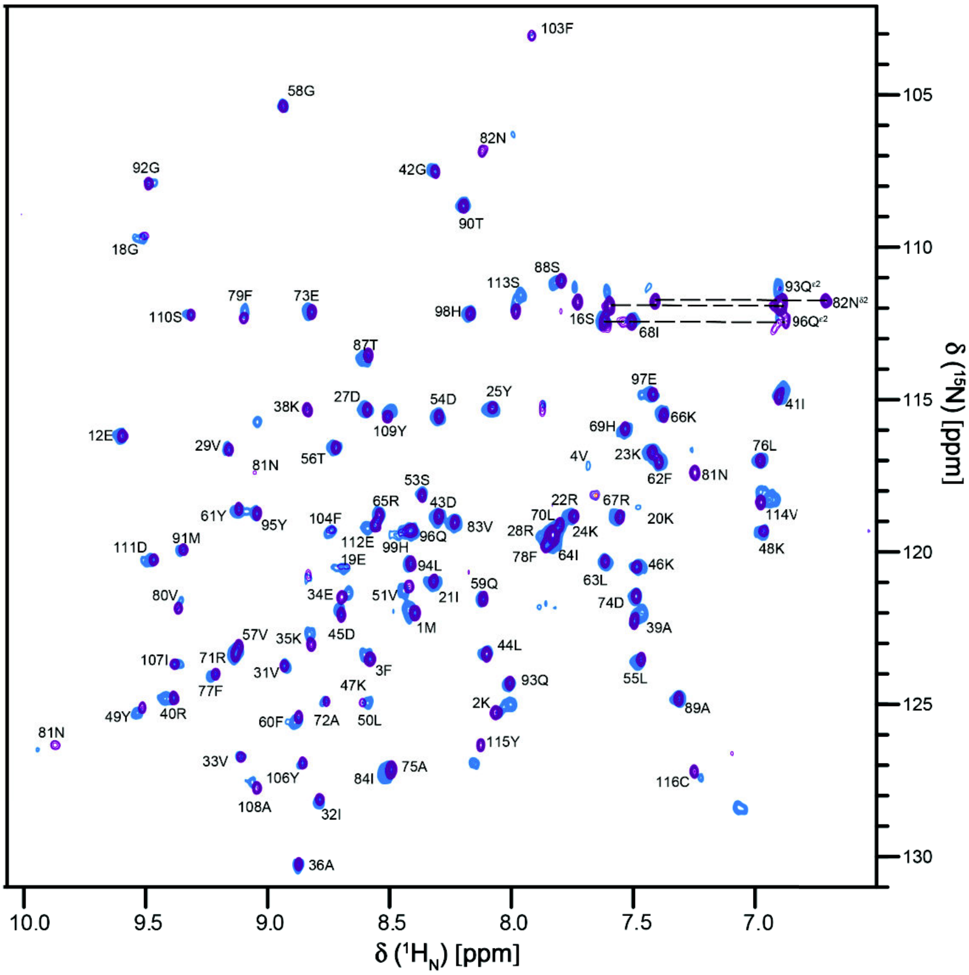


Figure S2. ^{1}H-^{15}N-TROSY HSQC spectra of non-lipidated ^{15}N-labeled GABARAP (G116CΔL117) (blue) and lipidated ^{15}N-GABARAP (G116CΔL117) attached to nanodiscs (purple). Lipidation was achieved by coupling to the thiol-reactive compound DPPE-MBP (1,2-dipalmitoyl-*sn*-glycero-3-phosphoethanolamine-N-[4-(p-maleimido-phenyl) butyramide]).

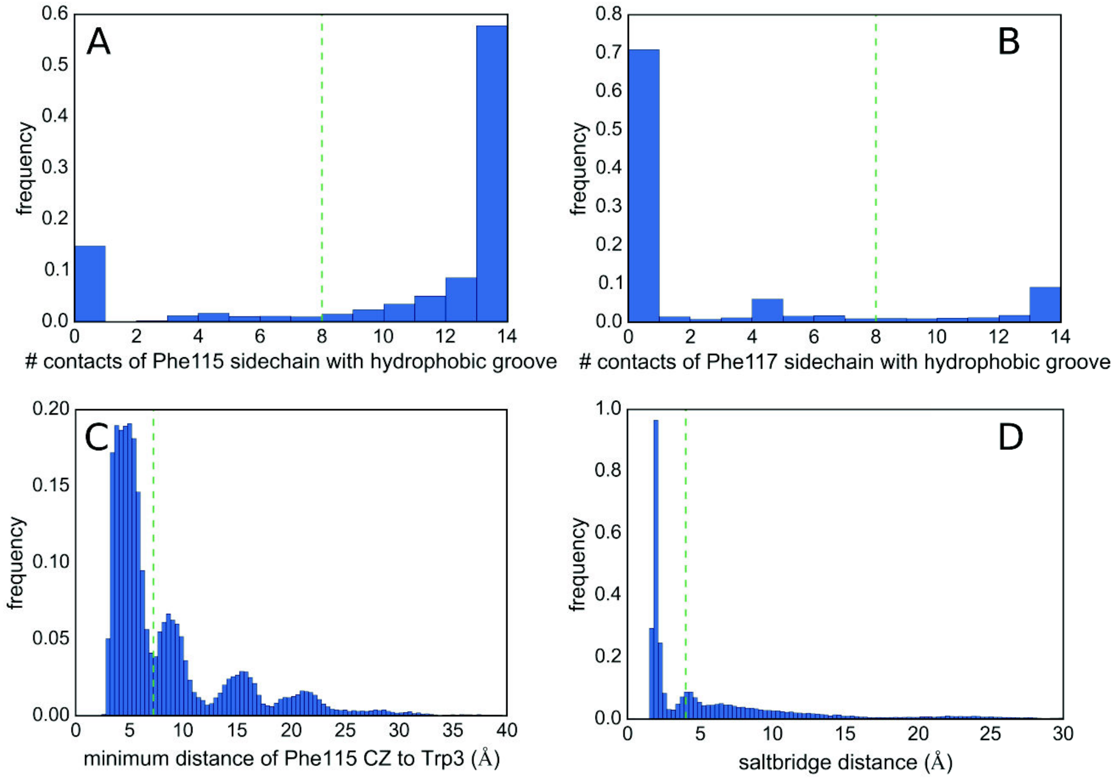


Figure S3. Histograms for the number of contacts of (A) the side chain of Phe115 and (B) the side chain of Phe117 with the hydrophobic groove formed by Met1, Val36, Phe79, and the aliphatic portion of Lys82 . (C) Distance distribution between the Cζ atom of Phe115 and any of the Trp3 atoms. (D) Distance distribution between the N atom of Met1 and the closest of the two carboxylate oxygens of Phe117. Histograms (A) to (D) were used for the definition of the criteria (1) to (4) in order to cluster the structures sampled in the target replica of the HREMD simulation. The cut-off values are indicated by dashed green lines.

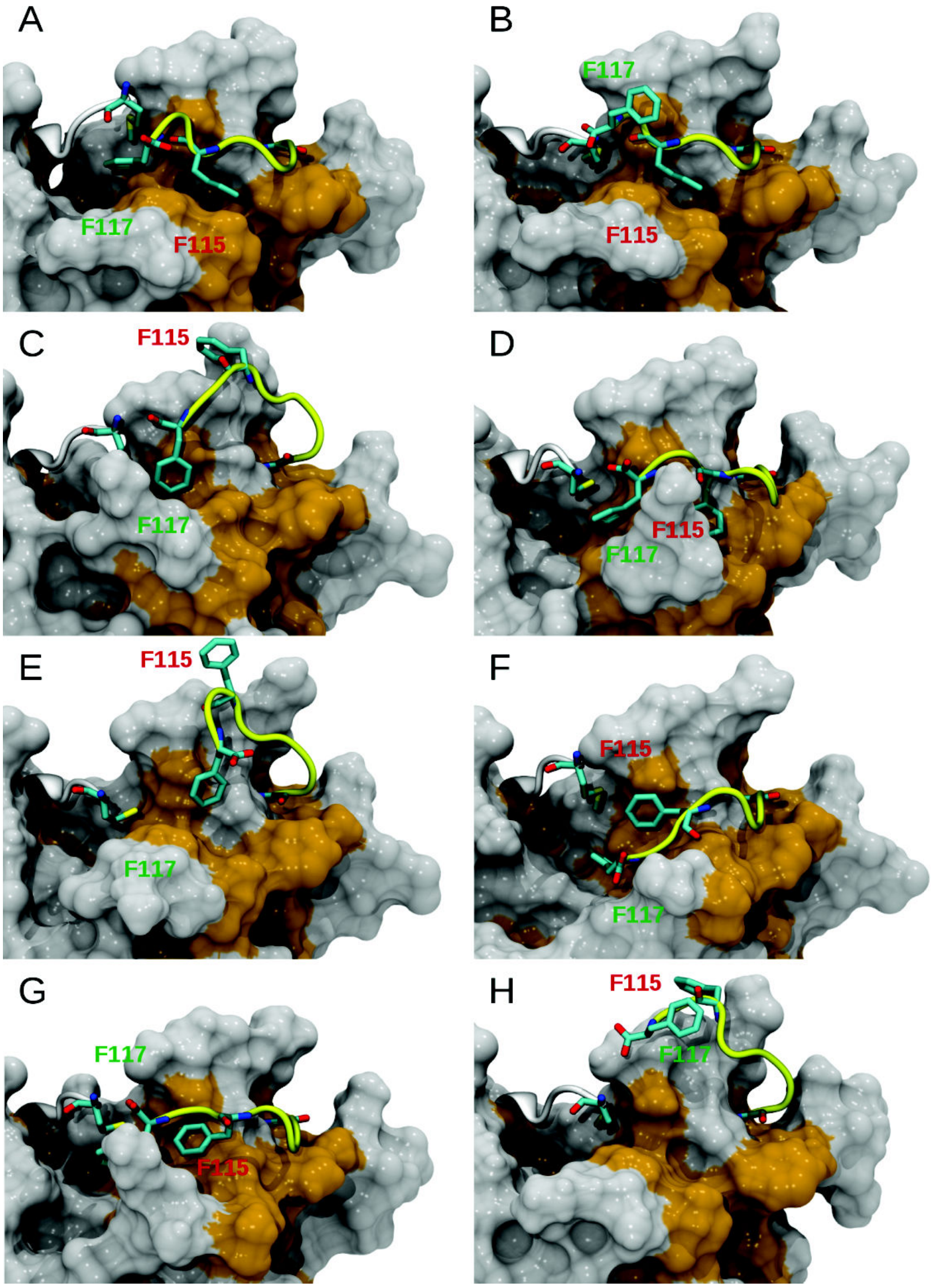


Figure S4. Representative structures for states 5 to 12 (panels A to H) obtained from the HREMD simulation. Most residues of GATE-16 are shown in surface representation with hydrophobic residues being highlighted in orange. Residues Met1-Trp3 and Gly111-Phe117 are shown in cartoon and the C-terminus is colored yellow. The side chains of the residues of particular interest (Met1, Phe115, Phe117) are shown explicitly.

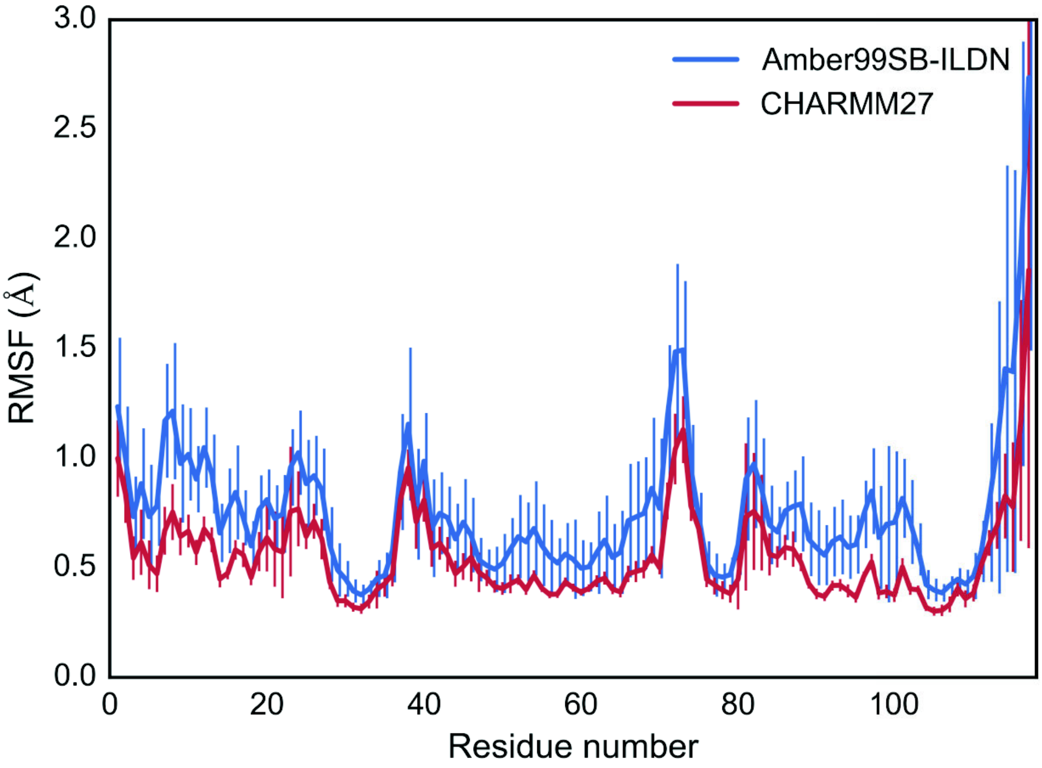


Figure S5. Cα root-mean-square fluctuations (RMSFs) for MD simulations with Amber99SB-ILDN and CHARMM27. RMSF values for Amber99SB-ILDN have been averaged over the 29 20-ns MD runs. For CHARMM27, the three 200-ns MD runs were first divided into subtrajectories of 20 ns length, yielding 30 subtrajectories. The RMSF values were determined for each subtrajectory, where the structures of each subtrajectory were superimposed on the initial structure of the subtrajectory in question, and the RMSFs then averaged over the 30 subtrajectories. Error bars indicate one standard deviation. For Amber99SB-ILDN the standard deviations are greater than for CHARMM27 as the starting structures for these 29 MD simulations were taken from the preceding HREMD simulation. Thus the structures are more diverse than in the three MD simulations with CHARMM27, which were all initiated from the same crystal structure.

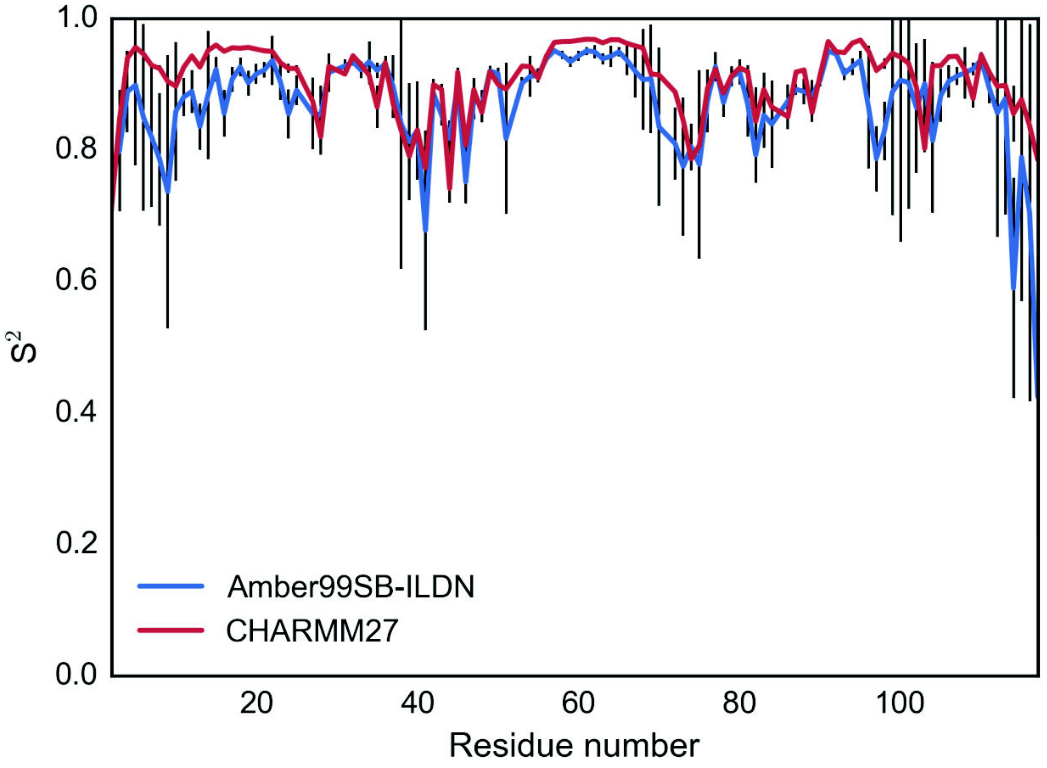


Figure S6. Comparison of S^2 value estimates computed from MD simulations with the Amber99SB-ILDN and CHARMM27 force fields. For CHARMM27, the three 200-ns MD runs were divided into subtrajectories of 20 ns length and the S^2 values calculated from those and subsequently averaged. The calculation of S^2 for Amber99SB-ILDN is explained in the main text. The error bars (black) indicate one standard deviation for the S^2 values obtained for Amber99SB-ILDN.

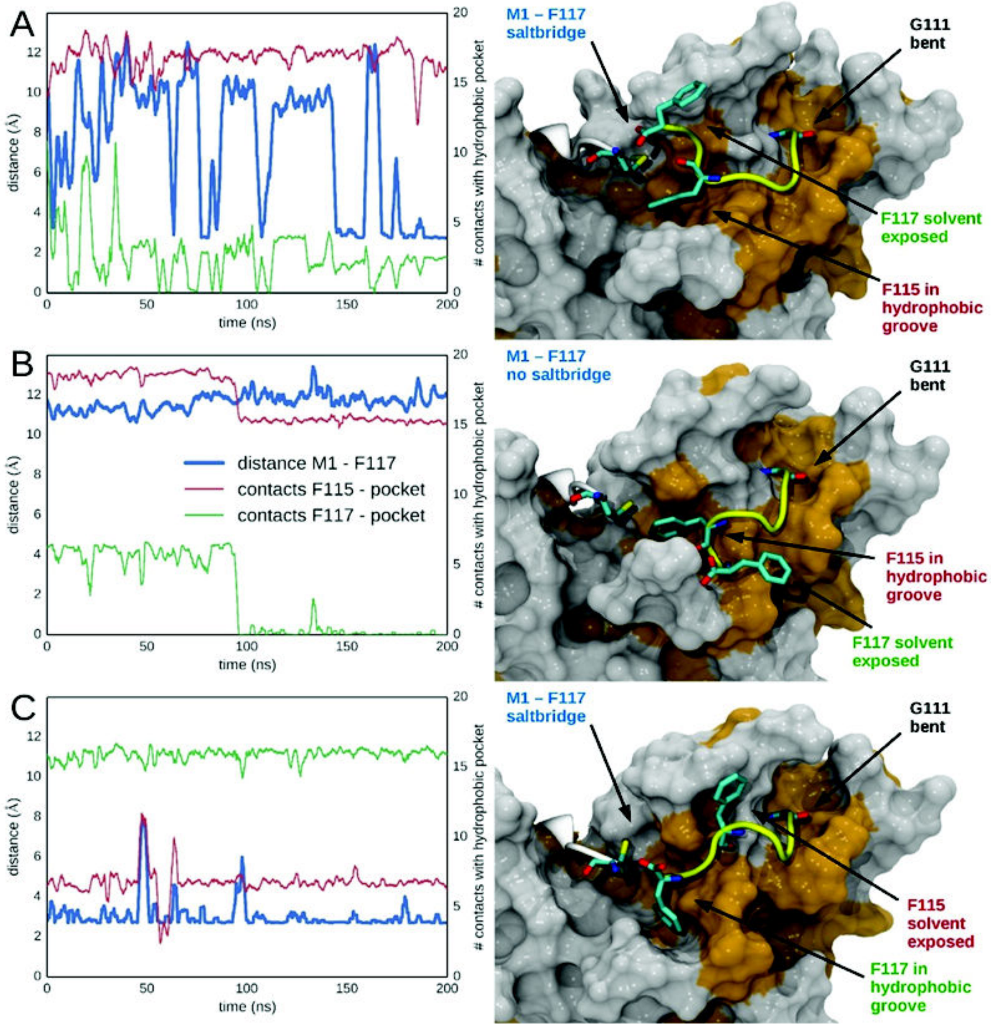


Figure S7. Conformational fluctuations obtained in the course of three 200-ns MD simulations (corresponding to panels A, B and C) of GATE-16 with CHARMM27. The diagrams on the left show the minimum distance between the nitrogen of Met1 and the carboxylate oxygens of Phe117 (blue, left y-axis), as well as the number of contacts that Phe115 (red) and Phe117 (green) make with the hydrophobic groove (right y-axis). The pictures on the right show representative structures of each trajectory. Text colors correspond to the colors in the plots to the left; in the surface representations hydrophobic residues are highlighted in orange; the C-terminus is shown in yellow.

6.3 Appendix C: Supporting Information – Molecular Dynamics Simulations Reveal Key Roles of the Interleukin-6 Alpha Receptor in the Assembly of the Human Interleukin-6 Receptor Complex

Supporting Information

Molecular Dynamics Simulations Reveal Key Roles of the Interleukin-6 Alpha Receptor in the Assembly of the Human Interleukin-6 Receptor Complex

Oliver Schillinger,[†,§] *Vineet Panwalkar,*[†,§] *Birgit Strodel,*[†,¶] *and Andrew J. Dingley*[†,§,*]

[†]ICS-6 (Strukturbiochemie), Forschungszentrum Jülich, 52425 Jülich, Germany

[§]Institut für Physikalische Biologie, Heinrich-Heine-Universität, 40225 Düsseldorf, Germany

[¶]Institut für Theoretische Chemie und Computerchemie, Heinrich-Heine-Universität, 40225 Düsseldorf, Germany

Table S1. Parameters Used in the MD Simulations.

property	**apo hIL-6**	**hIL-6Rα**	**apo vIL-6**
# atoms	34 204	73 683	40 684
# water molecules	10 404	22 410	12 536
# Na ions	41	89	52
# Cl ions	41	89	50
mean volume [nm^3][a]	339.2 ± 1.0	731.7 ± 1.5	404.9 ± 1.1
min. protein distance to its periodic image [nm]	2.21	1.74	2.16

[a] Mean values provided with one standard deviation.

Table S2. Comparison of Root Mean Square Deviations (RMSD) of Chemical Shift Predictors.

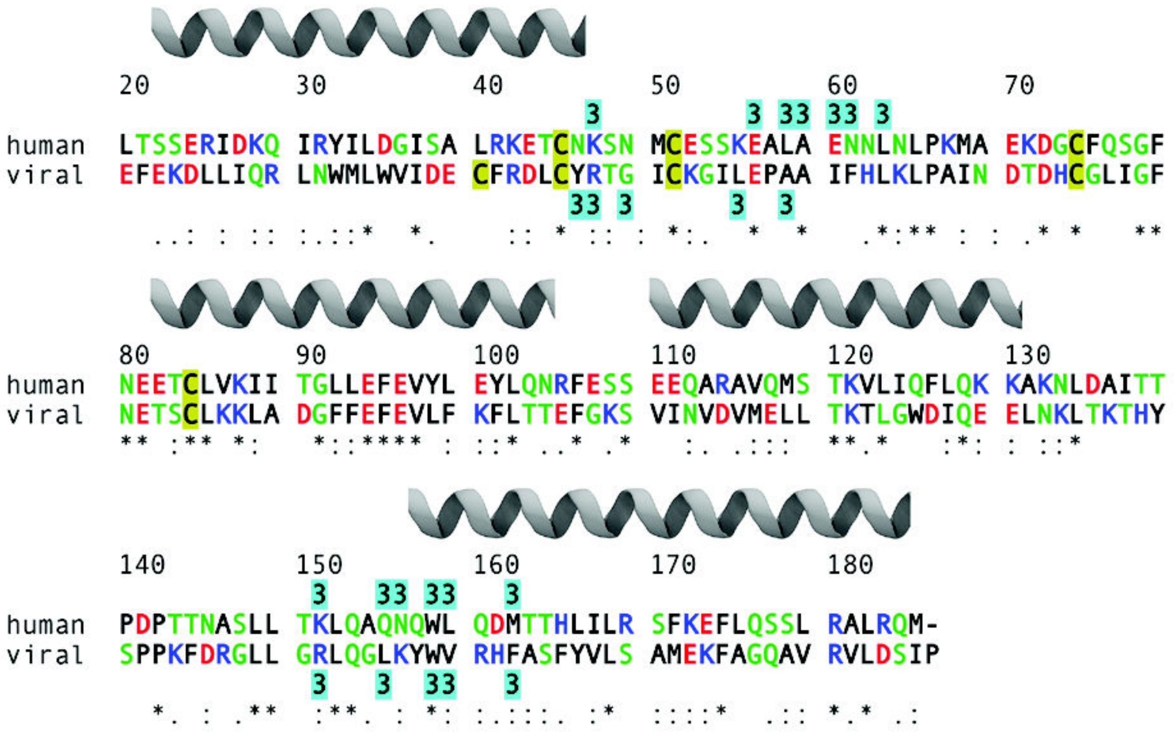
atom type	SHIFTX2 vs. SPARTA+[a]	SHFTX2 vs. hIL-6 exp.[b]	SPARTA+ vs. hIL-6 exp.[c]
Cα	0.37	1.12	1.12
C'	0.32	1.37	1.39
N	0.95	2.22	2.10
HN	0.14	0.39	0.38

[a] RMSD of both predictions for hIL-6 MD simulations.

[b] RMSD of SHIFTX2 predictions to hIL-6 experimental values.

[c] RMSD of SPARTA+ predictions to hIL-6 experimental values.

Conserved (*):	41 residues, 24.7%
Strongly similar (:):	48 residues, 28.9%, cumulative: 53.6%
Weakly similar (.):	21 residues, 12.7%, cumulative: 66.3%



```
       20         30         40         50         60         70
                                    3         3 33 33 3
human  LTSSERIDKQ IRYILDGISA LRKETCNKSN MCESSKEALA ENNLNLPKMA EKDGCFQSGF
viral  EFEKDLLIQR LNWMLWVIDE CFRDLCYRTG ICKGILEPAA IFHLKLPAIN DTDHCGLIGF
                                   33 3      3  3
         ..: : :: :.::*  *.    :: * ::  :*:.  *  *   .*:** :  :.* *   **

       80         90         100        110        120        130
human  NEETCLVKII TGLLEFEVYL EYLQNRFESS EEQARAVQMS TKVLIQFLQK KAKNLDAITT
viral  NETSCLKKLA DGFFEFEVLF KFLTTEFGKS VINVDVMELL TKTLGWDIQE ELNKLTKTHY
       ** :** *:   *::**** : ::* ..* .*   :. .:::  **.*   :*: : ::*

       140        150        160        170        180
                   3   33 33   3
human  PDPTTNASLL TKLQAQNQWL QDMTTHLILR SFKEFLQSSL RALRQM-
viral  SPPKFDRGLL GRLQGLKYWV RHFASFYVLS AMEKFAGQAV RVLDSIP
                   3   3  33   3
         *. : .**  :**. : *: :.:::. :*  ::::*  .:: *.* .:
```

Figure S1. Sequence alignment of human and viral IL-6. The numbering is based on hIL-6 (PDB ID: 1IL6). Helices are indicated above the alignment. Residues of site IIIa are marked with cyan above (human) and below (viral) the alignment. Conserved residues are marked with ‘*’, strongly similar residues with ‘:’, and weakly similar residues with ‘.’. Hydrophobic amino acids are colored black, hydrophilic green, positively charged blue, negatively charged red and cysteines with a yellow background. Sequence alignment was performed by Clustal Omega (1.2.1).

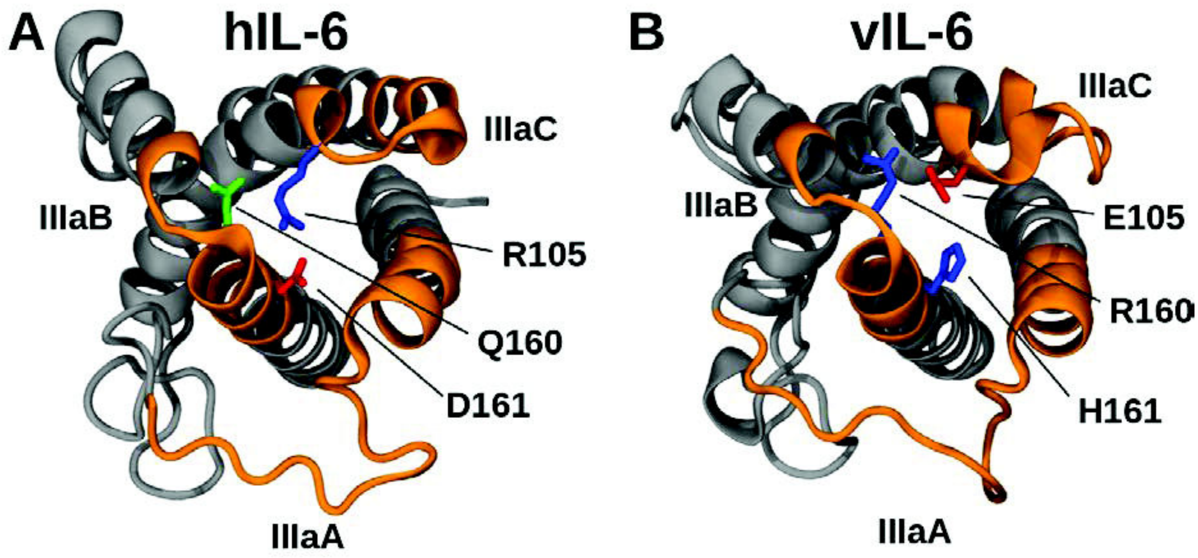


Figure S2. The three IL-6 subsites (orange) of binding site IIIa. The crucial exchange at sequence position 105 of hIL-6 (A) and vIL-6 (B) is shown as a stick model, along with residues that are strongly affected by the amino acid exchange.

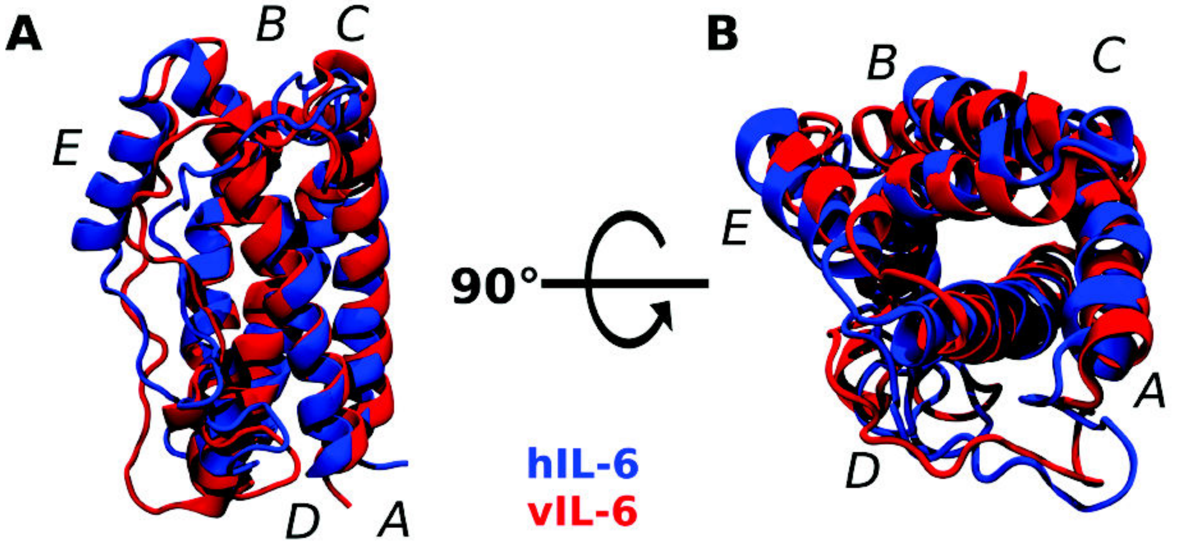


Figure S3. Structural alignment of hIL-6 (blue, PDB ID: 1IL6) and vIL-6 (red, PDB ID: 1I1R). Superposition of the structures was performed by minimizing the backbone RMSD of all common helix residues. The helix annotation corresponds to Figure 3. The backbone RMSD for helices is 1.76 Å and 4.65 Å for the loops. (A) Side view of the superposition and (B) rotation of the molecule by 90° around the horizontal axis.

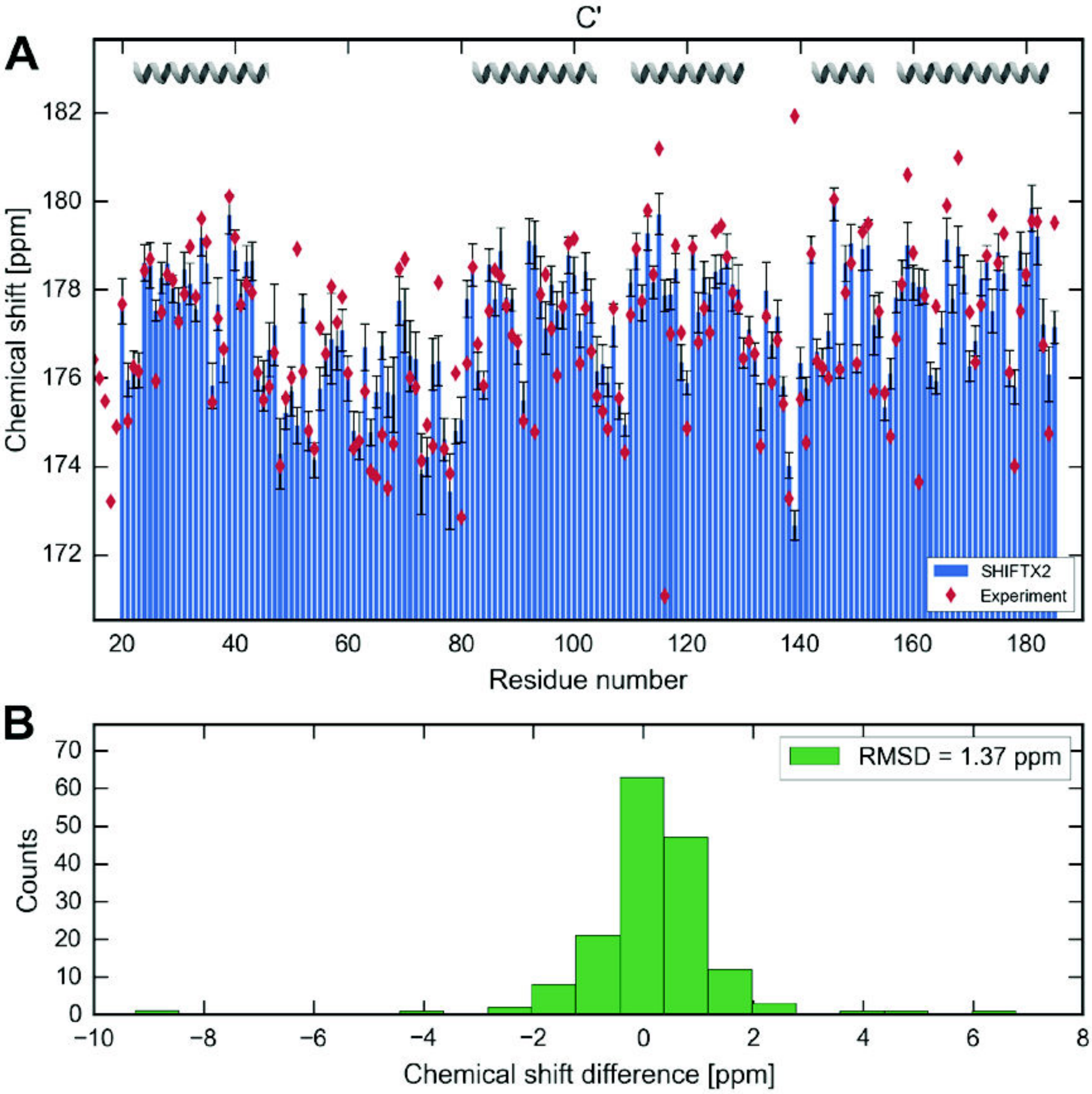


Figure S4. (A) Backbone C' chemical shifts measured by NMR (red diamonds) and predicted with SHIFTX2 using the MD simulations data (blue bars). The error bars represent one standard deviation over all simulation time steps. RMSD denotes the root mean square deviation of predicted and experimental shifts. (B) Deviation of predicted and experimental shifts per residue. Twenty evenly spaced bins between the minimum and maximum values were chosen. The positions of the α-helices are indicated at the top.

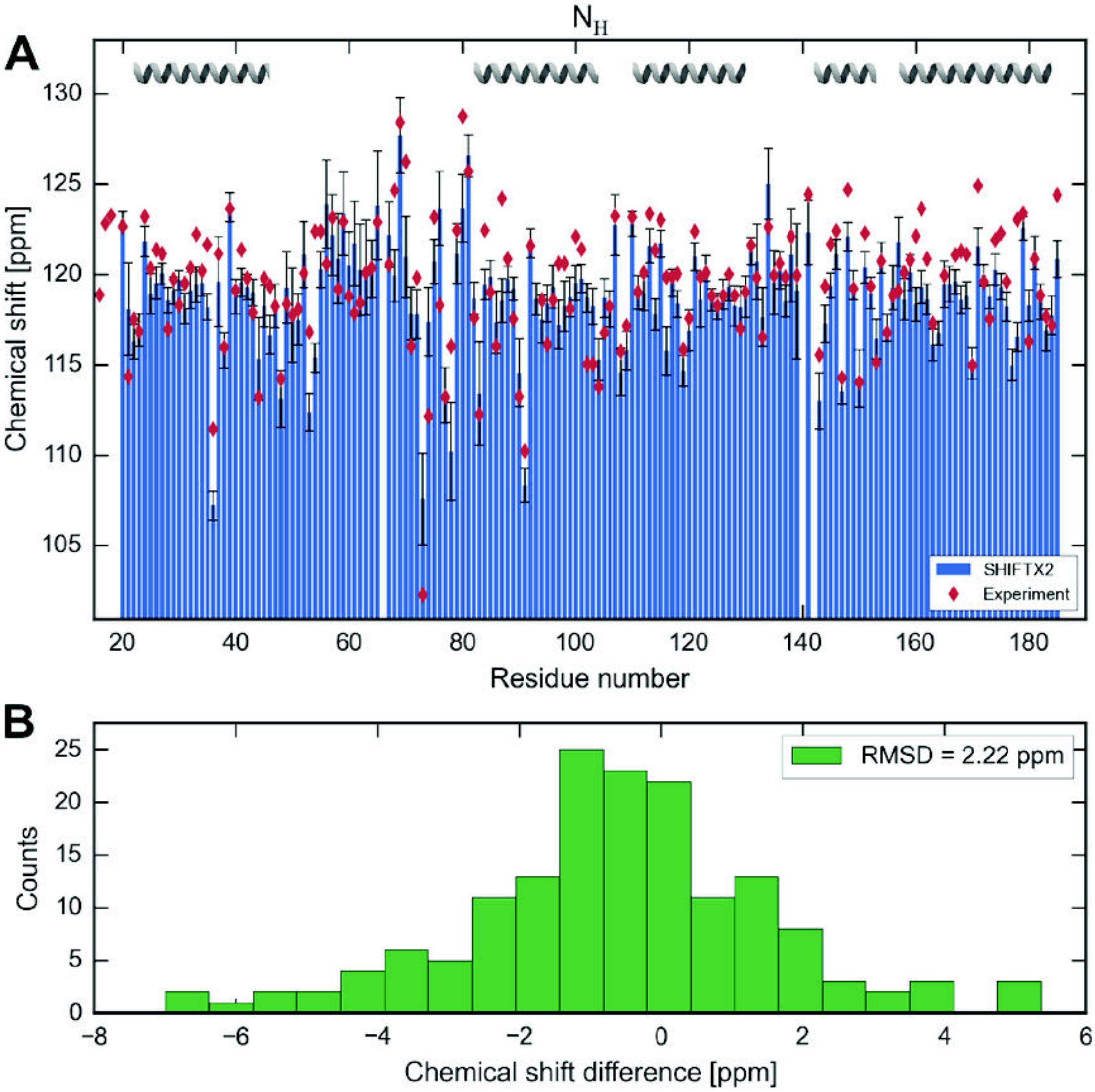


Figure S5. (A) Backbone N chemical shifts measured by NMR (red diamonds) and predicted with SHIFTX2 using the MD simulations data (blue bars). The error bars represent one standard deviation over all simulation time steps. RMSD denotes the root mean square deviation of predicted and experimental shifts. (B) Deviation of predicted and experimental shifts per residue. Twenty evenly spaced bins between the minimum and maximum values were chosen. The positions of the α-helices are indicated at the top.

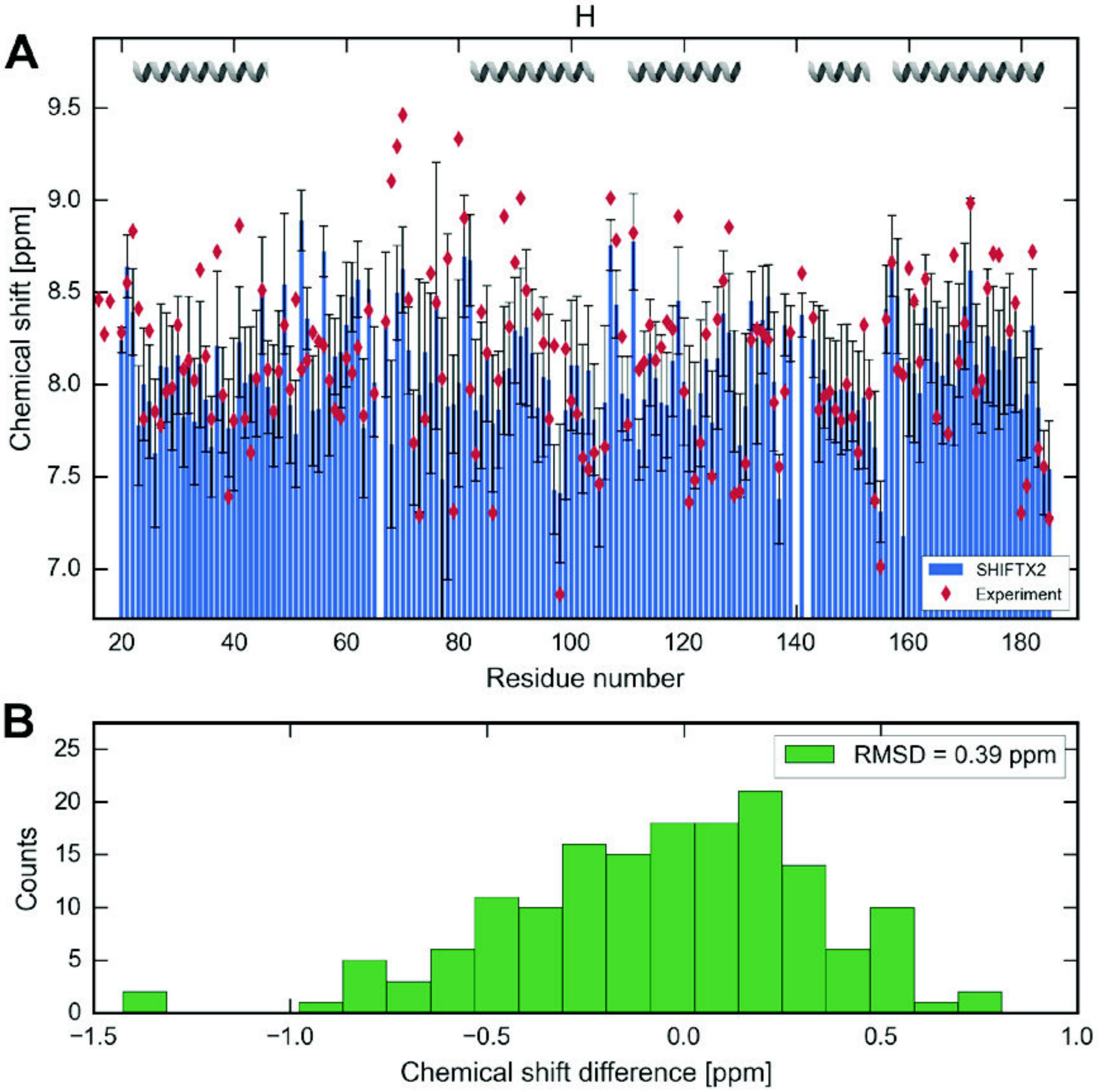


Figure S6. (A) Backbone H_N chemical shifts measured by NMR (red diamonds) and predicted with SHIFTX2 using the MD simulations data (blue bars). The error bars represent one standard deviation over all simulation time steps. RMSD denotes the root mean square deviation of predicted and experimental shifts. (B) Deviation of predicted and experimental shifts per residue. Twenty evenly spaced bins between the minimum and maximum values were chosen. The positions of the α-helices are indicated at the top.

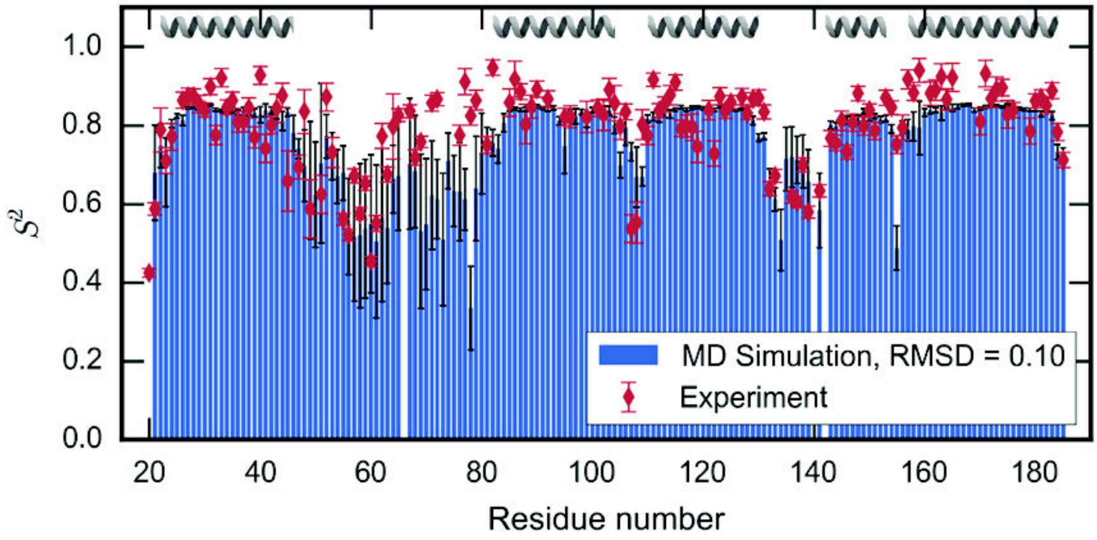


Figure S7. S^2 order parameters of backbone amide groups derived from MD simulations (bars) and ^{15}N NMR relaxation data (symbols). The positions of the α-helices are indicated at the top.

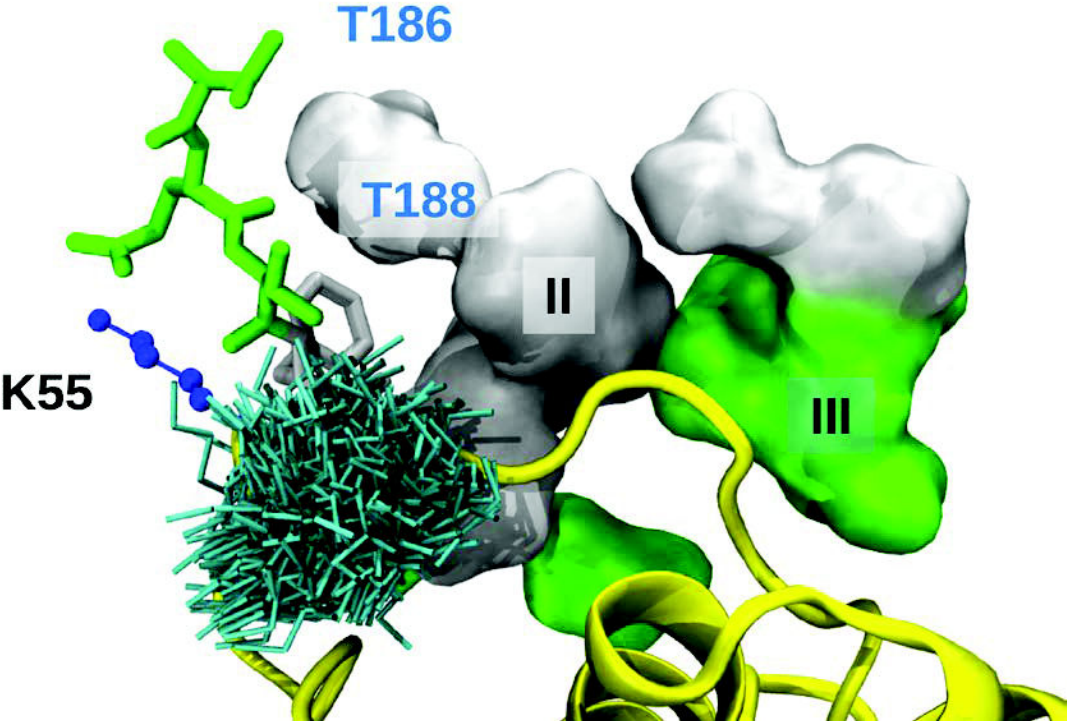


Figure S8. Binding interface I of hIL-6 (yellow cartoon) and the IL-6Rα analogous to **Figure 7**. The predominant position of Lys55 during the IL-6Rα bound MD simulation is shown in blue ball and stick representation, and 100 evenly distributed conformations over the 1 μs simulation of apo hIL-6 in cyan stick representation. IL-6Rα interfaces II and III are indicated, and shown in surface representation.

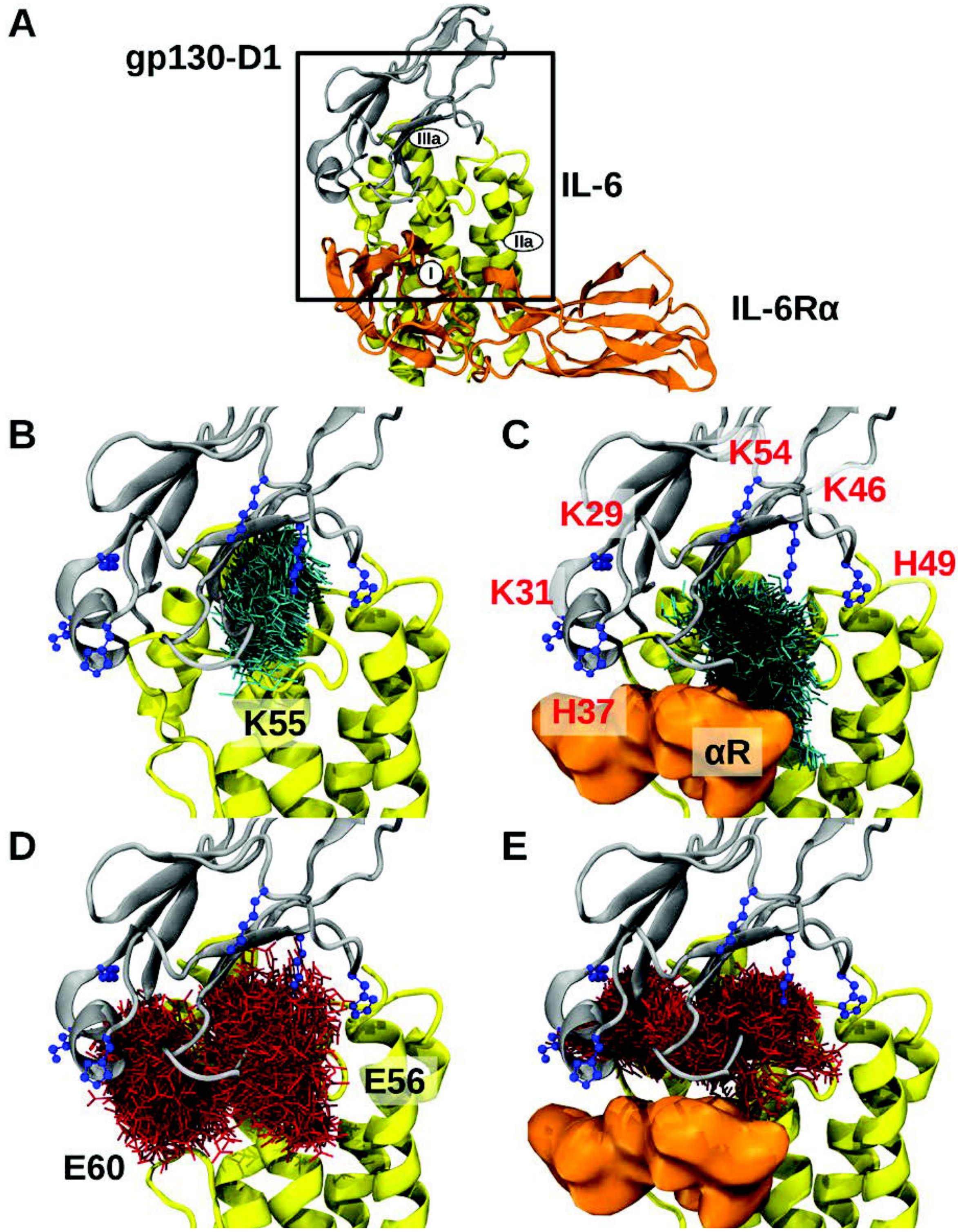


Figure S9. Front view (in accordance with the upper row in **Figure 3**) of IL-6 (yellow cartoon) and the D1 domain of gp130 (gray cartoon). (A) The black square indicates the viewpoint of this figure. Superimposed on the receptor bound crystal structure are distributions of conformations of Lys55 (cyan stick representation, panels B and C), and

Glu56 and Glu60 (red stick representation, panels D and E). Conformations from the apo state MD simulation are on the left (panels B and D) and conformations from the IL-6Rα bound state on the right (panels C and E). The IL-6Rα binding interface is shown in orange surface representation. Positively charged sidechains of the gp130 receptor are shown in blue ball and stick representation, and annotated in red in panel C only.

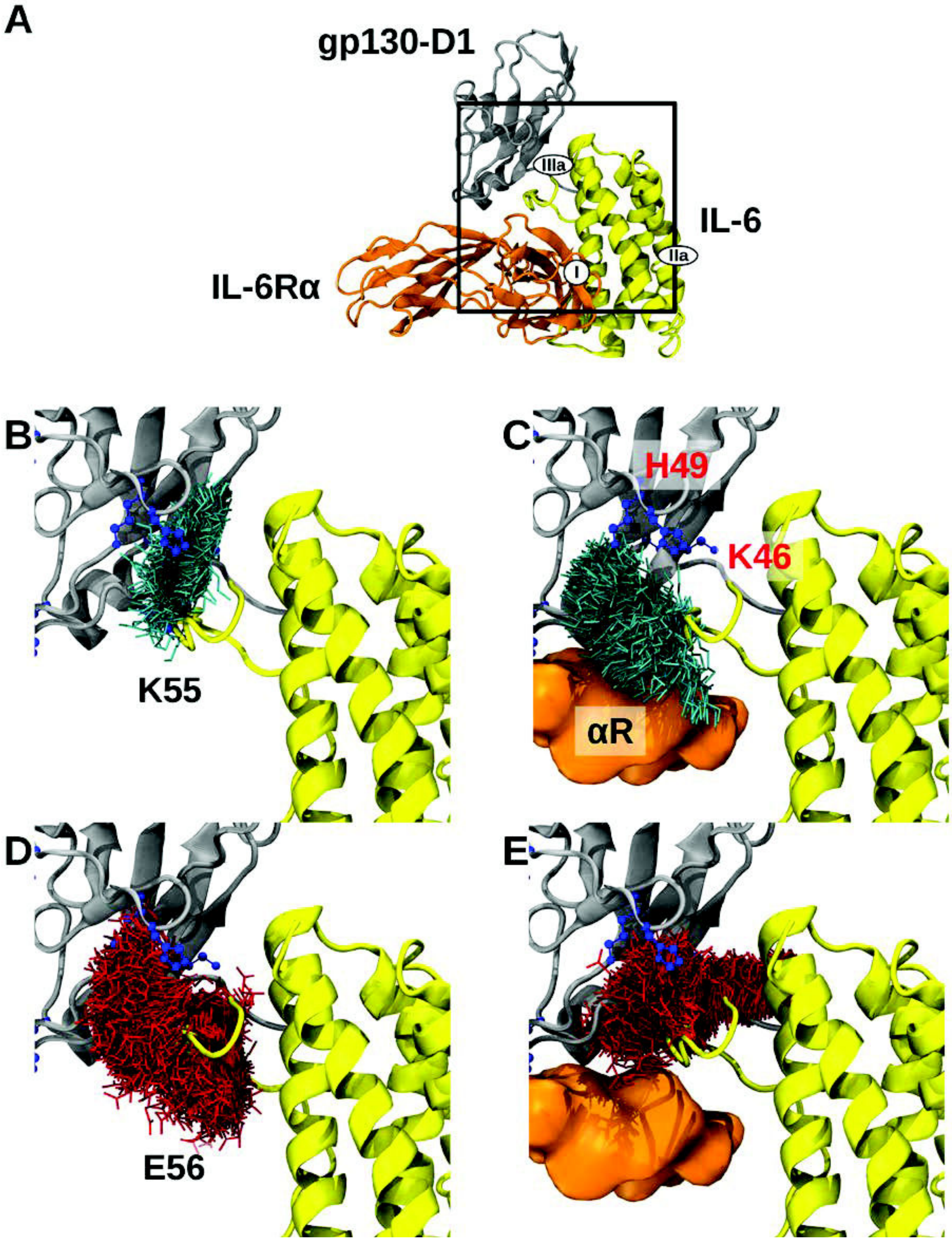


Figure S10. Side view after a 90° rotation around the axis of the four-helix bundle with respect to **Figure S9**. (A) The black square indicates the viewpoint of this figure. Labeling and representations in (B−E) are the same as in **Figure S9**.

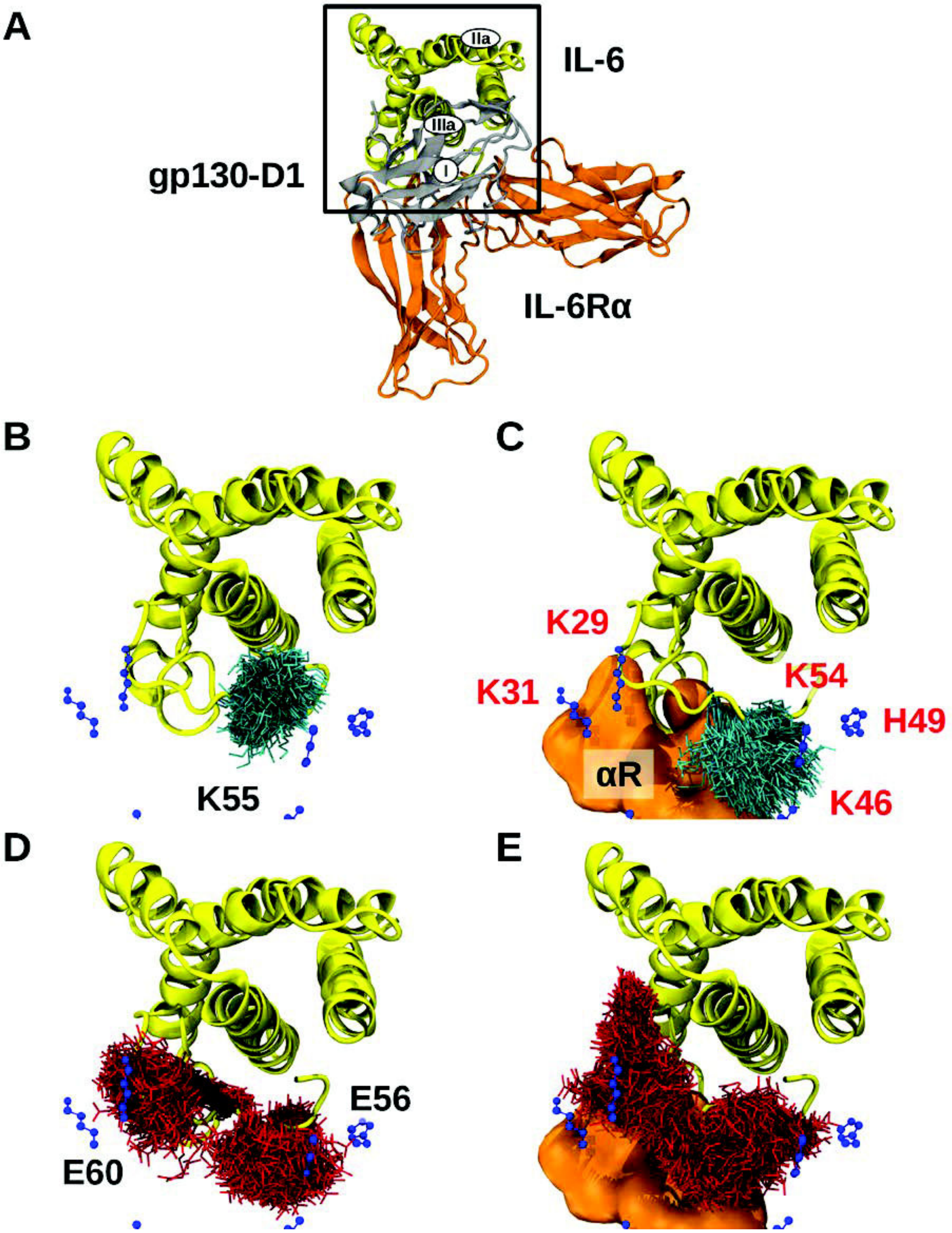


Figure S11. Top view on site IIIa from the perspective of gp130-D1 showing the same residues as **Figure S9** and analogous to the bottom row in **Figure 3**. (A) The black square indicates the viewpoint of this figure. Labeling and representations in (B–E) are the same as in **Figure S9**.

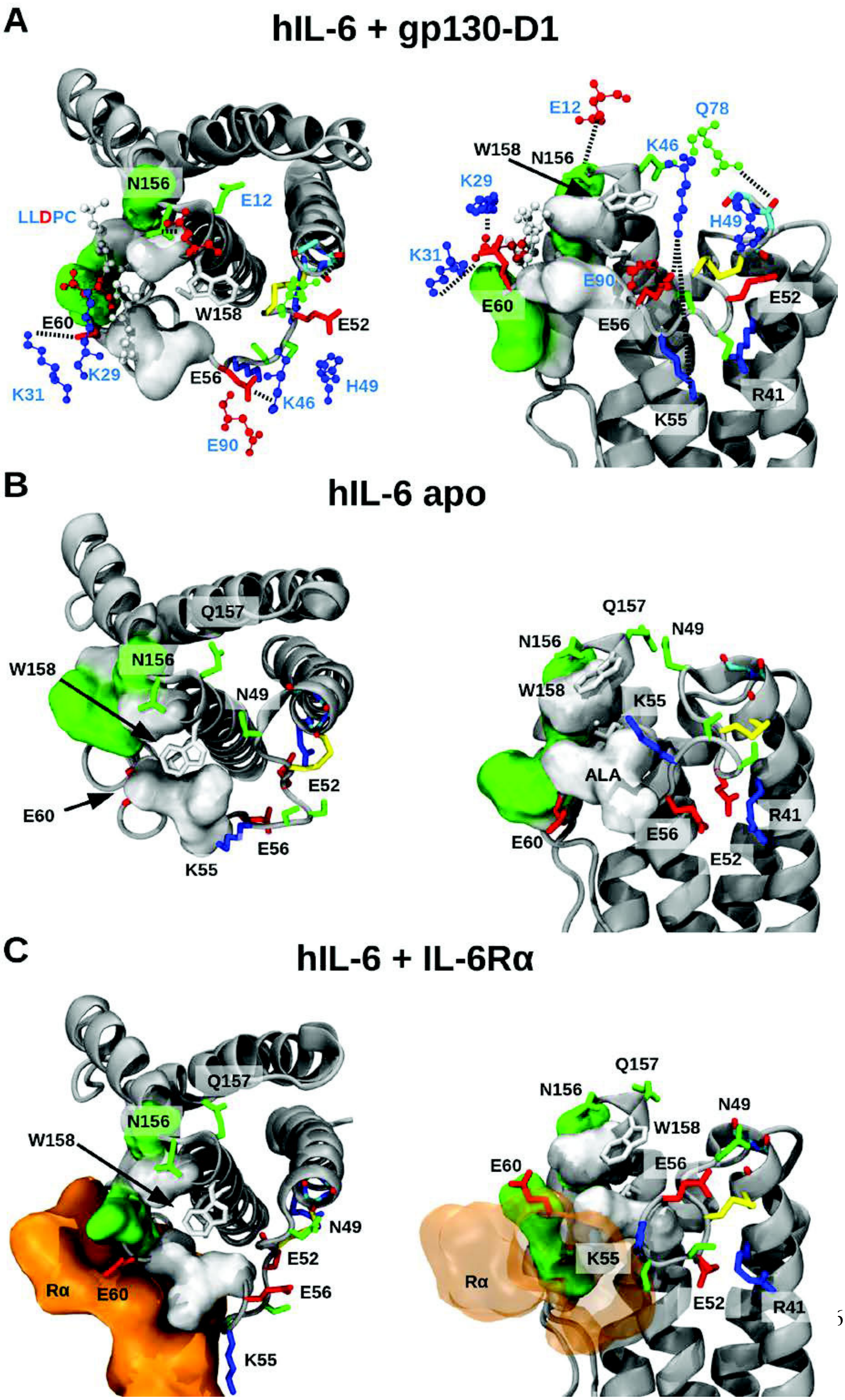

A
hIL-6 + gp130-D1
N156
E12
LLDPC
E60
W158
E52
K29
E56
K31
K46
H49
E90
E12
Q78
W158
N156
K46
K29
H49
K31
E90
E60
E56
E52
R41
K55
B
hIL-6 apo
Q157
N156
W158
N49
E52
E60
E56
K55
Q157
N156
N49
W158
K55
ALA
R41
E56
E60
E52
C
hIL-6 + IL-6Rα
Q157
N156
W158
N49
E52
Rα
E60
E56
K55
Q157
N156
W158
N49
E60
E56
K55
Rα
E52
R41

Figure S12. IL-6 binding interfaces with gp130-D1 for hIL-6. (A) Shows the receptor bound crystal structure, (B) the structure of apo IL-6, and (C) the structure of α-receptor bound. Residues of hIL-6 that interact with receptors are shown in stick representation and annotated with black text, and residues of gp130-D1 are shown in (A) in ball and stick representation and annotated with blue text. Hydrophobic residues and residues involved in binding the gp130 N-terminus are displayed as surface representation. Residue color coding: basic (blue), acidic (red), hydrophilic uncharged (green), and hydrophobic (gray). The conserved cysteine bridge between residues 45 and 51 is shown in yellow and the backbone residues of helix A in IL-6 are color coded by element: carbon (cyan), oxygen (red). The orientations in this figure show site IIIa from the perspective of gp130-D1 (left) and site I from the perspective of IL-6Rα (right), corresponding to the bottom and top panels of **Figure 3**, respectively.

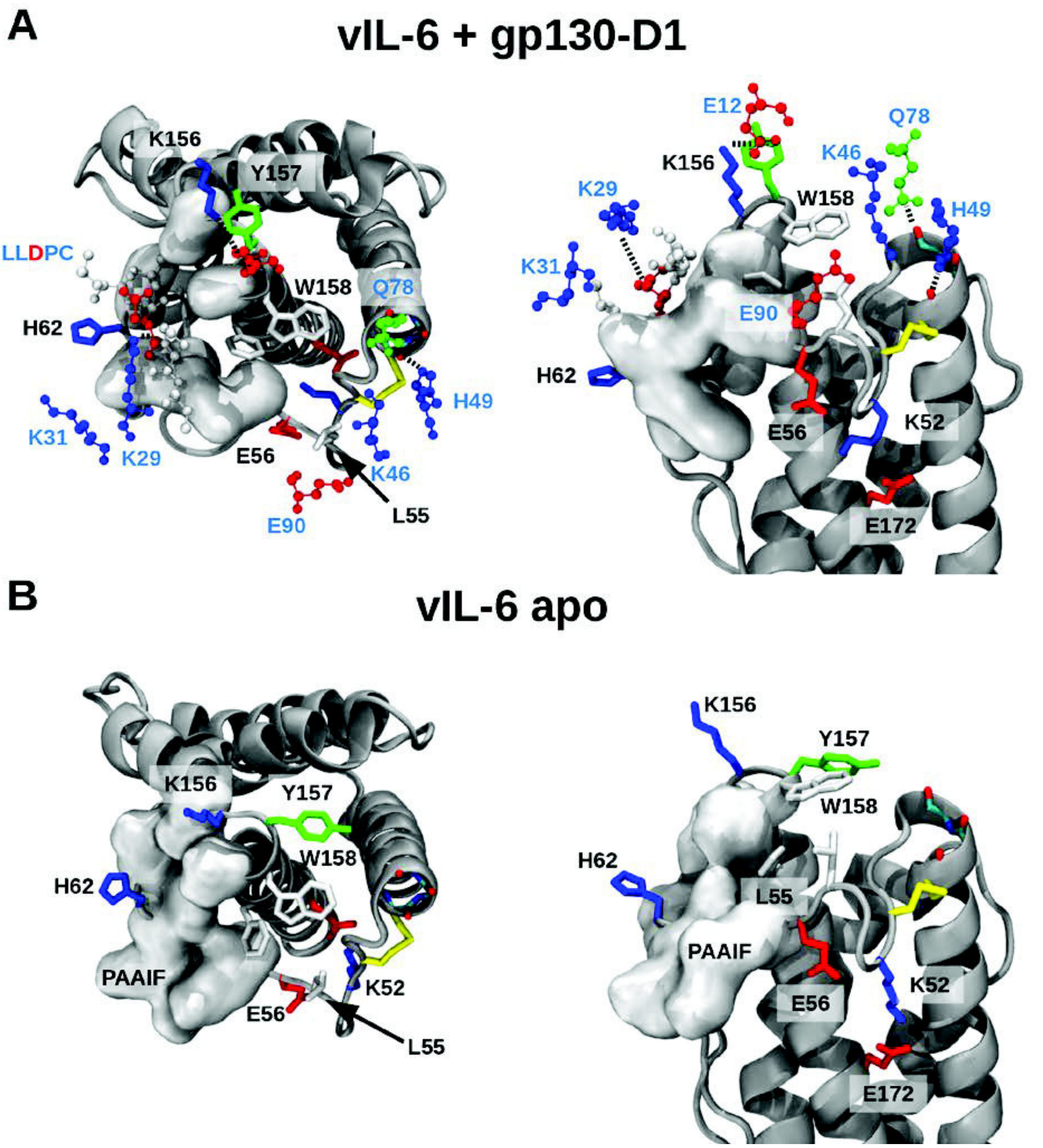


Figure S13. IL-6 binding interfaces with gp130-D1 for vIL-6. (A) Shows the receptor bound crystal structure and (b) the apo state. Color coding corresponds to **Figure S12**.

6.4 Appendix D: Supporting Information – Integrated NMR, Fluorescence and MD Benchmark Study of Protein Mechanics and Hydrodynamics

Supporting Information

Integrated NMR, Fluorescence and MD Benchmark Study of Protein Mechanics and Hydrodynamics

Christina Möller,[1,2] Jakub Kubiak,[3#] Oliver Schillinger,[2,4#] Ralf Kühnemuth,[3] Dennis Della Corte,[2] Gunnar Schröder,[1,2] Dieter Willbold,[1,2] Birgit Strodel,[2,4*] Claus A. M. Seidel,[3*] and Philipp Neudecker[1,2*]

[1]Institut für Physikalische Biologie and BMFZ, Heinrich-Heine-Universität Düsseldorf, Germany

[2]Institute of Complex Systems: Structural Biochemistry, Forschungszentrum Jülich, Germany

[3]Lehrstuhl für Molekulare Physikalische Chemie, Heinrich-Heine-Universität Düsseldorf, Germany

[4]Institut für Theoretische Chemie und Computerchemie, Heinrich-Heine-Universität Düsseldorf, Germany

# Contributed equally.

* Corresponding authors.

Supplementary Information

1. Fitting procedure of NMR translational diffusion data

The NMR translational diffusion data was fitted using MATLAB 2014a, The MathWorks, Natick, 2014. First, the signal intensity of ^{1}H methyl resonances of GABARAP was integrated from 0.27 to 1.05 ppm for every gradient strength and fitted by a nonlinear 2 parameter fit according to

$$I_{methyl}(G_{rel}) = I_{methyl}(0)\, e^{-d_{gabarap}G^2_{rel}} \quad \text{(S1)}$$

in order to determine the reference intensity $I_{methyl}(0)$ and the relative translational diffusion coefficient $d_{gabarap}$ of GABARAP. Afterwards, the integrated signal intensities of the reference molecule dioxane were determined from 3.753 to 3.763 ppm and $I_{dioxane}(0)$, $I_{gabarap}(0)$ as well as the relative diffusion coefficient $d_{dioxane}$ were determined by a nonlinear 3 parameter fit using the following equation:

$$I = I_{dioxane}(0)\, e^{-d_{dioxane}G^2_{rel}} + I_{gabarap}(0)\, e^{-d_{gabarap}G^2_{rel}} \quad \text{(S2)}$$

The diffusion coefficient D is related to the solvent viscosity η and the hydrodynamic radius r_H of the molecule according to the Stokes-Einstein law [1]

$$d \propto D = \frac{k_B T}{6\pi\eta R_h}, \quad \text{(S3)}$$

where k_B is the Boltzmann coefficient. Therefore the hydrodynamic radius of GABARAP is described by

$$R_{h,gabarap} = \frac{d_{dioxane}}{d_{gabarap}} R_{h,dioxane}, \quad \text{(S4)}$$

where $R_{h,dioxane}$ is assumed to be 2.12 Å [2].

The diffusion coefficient D can furthermore be determined from the following equations

$$D_{gabarap} = \frac{d_{dioxane}}{\gamma_H^2 \delta^2 \left(\Delta - \frac{\delta}{3}\right) G^2_{\max}} \quad \text{(S5a)}$$

$$D_{gabarap} = \frac{d_{gabarap}}{\gamma_H^2 \delta^2 (4\Delta - \delta)\pi^2 G^2_{\max}} \quad \text{(S5b)}$$

for the PG-SLED experiment with individual rectangular (S5a) or bipolar (S5b) pulses, where G_{max} is the maximum gradient strength in T/cm, δ is the gradient time, and Δ is the time separation between the leading edges of the two diffusion pulsed gradients.

2. Calculations of crowding parameters

In order to investigate inter-molecular interactions, e.g. crowding, we use convention used by Roosen-Rungea et al (ref. [3]). Concentration of GABARAP and BSA was determined by measuring optical absorption at 280 nm ($\varepsilon_{280nm,GABARAP} = 11920\ M^{-1}cm^{-1}$; $\varepsilon_{280nm,BSA} = 49915\ M^{-1}cm^{-1}$). Protein and dextran concentrations were recalculated into volume fractions using:

$\varphi = \frac{\vartheta * m}{V_{sol} + \vartheta * m}$, where φ is volume fraction, V_{sol} is volume of solvent, m is mass of cosolvent (GABARAP, BSA or dextran), and ϑ is specific volume of cosolvent molecules (inverted molecular density, $\vartheta = 1/d$); ϑ of BSA and GABARAP is 0.735 cm^3/g [4] and ϑ of dextran is 0.625 cm^3/g [5].

Rotational correlation time was recalculated into relative rotational diffusion coefficient $D_{rot} / D_{rot,0} = \rho_{global,0} / \rho_{global}$ in order to make analysis consistent with [6].

Linear regressions were fitted to the individual data sets using

$$D_{rot} / D_{rot,0} = 1 - h * \varphi \ , \qquad \text{(S6)}$$

where h is parameter indicating strength of interaction (decrease of diffusion), (h_{GABARAP} 41.2±1.6, h_{BSA} 13.9±1.0, $h_{\text{Dextran10}}$ 7.5±0.5, $h_{\text{Dextran40}}$ 9.4±0.6. The theoretical relation for hard spheres is $h = 0.41$ to 0.7, depending on size ratio [6].

3. Model of oligomerization

Oligomerization was quantified using simplified dimerization model. We assume that (*i*) only dimers are formed and (*ii*) observed ρ depends linearly on dimer fraction y:

$$\begin{aligned} \rho &= y\rho_{\dim} + (1-y)\rho_{mon} \\ &= \rho_{mon} + (\rho_{\dim} - \rho_{mon})y \end{aligned} \qquad \text{(S7)}$$

With global rotation time of monomeric GABARAP $\rho_{mon} = 9$ ns and assume prolate shape of dimer (2:1), global rotation time for dimer is $\rho_{dim} = 27$ ns.

With the formula:

$$\Rightarrow \ y = 1 + \frac{K_D}{4G_{tot}} - \sqrt{\left(1 + \frac{K_D}{4G_{tot}}\right)^2 - 1} \qquad \text{(S8)}$$

we estimate dissociation constant for assumed dimer formation to be K_D = 3.0±0.3 mM; y(1 mM) = 0.3 (see Figure S9).

4. Translational diffusion measured with FCS

FCS measurements at 22.5°C±0.5°C were performed using a confocal microscope (water-immersion objective Olympus UPlanApo 60x, NA 1.20; pinhole 70 µm) with cw excitation (diode laser Cobolt 06-MLD 488 nm), fluorescence signal was split using a polarizing beam-splitter and recorded by two detectors (Perkin Elmer SPCM-AQRH-14) through band-pass filter (BP 525/39). In order to avoid saturation effects inside the illumination volume, a power series with increasing excitation power density was performed. The excitation power density was estimated by measuring power at the objective and assuming uniform power density within the focal spot of diameter R_{PSF}.

$$R_{PSF} = \omega_{xy} = \left(4 \times t_{diff,Rh110} \times D_{Rh110}\right)^{-2} \qquad \text{(S9)}$$

Cross-correlation curves generated by a hardware correlator, devoid of polarization effects, were fitted using $G_{\mathrm{diff}}(t_{\mathrm{c}})$ $G_{\mathrm{b}}(t_{\mathrm{c}})$ of eq 23. Diffusion times of GABARAP F62C-BFL and Rhodamine 110 (as diffusion standard) for power densities smaller than $2*10^7$ W cm^{-2} were averaged (see Figure S4) and taken for estimation of the translational diffusion coefficient D:

$$D = \frac{\omega_{xy}^2}{4 \times t_{diff}} \qquad \text{(S10)}$$

The diffusion coefficient of Rhodamine 110 at 22.5±0.5°C is $D_{\mathrm{Rh110}} = 4.3 \pm 0.3 * 10^{-10}$ m^{2}s^{-1} [7]. The hydrodynamic radius R_h was calculated assuming an ideal sphere using eq S3 (for translational diffusion) and eq 37 with ρ_{global} from Table 2 (for rotational diffusion).

5. Application of the Aikaike information criterion for the fitting of C(t_c) from MD data

The averaged total correlation functions C(t_c) calculated from the MD trajectories were fitted to multi-exponential decays given in eq 16. However, C(t_c) was not fitted for all bond vectors using three decays. Instead, the number of decays, ranging between one and three, was determined for each bond vector individually and the best-fit model was selected based on the Aikaike information criterion (AIC) [8]. The model with the minimum AIC value ensures a good fit quality with a small weighted residual sum of squares (RSS), while favoring simple models with a small number of parameters, *k*:

$$AIC = 2k + n \times \ln\left(RSS\right) \tag{S9}$$

where *n* is the number of data points. For some residues, however, similar AICs for different models, i.e., different numbers of exponential decays (*m* = 1, 2 or 3) were obtained, making a selection of the best model difficult. Moreover, we realized that the averaging of C(t_c) is affected by rounding errors due to the finite precision of floating point values, which depend on the order the individual correlation functions during the calculation of the average. Even though these errors are small, they propagate through the fitting procedure and give rise to differences in the fitted model parameters, which also complicates the selection of the best model. To circumvent this problem, averaged correlation functions $C_i(t_c)$ were calculated hundred times (i = 1, 2, …,100) for each NH-bond vector while i = 1, 2, …, 20 was chosen for the sidechain bond vectors to reduce the computational effort. The correlation functions from individual subtrajectories were averaged in random order resulting in different rounding errors. Each $C_i(t_c)$ was then fitted with each model *m* = 1, 2 and 3 and a normalized probability $p_{i,m}$ of being the best model was computed:

$$p_{i,m} = \frac{\exp[(AIC_i^{\min} - AIC_{i,m})/2]}{\sum\limits_m \exp[(AIC_i^{\min} - AIC_{i,m})/2]} \tag{S10}$$

where AIC_i^{min} is the minimum AIC value obtained for the three models for a given $C_i(t_c)$. The mean probability for each model was computed by averaging over the hundred correlation functions $C_i(t_c)$, i.e., $\overline{p_m} = p_m / 100$.

For most N–H bond vectors, the standard deviation from this mean is negligible and $\overline{p_m}$ is close to 1 for one of the three models and close to 0 for the other two, providing a clear choice for the best model. Only for few cases the models ended up with mean probabilities significantly different from 0 and 1. In all cases, the best model m was selected as the one with the largest mean probability and S^2 values were then computed as averages over $S_{i,m}^2$ values from the fits of the $C_i(t_c)$ curves, i.e., fits were not considered when a model other than m had the lowest AIC value. The chosen models always had a mean probability of at least 70 % for all but four residues (Lys6: 69%, Glu12: 51%, Arg40: 54%, Asn82: 52%).

6. Information on the software MOP*S*^{2}

The Python program MOP*S*^{2} (Molecular Order Parameters *S*^{2}) was developed in the context of the research presented in this paper. Its purpose is to compute bond vector *S*^{2} order parameters from MD trajectories. The software and detailed installation instructions are available at https://github.com/schilli/MOPS. As the computation of the bond vector correlation functions and the computation of order parameters from the correlation functions can both take time, their computations have been split into separate steps in the workflow.

If the MOP*S*^{2} executable is used (instead of the Python application programming interface for writing tailored Python scripts) the correlation functions can be used with a command similar to:

```
MOPS corr --top <topology file> --trj <list of trajectory
files> --corrpath <directory> --length 10000 --fit backbone
```

The topology file is typically an RCSB PDB file and the trajectories can be in any format supported by the program MDTraj (http://mdtraj.org), such as pdb, xtc, trr, dcd and more. The list of trajectories can be specified with automatic file name expansion, e.g., **`trajectory_*.xtc`**. The **`corrpath`** directory will be used as a storage place for compressed correlation functions and associated information. The **`length`** parameter specifies the length of sub-trajectories in picoseconds in which the input trajectories will be split to compute separate correlation functions. This time should at least be twice as long as the rotational correlation time of the protein. Longer times should not affect the result significantly. The **`fit`** parameter specifies a group of atoms on which a superposition of all frames of the trajectory should be performed to remove the global rotational motion. If omitted, no superposition is performed. The default is that MOPS2 calculates correlation functions for the N–H bond vectors. If one wants to calculate them for other bonds, such as side-chain bonds, one can specify the desired bond by providing the corresponding atom selections via **`--atomsel1 --atomsel2.`** All MDTraj selection strings (as described at http://mdtraj.org/latest/atom_selection.html) are valid and must be entered enclosed in quotation marks if the string contains spaces.

Order parameters can be computed from the correlation functions with several methods implemented into MOPS^2: with the `direct` method described in [Trbovic et al. Proteins (2008) 71: 684-684]; with the method called `mean` corresponding to the convergence value of the internal correlation functions, i.e., the mean of the last part of the internal correlation functions; with `LS` method, which uses the Lipari-Szabo model to calculating multi-exponential fits for the global correlation functions using a fixed number of exponentials for all residues; and with the `LSsel` method, which is similar to the `LS` method but allows to select the optimal number of exponentials during the fitting procedure based on the Akaike information criterion. The `LSsel` method was used in the reported research.

An example of a command line input for the order parameter computation with the `LSsel` method is given below. The other methods require mostly the same but fewer and easier to understand parameters that are described in each of the methods help output, which can be obtained with `MOPS <method> -h`.

```
MOPS LSsel --corrpath <corrpath> --outfile <outfilename> --internal --maxdecays <maxdecays> --nfits <nfits>
```

The path at which the correlation functions were stored with the `corr` method needs to be supplied as the `corrpath` parameter. The output will be written to the specified `outfilename`. The `internal` flag tells the program that the correlation functions were computed after superposition of each frame on the first frame of the trajectory. In the current work, the `internal` flag was not invoked as the correlation functions including global rotation were fitted. The `maxfits` parameter specifies the maximum number of exponentials to be use (usually less than 5, 3 was used in the current work) and `nfits` gives the number of fits with randomly averaged correlation functions to account for different results in borderline cases due to rounding errors, as described in detail in the previous section. The default value is 10; larger values increase the computation time linearly.

The order of the parameters provided when calling any of the methods implemented in MOPS^2 is not important. Detailed usage information can be accessed by calling MOPS^2 with the `-h` flag.

Supplementary Tables

TABLE S1

Parameters of the rotational diffusion tensor from ^{15}N NMR spin relaxation analysis at 25 °C.

Symmetry	Tensor parameters	Autocorrelation times
asymmetric	$D_{zz}=(1.933\pm0.003)\times10^7$ rad/s $D_{yy}=(1.680\pm0.003)\times10^7$ rad/s $D_{xx}=(1.598\pm0.002)\times10^7$ rad/s $D_a = D_{zz}-(D_{yy}-D_{xx})/2$ $=(0.294\pm0.004)\times10^7$ rad/s $D_r = (D_{yy}-D_{xx})/(2D_a)$ $=0.140\pm0.006$ $R=\sqrt{1+3D_r^2}=1.029$ α[a]$=(0.114\pm0.025)$ rad β[a]$=(1.048\pm0.007)$ rad γ[a]$=(0.811\pm0.011)$ rad	$\rho_{-2}=1/(6D_{rot}-2D_aR)$ $=10.19$ ns $\rho_{-1}=1/(4D_{xx}+D_{yy}+D_{zz})=10.00$ ns $\rho_0=1/(D_{xx}+4D_{yy}+D_{zz})=$ 9.76 ns $\rho_{+1}=1/(D_{xx}+D_{yy}+4D_{zz})=$ 9.08 ns $\rho_{+2}=1/(6D_{rot}+2D_aR)$ $=$ 9.07 ns
prolate axially symmetric	$D_{par}=D_{zz}$ $=1.933\times10^7$ rad/s $D_{per}=(D_{yy}-D_{xx})/2=1.639\times10^7$ rad/s $D_a=D_{par}-D_{per}$ $=0.294\times10^7$ rad/s	$\rho_{-1}=1/(6D_{per})$ $=10.17$ ns $\rho_0=1/(5D_{per}+D_{par})$ $=$ 9.87 ns $\rho_{+1}=1/(2D_{per}+4D_{par})=$ 9.08 ns
isotropic	$D_{rot}=(D_{xx}+D_{yy}+D_{zz})/3$ $=(1.737\pm0.001)\times10^7$ rad/s	$\rho_{global}=1/(6D_{rot})=(9.596\pm0.003)$ ns

[a] Euler angles, α, β, γ, are reported relative to the coordinate frame of PDB 1GNU.

TABLE S2

TCSPC fit parameters.

A. Fit parameters of fluorescence anisotropy decay $r(t_c)$ at 20°C modelled as a sum of 3 exponents with free ρ_3 according to eq 27

Residue	r_1	ρ_1 [ns]	r_2	ρ_2 [ns]	r_3	ρ_3 [ns]	$\chi^2_{r,sum}$	$\chi^2_{r,diff}$
V4C	0.14	0.20	0.10	1.99	0.13	9.45	1.0517	1.0297
E7C	0.13	0.23	0.12	1.48	0.12	6.86	0.9949	1.0414
K13C	0.13	0.26	0.13	2.00	0.11	10.88	0.9928	1.0590
I41C	0.20	0.21	0.11	1.23	0.06	9.76	1.0549	1.0670
F62C	0.05	0.23	0.07	2.60	0.25	8.87	1.0277	1.0471
G116C	0.10	0.20	0.09	1.50	0.18	7.36	1.0491	1.0141

B. Fit parameters of fluorescence anisotropy decay $r(t_c)$ at 20°C modelled with sum of 3 exponents, ρ_3 fitted globally according to eq 27.

Residue	r_1	ρ_1 [ns]	r_2	ρ_2 [ns]	r_3	ρ_3 [ns]	$\chi^2_{r,sum}$	$\chi^2_{r,diff}$
V4C	0.13	0.212	0.09	2.63	0.12	9.00	1.0517	1.0297
E7C	0.13	0.290	0.12	2.70	0.08	9.00	0.9943	1.0789
K13C	0.12	0.254	0.11	2.24	0.11	9.00	0.9931	1.0760
I41C	0.18	0.222	0.10	1.44	0.05	9.00	1.0549	1.0674
F62C	0.05	0.270	0.08	4.60	0.24	9.00	1.0279	1.0471
G116C	0.09	0.229	0.10	2.95	0.13	9.00	1.0525	1.0551

C. Fit parameters of fluorescence decay $F(t_c)$ at 20°C fit with sum of 3 exponents (eq 28).

Residue	x_1	τ_1 [ns]	x_2	τ_2 [ns]	x_3	τ_3 [ns]	$\langle\tau\rangle_x$ [ns]
V4C	0.84	5.84	0.11	2.82	0.06	0.31	5.20
E7C	0.89	5.93	0.08	3.08	0.03	0.53	5.53
K13C	0.84	5.89	0.12	3.25	0.04	0.58	5.34
I41C	0.84	5.65	0.12	3.29	0.04	0.57	5.17
F62C	0.91	6.02	0.06	3.23	0.03	0.49	5.69
G116C	0.65	5.59	0.18	2.30	0.17	0.39	3.77

TABLE S3

FCS fit parameters from model function given by eq 31, z_0/ω_0 was optimized in a free fit to the F62C-BFL data and kept fixed (f) for all other fits (Figure S2A). ρ_{global} for F62C-BFL was optimized by manually minimizing χ^2 (Figure S2A).

	F62C-BFL		**I41C-BFL**	
Parameters	**$G_{sp}(t_c)$**	**$G_{ps}(t_c)$**	**$G_{sp}(t_c)$**	**$G_{ps}(t_c)$**
χ^2	2.65	3.45	2.74	2.98
N	2.75		3.40	
t_{diff} [μs]	158		110	
(z_0/ω_0)	11 (fixed)			
b_1	0.087		0.100	
t_{b1} [μs]	41		30	
b_2	0.225		0.233	
t_{b2} [μs]	2.6		2.2	
b_3	0.026		0.013	
t_{b3} [μs]	0.12		0.12(f)	
a	0.929		0.929	
t_a [ns]	6.1		5.0	
b_{rot}	0.0017	0.0238	0.0049	0.0072
ρ_{global} [ns]	9.0(fixed)		9.0(f)	
C	-0.975(f)	-0.975(f)	-0.975(f)	-0.975(f)
S	0.663(f)	0.663(f)	0.663(f)	0.663(f)

TABLE S4

Rotational correlation times at 20.0°C and 25.0°C predicted by HYDROPRO based on the GABARAP structures available in the PDB with the following IDs: 1GNU, 3D32, 1KM7, 1KLV*, 1KOT*. Solvent viscosities were assumed to be 1.002 mPa s at 20.0°C and 0.8903 mPa s at 25.0°C [9].

[*] All conformers of NMR structural ensembles were analyzed individually and are reported as mean ± standard deviation.

PDB IDs:	ρ_{global} [ns] 20.0°C	ρ_{global} [ns] 25.0°C
1GNU	8.86	7.76
1KLV*	9.07±0.34	7.92±0.30
1KOT*	8.49±0.24	7.41±0.21
3D32	8.85	7.74
1KM7	9.19	8.00
average	8.85±0.25	7.73±0.22

Supplementary Figures

Figure S1A

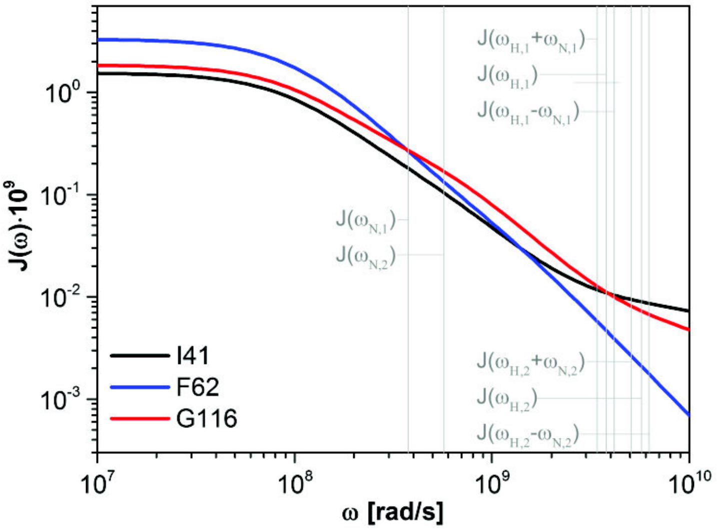


Figure S1B

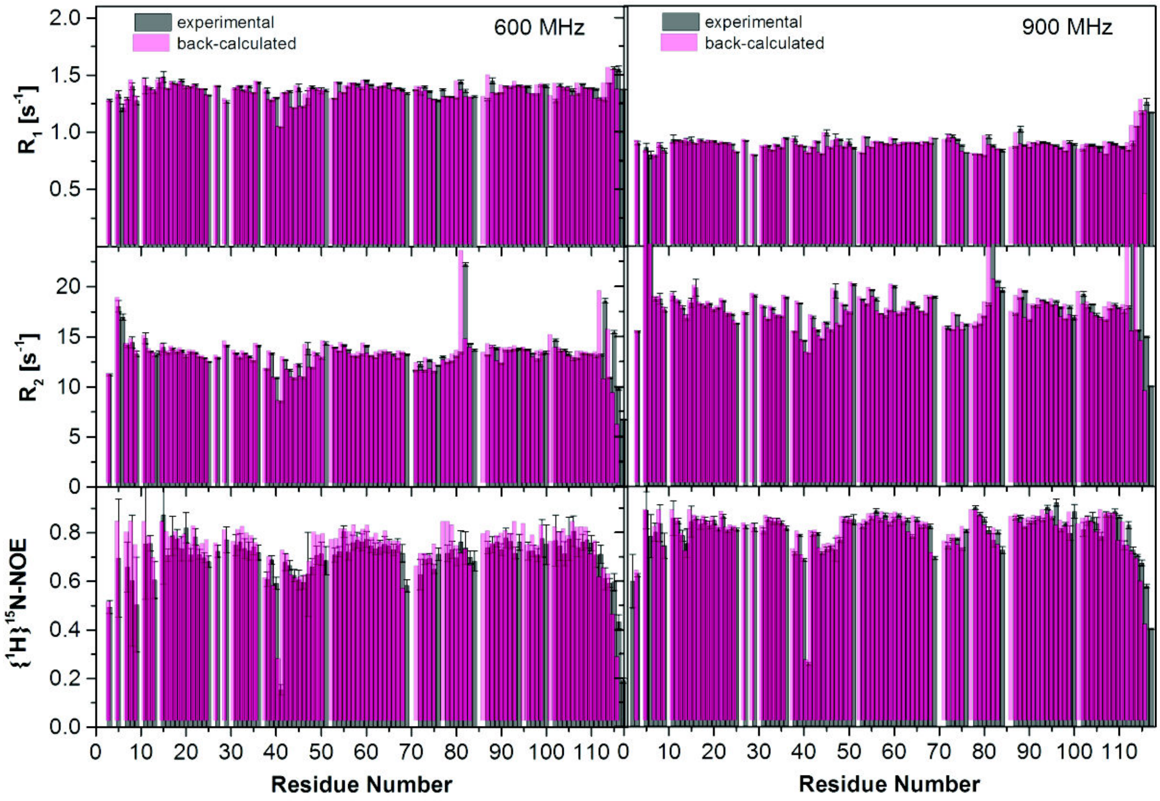


Figure S1C

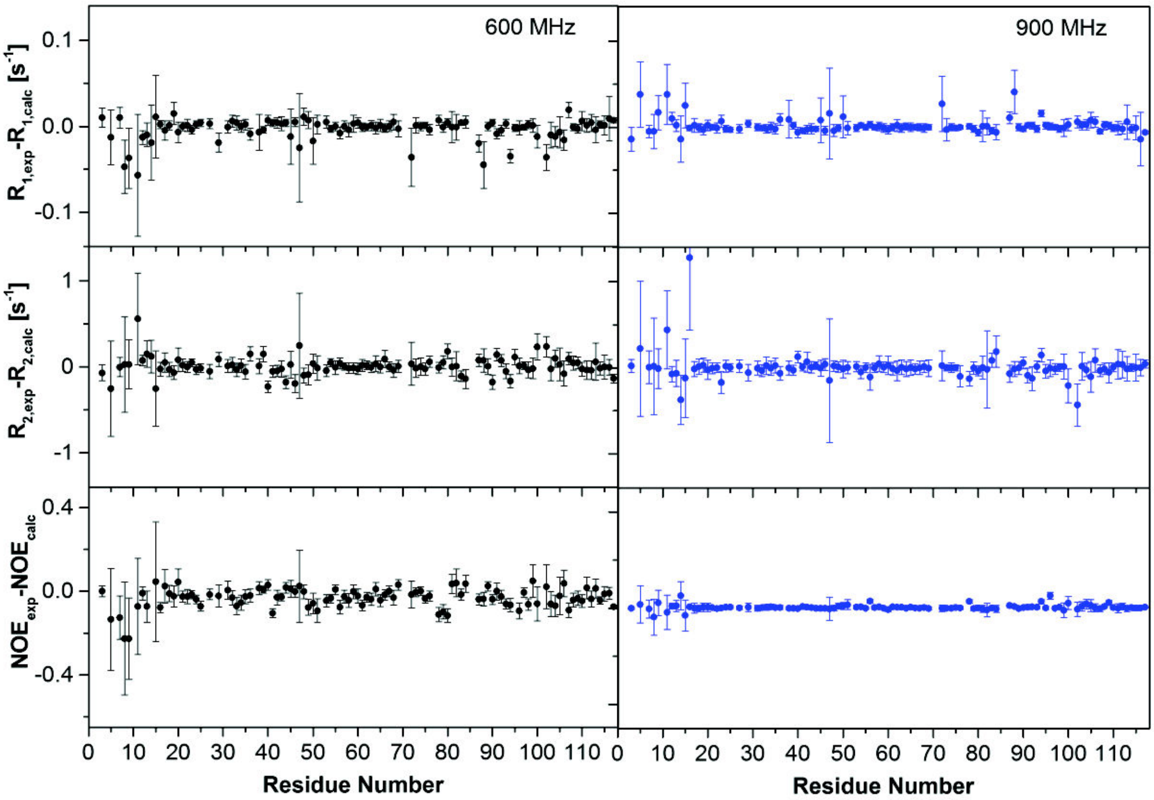


Figure 1: (A) Spectral density functions J(ω) of I41 (black), F62 (blue), and G116 (red) back calculated from the motion parameters obtained by the model-free analysis using eq 8 assuming isotropic diffusion. Grey vertical bars indicate the observable frequencies, where the spectral density was reported. Indices refer to the proton Lamor frequencies of 600 MHz (1) and 900 MHz (2). (B) Back-calculated spin relaxation rates R_1, R_2 and heteronuclear NOE values using from the motion parameters using eqs 9-11 assuming isotropic diffusion. (C) Difference of experimental and back-calculated relaxation data (red bars). The grey error bars indicate the experimental error for each residue number.

Figure S2A

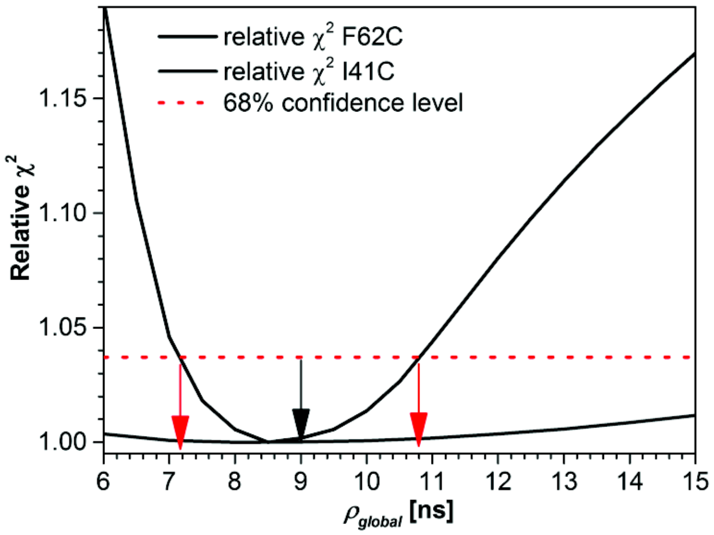


Figure S2B

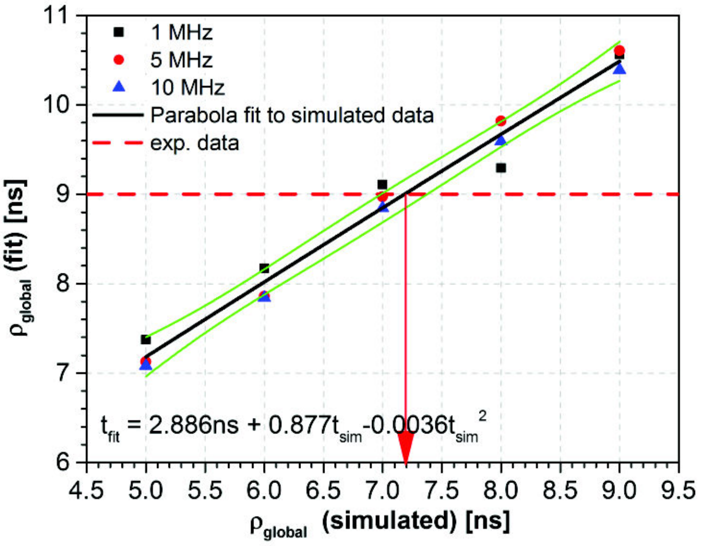


Figure S2: FCS fitting (A) 68% confidence level of t_{rot} for FCS fits for GABARAP F62C-BFL and I41C-BFL. Due to the very small asymmetry of ps and sp correlation no rotational correlation term could be found with sufficient significance for I41C. Rotational correlation times from the FCS fit (see eq 32 and Table S3) was recalculated using an empirical relation between simulated and fitted correlation times. (B) FCS calibration: The difference between t_{rot} and the fitted value arises from the proximity of the anti-bunching term given by fluorescence rate and dye excitation rate. The relation between these values was obtained by a series of simulations with given t_{rot} (5, 6, 7, 8 and 9 ns) and excitation rates (1, 5 and 10

MHz). Parameters of the Monte-Carlo simulation: τ_{e} = 5.9 ns, r_0 = 0.38, l_1 = 0.0308, l_2 = 0.0368. l_1 and l_2 take into account mixing of the polarization in the high NA objective and were obtained experimentally. The excitation rate was varied in the experimentally relevant range between 1 and 10 MHz to check for possible saturation effects (not observed). The simulation was stopped after detection of 10^7 photons. The stored photon train was correlated and fitted using $G_{\mathrm{a}}(t_{\mathrm{c}})$ $G_{\mathrm{rot}}(t_{\mathrm{c}})$ of eq 23.

Figure S3A

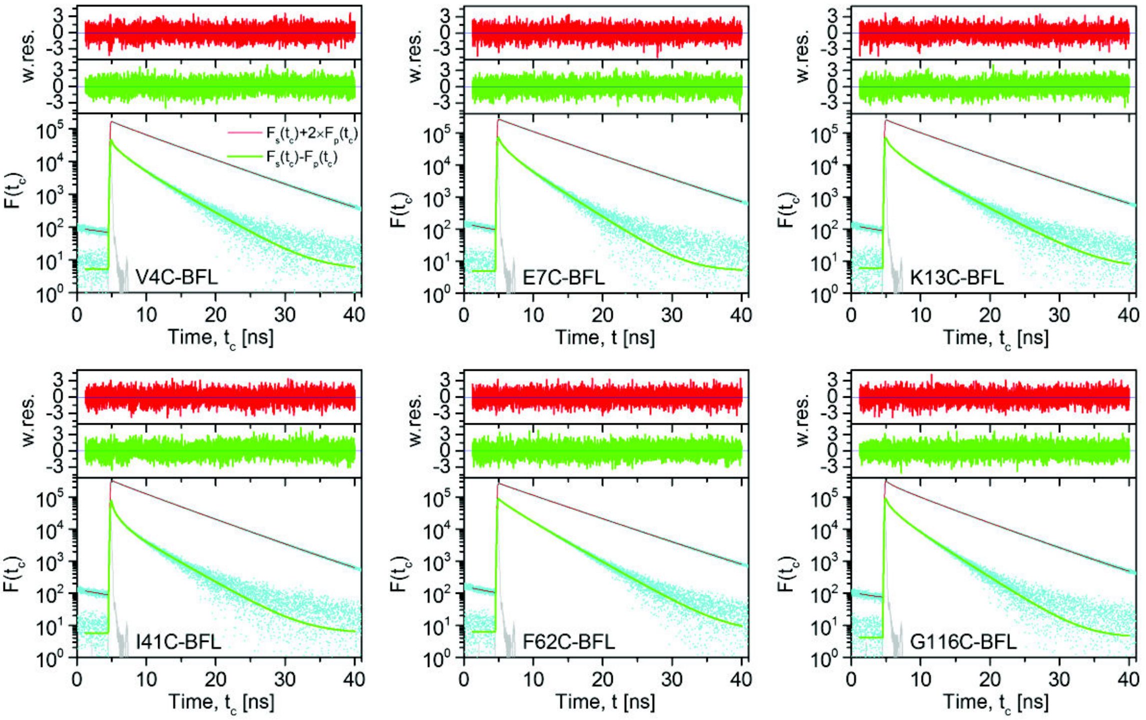

w.res.
3
0
-3
F(t_c)
10^0
10^1
10^2
10^3
10^4
10^5
F_s(t_c)+2×F_p(t_c)
F_s(t_c)-F_p(t_c)
V4C-BFL
E7C-BFL
K13C-BFL
I41C-BFL
F62C-BFL
G116C-BFL
0
10
20
30
40
Time, t_c [ns]
Time, t [ns]

Figure S3B

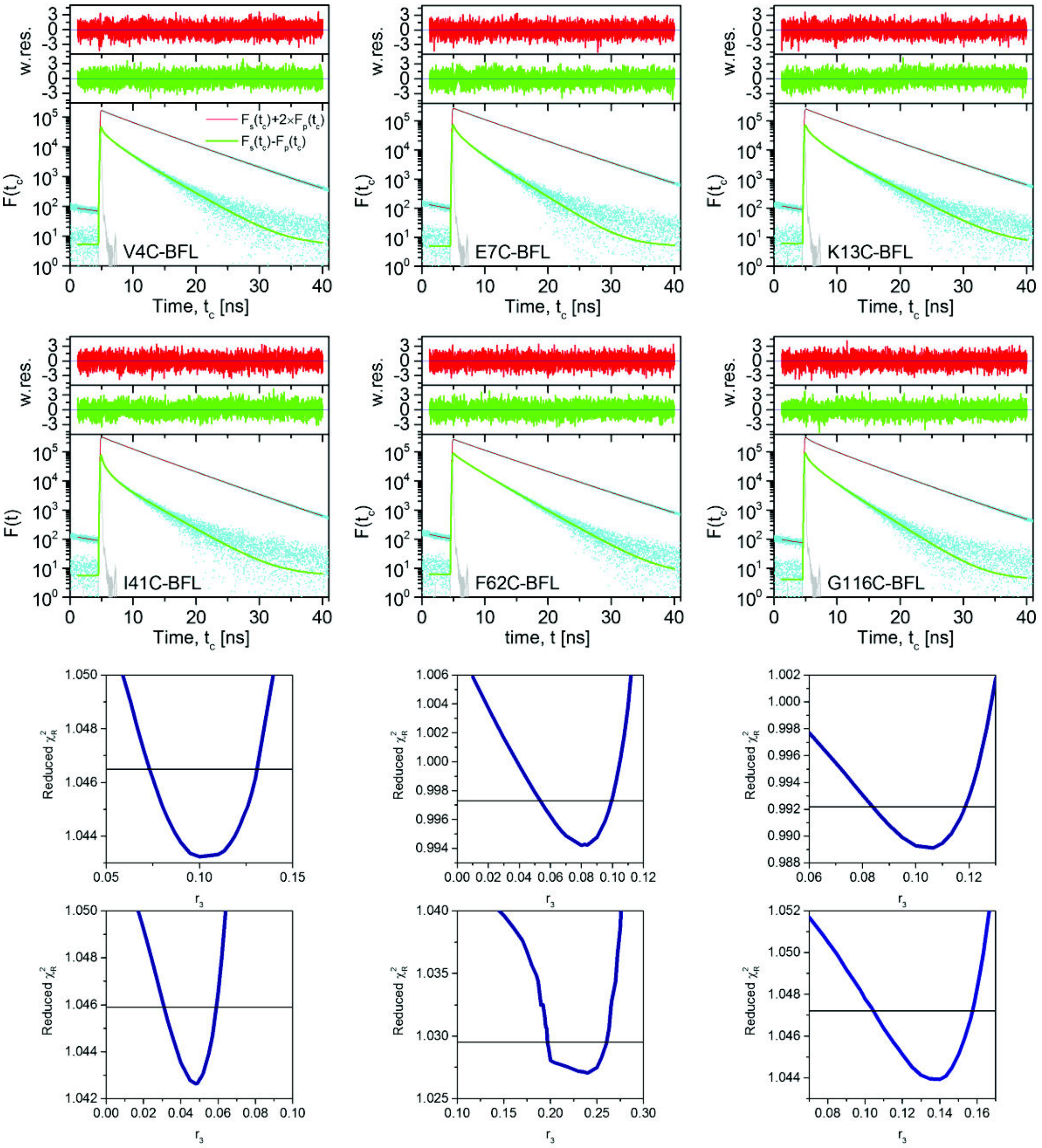


Figure S3: Anisotropy fitting. (A) Individual fits of the polarization-resolved fluorescence decay (sum and difference fit, see eqs 27A and 27B) with 3 exponentials describing the fluorescence decay (eq 28) and 3 exponentials for the anisotropy decay (eqs 29). (B) Global rotation was fitted jointly for all 6 variants in order to obtain comparable amplitudes of ρ_{global} and minimize noise in S^2 estimation. χ_R^2 surface scans for ρ_∞ with marked 68% confidence levels are shown next to the fits. Results of the fits are collected in Tab S2.

Figure S4

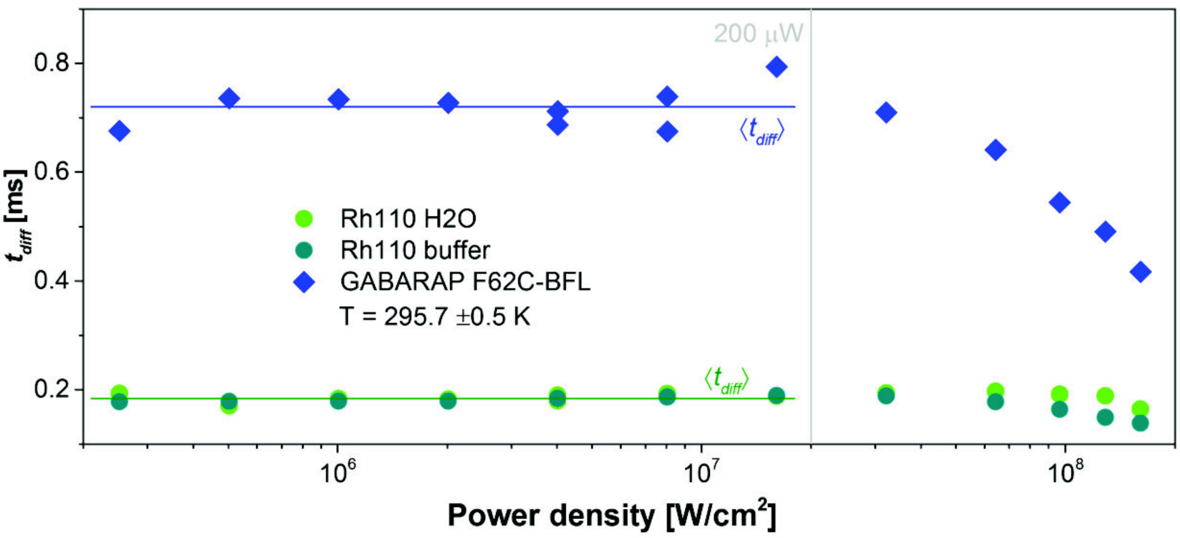


Figure S4: Power series FCS. The diffusion time t_{diff} of GABARAP F62C-BFL at increasing excitation power density was measured to correct for possible saturation effects (decreasing diffusion time). Measurements at excitation powers smaller than 200 µW were used to calculate the average diffusion times (horizontal lines). Rhodamine 110 in H_2O and in buffer (with 0.6 µM GABARAP) was used as reference (diffusion coefficient D_{trans} of Rh110 in H_2O at 295.7K is 4.3±0.3 $10^{-10}m^2s^{-1}$ [7]).

Figure S5

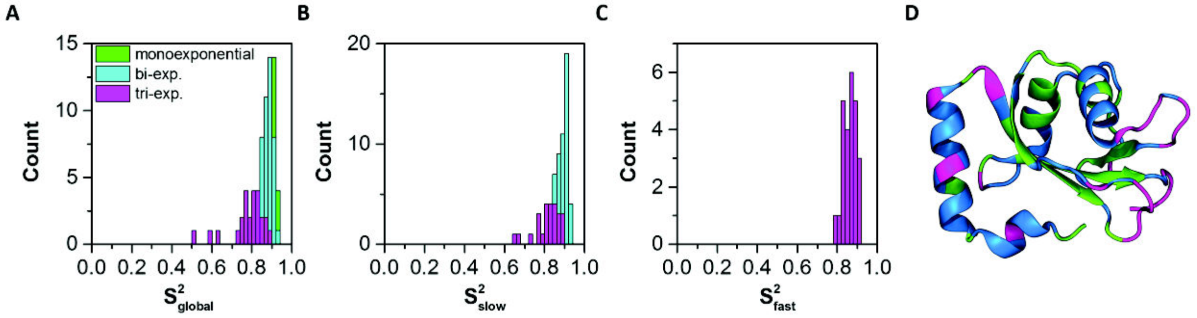


Figure S5: Order parameter histograms of the values shown in Figure 4A. S^2 values were extracted from MD trajectories by fitting the global correlation functions to one (green), two (blue), or three (magenta) exponential decays. The most complex model that did not overfit the bond vector correlation function was selected for each amino acid. This accounts for one global rotational correlation time ρ_{global} and maximum two internal correlation times ρ_1 and ρ_2. The S^2 order parameters represent the amplitudes distributions of the underlying decay processes.

Figure S6

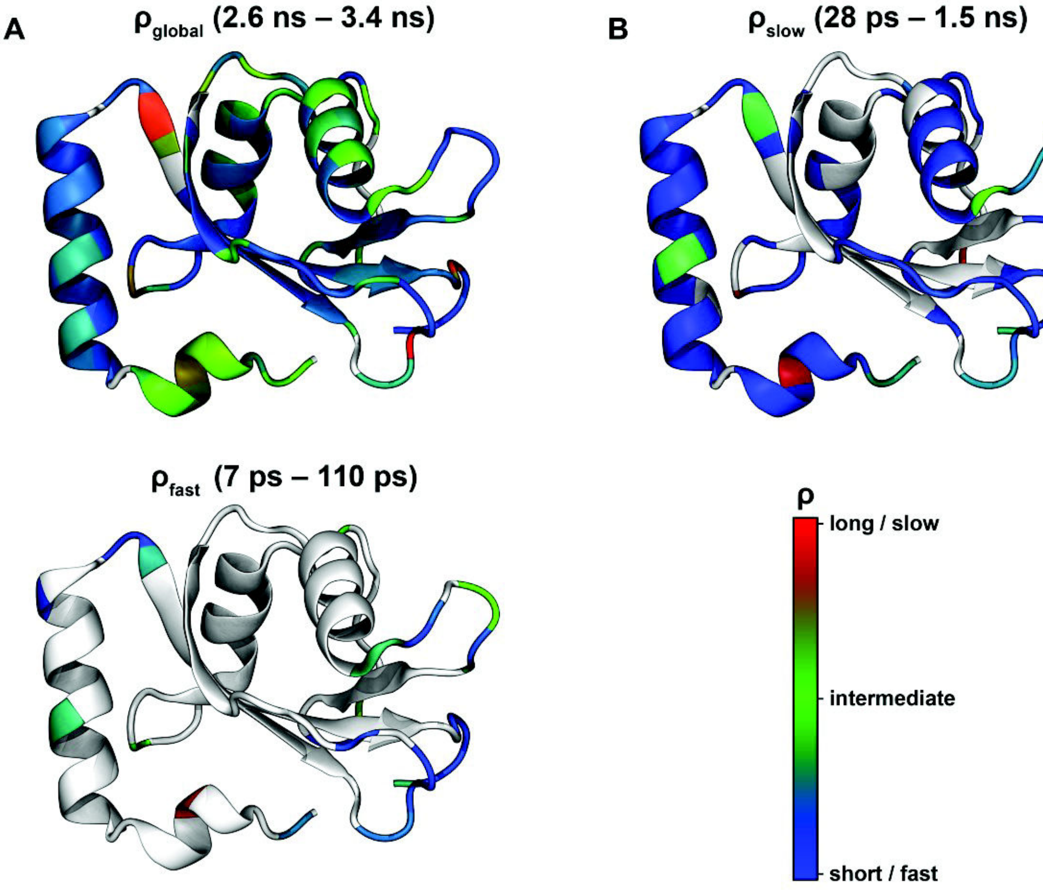


Figure S6: ρ_{global} (A), ρ_{slow} (B) and ρ_{fast} (C) mapped onto the protein structure. The minimum and maximum values of the color scale are set to the minimum and maximum values in each figure respectively.

Figure S7

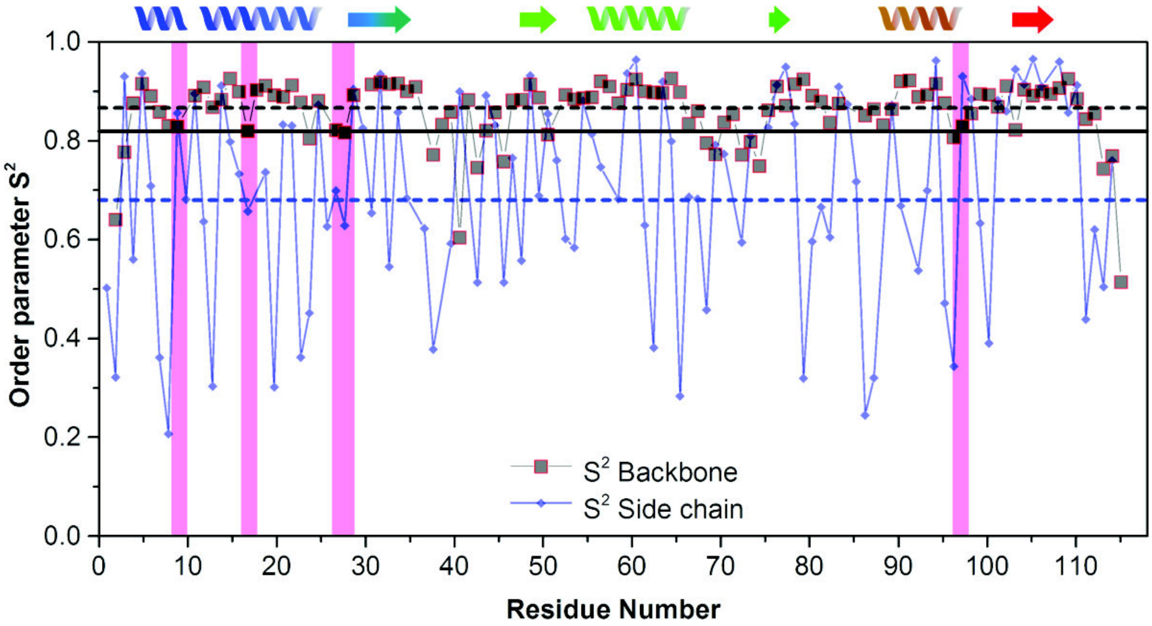


Figure S7: Detection of molecular hinges based on S^2 order parameters obtained from MD for the N–H bond vectors (grey squares) and side chains (blue diamonds). Hinges require a flexible backbone wiFFigth an S^2 value below the average backbone S2 (black dashed line) minus one standard deviation (black solid line) but can only occur in regions outside flexible loops. The shaded areas (pink) indicate the residues (8, 17, (27,28) and 97) that fulfill these requirements. The average S2 value for the side chains is also indicated (blue dashed line). The regular secondary structure elements are indicated above the plot.

Figure S8

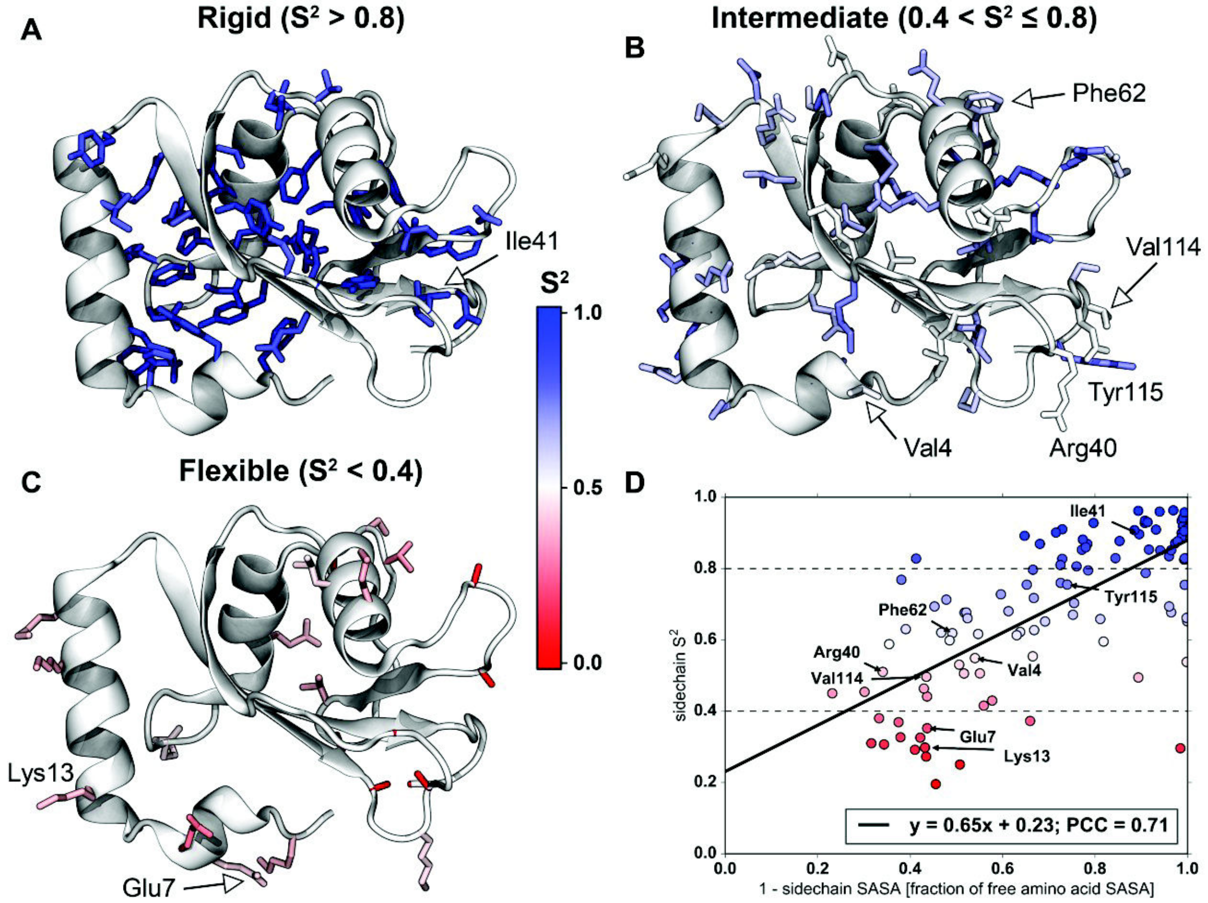


Figure S8: Side chain S2 values mapped on the structure of GABARAP. Panels A to C highlight rigid, intermediate and flexible residues in terms of side chain S2 values respectively. Panel D shows the correspondence of sidechain S2 values with the degree of burial in the protein. The latter is quantified as 1 minus the fraction of free side chain solvent accessible surface area (SASA).

Figure S9

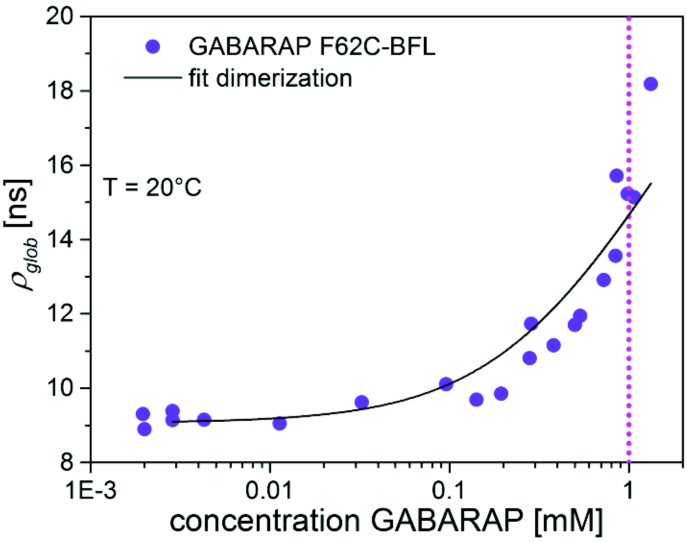


Figure S9: The global rotational correlation time, ρ_{global}, depends on GABARAP concentration. Assuming a simple dimerization model, prolate shape of the dimer (2:1 resulting in $\rho_{\text{dim}} = 27$ ns) and linear dependency of ρ on dimer fraction $\rho = y\rho_{\text{dim}} + (1 - y)\rho_{mon}$, we estimate the dissociation constant to be in the low mM range (K_d = 3.0 ±0.3 mM).

Figure S10

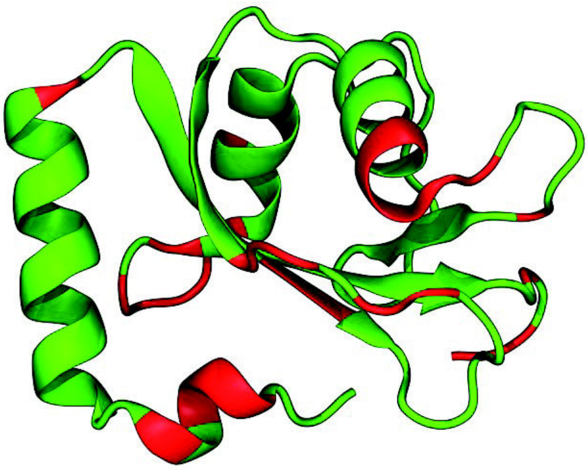


Figure S10: Residues with flexible backbone φ/ψ angles (red) as determined from dihedral angle principle component analysis (PCA). A weighted sum was computed of the eigenvectors with eigenvalues > 0.2. All residues with a maximum dihedral angle vector component larger than 0.05 were assigned as flexible (red), the others as rigid (green).

References

1. Cantor, C.R. and P.R. Schimmel, *Biophysical Chemistry: Part I: The Conformation of Biological Macromolecules*. 1980: W. H. Freeman.
2. Wilkins, D.K., et al., *Hydrodynamic radii of native and denatured proteins measured by pulse field gradient NMR techniques.* Biochemistry, 1999. **38**(0006-2960 (Print)): p. 16424-31.
3. Roosen-Runge, F., et al., *Protein self-diffusion in crowded solutions.* Proc Natl Acad Sci U S A, 2011. **108**(29): p. 11815-20.
4. Fischer, H., I. Polikarpov, and A.F. Craievich, *Average protein density is a molecular-weight-dependent function.* Protein Sci, 2004. **13**(10): p. 2825-8.
5. Banks, D.S. and C. Fradin, *Anomalous diffusion of proteins due to molecular crowding.* Biophys J, 2005. **89**(5): p. 2960-71.
6. Bernado, P., J. Garcia de la Torre, and M. Pons, *Macromolecular crowding in biological systems: hydrodynamics and NMR methods.* J Mol Recognit, 2004. **17**(5): p. 397-407.
7. Gendron, P.O., F. Avaltroni, and K.J. Wilkinson, *Diffusion coefficients of several rhodamine derivatives as determined by pulsed field gradient-nuclear magnetic resonance and fluorescence correlation spectroscopy.* J Fluoresc, 2008. **18**(6): p. 1093-101.
8. Burnham, K.P. and D.R. Anderson, *Multimodel inference understanding AIC and BIC in model selection.* Sociological methods & research, 2004. **33**(2): p. 261-304.
9. Cho, C.H., et al., *Thermal Offset Viscosities of Liquid H2O, D2O, and T2O.* Journal of Physical Chemistry B, 1999. **103**(11): p. 1991-1994.