Kinetics of Disordered Proteins and their Interactions

Thomas David Ferdinand Löhr

Department of Chemistry
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

St Edmunds College

September 2021

Declaration

I hereby declare that this thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the chapter prefaces and specified in the text. It is not substantially the same as any work that has already been submitted before for any degree or other qualification except as declared in the preface and specified in the text. The thesis does not exceed 60,000 words including abstract and footnotes, but excluding table of contents, figure captions, list of figures, diagrams, list of abbreviations, bibliography, appendices, and acknowledgements, as prescribed by the Department of Chemistry degree committee.

Thomas David Ferdinand Löhr
March 2022

Kinetics of Disordered Proteins and their Interactions

Thomas David Ferdinand Löhr

Disordered proteins and regions are highly prevalent in the human proteome, and are often implicated in disease. However, methods to study these systems in detail are lacking, and the potential for thermodynamic and kinetic characterisation using experimental methods is limited. Molecular simulations and associated analysis methods have advanced to the point where investigating disordered proteins and their interactions with other (bio-)molecules on an atomistic scale is now possible. Amyloid-β 42 (Aβ42) is an aggregation-prone biomolecule implicated in Alzheimer's disease, and recent work has shown that small molecules can inhibit the aggregation by dynamically binding to the monomeric form of this disordered protein.

In this work I performed long-timescale simulations of Aβ42 with and without the addition of small molecules, and analysed the kinetics of the system using a neural network and a probabilistic state definition. Without a small molecule, the system occupies several states and transitions occur on the range of microseconds. With the small molecule 10074-G5, the dominant disordered state increases in population, and transitions out of this state become slower. Additionally, the conformational entropy of the protein backbone is increased, with the small molecule forming nanosecond-lifetime π-stacking interactions with aromatic side chains. These findings are consistent with nuclear magnetic resonance experiments, and indicate the possibility of designing molecules with high specificity.

Another approach to targeting aggregation prone proteins such as Aβ42 consists of using specially engineered single-domain antibodies (sdAbs) with a modified complementarity determining region (CDR). These complementarity determining regions (CDRs) are often disordered and their dynamics are poorly understood. I performed enhanced sampling simulations of both an sdAb designed using a sequence-matching method as well as one developed with a structural approach to better understand their conformational space and provide information to improve selectivity and specificity of designed antibodies.

These results show that it is possible to provide a comprehensive characterisation of the kinetics and thermodynamics of disordered proteins in terms of kinetic ensembles, which are defined by the structures and corresponding populations in their different states together with the transition rates between these states.

To my family

Acknowledgments

On a sunny spring day in 2017 I arrived at the Chemistry department to discuss a possible PhD with Michele. After an enthusiastic 30-minute chat, half of which was spent talking about German politics, he encouraged me to apply to the program. I returned home to work on my application, worrying about my not-quite stellar undergraduate grades and the nerve-wracking interview that was surely going to come. When I talked to Michele next, I anxiously asked him about when we were going to have the interview, only to be told: "What do you mean, we already had it!"

Thank you Michele, for making things easier, and for giving me so much freedom and encouraging me to experiment to my heart's content. It is hard for me to imagine a better environment for a PhD.

I wouldn't be here if it weren't for Prof. Carlo Camilloni. Carlo welcomed me into his (at that point) tiny group, and took the time to teach my inexperienced self all about simulations, he sat down with me to pair-program, and always respectfully told me why my ill-thought-out ideas couldn't possibly work. Thank you Carlo for continuing to be a mentor to me after all this time!

I'm incredibly indebted to my colleague and friend Dr. Gabi Heller, who supported me so much throughout my PhD and with whom I hope to collaborate far into the future. Interactions with her have never failed to spark scientific excitement. I would also like to thank her for proofreading parts of this thesis.

I would like to thank Dr. Kai Kohlhoff for our stimulating weekly meetings and all the resources he was able to provide me through Google. I am also grateful to Dr. Max Bonomi, for the many straight-to-the-point discussions about simulations and intricacies of Plumed. I'm also indebted to my frequent collaborator and friend Dr. Faidon Brotzakis, who has been involved in countless exciting projects and ideas.

I am grateful I was able to work with many talented and enthusiastic students over the years: Kate Zator, Concetta Cozza, Eric Wang, Steven Truong, Jan Mičan, Adelie Louet.

I would also like to thank my friends from the CMD and Cambridge: Oded for the countless climbing sessions, coffee breaks, pub visits, and of course the math; the crossword gang, with Prashanth, Oded, Sam, and other guests for keeping ourselves (but especially me) sane during lockdown; *The Breakfast Club*, with Michele, Olli, David, and Kevin for the countless delicious and not-so-delicious brunches; Gia for all the fun in- and outside of Italy; my other office mates from 142a for getting

me addicted to coffee; Swapy for random punting encounters; Marc for inspiring conversations and fun activities in and out of the lab; Michael for exciting last minute projects; Ryan for being a wonderful source of positive energy; Alessia and Ema for a much-needed Italian resurgence in the lab; and my floor-mates from Swirles Court, who were something of a family to come home to in my first year.

Finally, I'm very grateful to my family for gifting me that fateful chemistry set one Christmas and always encouraging me to continue on that path.

Abbreviations

Aβ42	amyloid-β 42
AMM	augmented Markov model
CV	collective variable
CDR	complementarity determining region
CSP	computational structure prediction
CASP	Critical Assessment of protein Structure Prediction
cryo-EM	cryo electron microscopy
DSSP	Dictionary of Protein Secondary Structure
DMD	dynamic mode decomposition
FAST	fluctuation-amplification of specific traits
FRET	Förster resonance energy transfer
haMSM	history-augmented Markov state model
HDX	hydrogen-deuterium exchange
LINCS	Linear Constraint Solver
MFPT	mean first-passage time
MD	molecular dynamics
MDS	multidimensional scaling
MSA	multiple-sequence alignment
NMR	nuclear magnetic resonance
NOE	nuclear Overhauser effect
pre-mRNA	pre-messenger ribonucleic acid
PCCA	Perron cluster-cluster analysis
PCA	principal component analysis
PDB	Protein Data Bank
QM/MM	quantum mechanics / molecular mechanics
RDC	residual dipolar coupling
RMSD	root-mean-square deviation
RNA	ribonucleic acid
SAXS	small angle X-ray scattering
sdAb	single-domain antibody
TICA	time-structure independent component analysis
TIC	time-structure independent coordinate

UMAP	uniform manifold approximation and projection
VAMP	variational approach to Markov processes

Symbols

$\langle \cdot \rangle$	Ensemble average (Equations 1.3 and 1.6).
$\|\cdot\|_F^2$	Frobenius norm of a matrix.
$\mathbb{E}_t[\cdot]$	Expectation value over t.
$\mathcal{O}(\cdot)$	Error order.
$\mathcal{X}^n$	The set of all structures for an n-atom system $\prod_i^n \mathbb{R}^3$ (Equation 1.1).
$\mathbf{x}$	A single structure $\mathbf{x} \in \mathcal{X}^n$.
w_i	Probability of observing a structure i.
E^m_{Struc}	Structural ensemble $\{(\mathbf{x}_1, \ldots, \mathbf{x}_m) \mid \mathbf{x}_i \in \mathcal{X}^n\}$ (Equation 1.2).
E^m_{Thermo}	Thermodynamic ensemble $\left(E^m_{\text{Struc}}, \boldsymbol{w}\right)$ (Equation 1.4).
$E^{(m,k)}_{\text{Kinetic}}$	Kinetic ensemble $\left(E^m_{\text{Struc}}, (\chi_1, \ldots, \chi_m), \mathbf{P}\right)$ (Equation 1.7).
$\mathcal{H}(\mathbf{x}, \mathbf{p})$	Hamiltonian operator (Equation 1.9).
$V(\mathbf{x})$	Molecular dynamics force field potential (Equation 1.8).
$F(\mathbf{x})$	Molecular dynamics forces $-\nabla V(\mathbf{x})$.
$\mathbf{p}$	Momenta of the system.
$\boldsymbol{v}$	Velocities of the system.
$\mathbf{a}$	Accelerations of the system.
Δt	Time step taken in molecular dynamics.
ξ	Collective variable (CV) $\xi : \mathcal{X}^n \to \mathbb{R}$ (see Section 1.3.3).
τ	Lag time in a kinetic model.
$\mathcal{S}$	State space of a Markov model.
s	Individual state of a Markov model $s \in \mathcal{S}$.
χ	Mapping from structure to states $\chi : \mathcal{X}^n \to \mathcal{S}$.
$\mathbf{Z}(\tau)$	Count matrix in a Markov model.
$\mathbf{P}(\tau)$	Transition matrix in a Markov model (Equation 1.16).
$\boldsymbol{\pi}$	Equilibrium distribution of the kinetic model.
λ_i	ith eigenvalue of a matrix (Equation 1.19).
$\mathbf{r}_i$	ith eigenvector of a matrix (Equation 1.19).
t_i	ith implied timescale (Equation 1.20).
$\mathbf{C}(\tau)$	Feature covariance matrix (Equation 1.24).
$\mathcal{K}$	Infinite-dimensional, linear Koopman operator.
$\mathbf{K}(\tau)$	Finite-dimensional Koopman matrix (Equations 1.21 and 1.27).

Contents

List of Figures

1 Introduction

Happy families are all alike;
every unhappy family is unhappy in its own way.
— Leo Tolstoy, in *Anna Karenina*

Proteins are one of the most fundamental structures that make up life. They perform an enormously diverse array of functions, from catalysing biochemical reactions, to allowing cell replication, transporting ions and molecules across membranes, and supporting intra- and extracellular structure. Proteins consist of amino acids, linked by amide bonds (Figure 1.1). There are 21 *proteinogenic*, i.e. naturally occurring, amino acids, and their combinations allow for such a wide range of behaviour and structure. Indeed, the central dogma of structural biology is often stated as follows[1]:

Structure determines function.

Most proteins fold into single, unique structures that allow them to carry out their functions. For example *kinases* will feature a pocket-like structure to allow the transfer of a phosphate group to another biomolecule[2]. *Keratin* is a particularly strong structural protein making up hair, nails, and similar materials - it gains its strength through disulfide bonds and large hydrophobic patches allowing stable polymerization[3]. On the other hand, some proteins feature flexibility as an intrinsic component of their function. One such instance is the chaperone *Hsp33*, which regulates it's specificity through self-binding to a metastable region[4].

In the following sections I will dive deeper into protein disorder, it's purpose and biological implications. I will also talk about how to target disorder to potentially

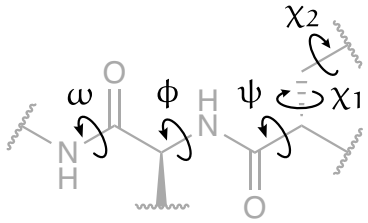


Figure 1.1: The peptide bond with flexible (ϕ, ψ, χ_i) and mostly rigid (ω) dihedrals shown.

treat disease, and methods to probe proteins with high flexibility, with a focus on biomolecular simulation to recover kinetic information.

1.1 Disordered structure?

Disorder in structural biology has been historically neglected and often glanced over. This is despite the fact that approximately one third of the human proteome is disordered in some fashion[5], and the precise function of many of those proteins being poorly understood. Interestingly, the proportion of disordered regions is increased in eukaryotic organisms as opposed to prokaryotes and archaea, and varies as a function of different metrics of complexity[6]. A significant reason for the neglect is certainly the difficulty in characterising these disordered proteins in both their structure and function. Many are postulated to be promiscuous binders and involved in many subtle interactions[7]. At the same time, their disordered nature makes structural determination difficult, as many experimental techniques will result in averaging over long timescales or over the full ensemble of structures, thus obscuring less-populated states or processes occurring on short timescales. Single-molecule techniques can allow more direct observation of conformations with low populations, but are often limited by their requirement for some form of label[8] or non-equilibrium conditions[9]. I will touch on common techniques to characterise both folded and disordered proteins later in this chapter, but first I will give some biological motivation for studying these systems.

1.1.1 Function and dysfunction of protein disorder

One of the more commonly encountered examples of disorder in the human proteome is as a flexible segment in a multi-domain protein complex. This can allow two or more globular, folded protein domains to associate to temporarily form a larger unit with some specific activity. This tether thus ensures that this process can happen quickly, as the protein doesn't have to wait to associate with a completely separate protein. One example of a system like this is the U2AF splicing factor, part of the human spliceosome assembly responsible for editing pre-messenger ribonucleic acid (pre-mRNA)[10]. Upon encountering sequence variations in the ribonucleic acid (RNA), the two RNA recognition motifs can undergo domain rearrangements to modulate splicing activity[11]. The strength of this process is possible due to precise evolutionary tuning of the disordered linker joining the two motifs. The authors of the study were able to elucidate this mechanism using nuclear magnetic resonance (NMR) spectroscopy.

An instance of extreme disorder can be seen in the interaction of the two disordered proteins histone H1 and prothymosin-α[12], both involved in key cellular functions. Schuler and co-workers combined Förster resonance energy transfer (FRET) with coarse-grained molecular dynamics simulations to reveal that both proteins remain disordered when bound, and yet display picomolar affinity to each other. This is possible due to their highly charged nature allowing strong electrostatic interactions without a sacrifice of conformational entropy upon binding.

Another common interaction type is known as *folding-upon-binding*. Here, a disordered protein transitions into a structured form when binding to a folded binding partner. This mechanism is exemplified in the measles virus between a disordered motif on the C-terminal domain of the nucleoprotein and the phosphoprotein X domain, recently studied using long-timescale molecular dynamics simulations[13]. In the transition state, only a few contacts are required to stabilise the intermediate, which remains mostly disordered, and eventually leads to a fully folded complex. Notably, if the disordered protein temporarily formed α-helices before binding, then these would often unfold again before completing the association process.

Additionally, disordered proteins have been shown to be drivers of *liquid-liquid phase separation* in multiple scenarios, allowing the resulting compartmentalization to regulate and isolate cellular processes[14, 15]. In particular, a droplet state can be stabilised by enthalpically-favourable non-specific side chain interactions and corre-

spondingly high entropy, whereas an imbalance could lead to the formation of the significantly more stable and dysfunctional amyloid (i.e. fibrillar) state. These droplet states are thought to occur in a significant fraction of the human proteome[16].

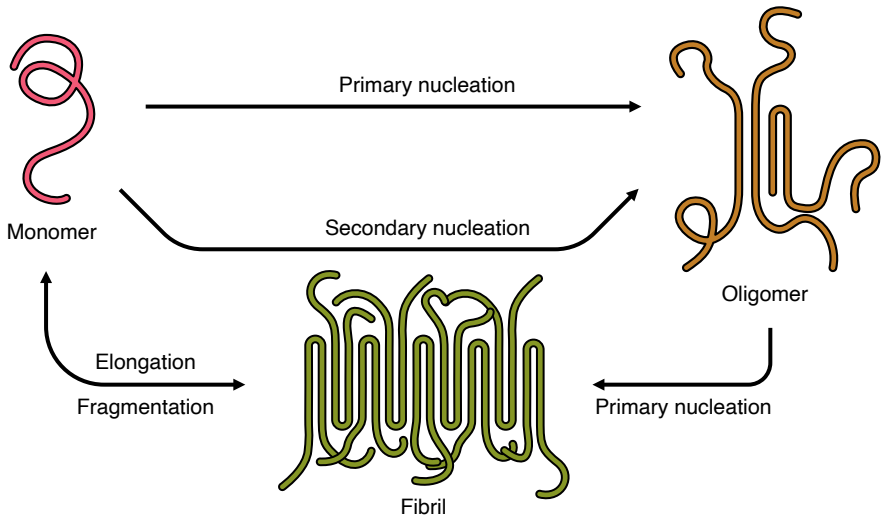


Figure 1.2: Schematic of the aggregation process of amyloid-β. The aggregation proceeds via numerous pathways and intermediates: Primary nucleation is the formation of oligomeric and fibrillar species directly from the monomeric protein, with a critical number of fibrils, they can catalyse the further formation of oligomers through secondary nucleation. The fibrils themselves are in an equilibrium of fragmentation and elongation processes with the monomer.

Until now we have mostly talked about the function of disorder, but it can also be prominently involved in disease. Many neurodegenerative illnesses feature the abnormal formation of fibrils and plaques in the brain through the aggregation of disordered proteins. A prime example of this process is the formation of oligomers and fibrils by the disordered protein amyloid-β, associated with Alzheimer's disease[17, 18]. This protein is the result of a cleavage process taking part on the cell membrane, resulting in peptides of lengths between 36 and 43 amino acids with different aggregation behaviours. The various aggregation intermediates form a network of complex processes such as primary and secondary nucleation, elongation, and fragmentation (Figure 1.2)[19]. The kinetics of these processes – in the case of amyloid-β at least – have been successfully modelled, so that the rate of the com-

ponent reactions can be estimated from only a readout of the total number fibrils over time[20]. These *kinetic aggregation assays* have made it possible to study the aggregation behaviour of amyloid-β under a diverse range of conditions such as the addition of small molecules[21].

1.1.2 Drug binding to disordered proteins

While the vast majority of small molecule drugs available today target specific structured binding pockets on folded proteins, there are few examples of small molecules directly binding disordered regions. This is despite the fact that disorder makes up so much of the human proteome and it's involvement in disease[22]. Reasons for this are the lack of understanding of the binding modalities, and the difficulty in characterising them[23]. I will briefly summarise some examples of drugs binding or interacting with disordered proteins.

A prominent system in this field is the interaction of the oncogenic proteins c-Myc and Max. Overactivation of c-Myc by Max is associated with various cancers[24], and this folding-upon-binding interaction has been investigated as a potential drug target. Numerous studies used FRET to quantify the association of the two proteins in the presence of various small molecules. Some of the subsequently discovered binding compounds feature planar conjugated motifs and can insert themselves into the interface, thus disrupting the interaction[25]. While some molecules in this and other screenings showed good specificity for c-Myc/Max[26], many others also had strong interactions with other related systems[27, 28] raising questions on specificity.

On the other hand, small molecules binding the monomeric disordered c-Myc protein, as opposed to interfering in it's interactions directly, have also been discovered. One such compound is 10058-F4[28] (Figure 1.4A), which has been studied both experimentally[29] and computationally[30]. The molecule was found to exhibit a 'specific-diffuse' binding mechanism, with a characteristic sequence specificity when compared to urea as a control molecule, but no fixed binding mode or any kind of folding-upon-binding behaviour. In addition, this interaction was shown to be entropically favourable using isothermal titration calorimetry and fluorescence titration experiments, with low enthalpic contributions[30].

Another example of molecules able to bind and favourably interact with disordered proteins are so-called *molecular tweezers*. These are typically rigid molecules featuring their own 'binding pocket' to accept a ligand, for example a positively

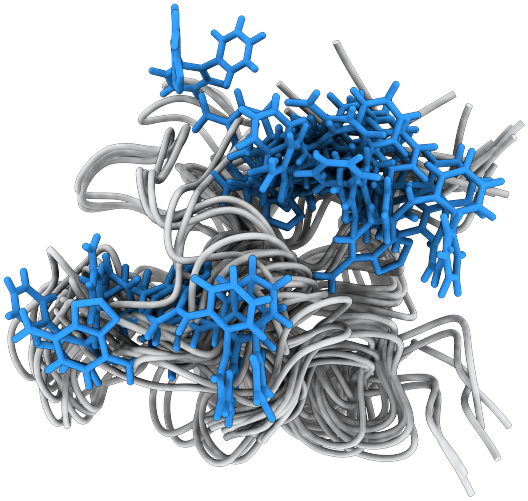


Figure 1.3: Interaction of 10074-G5 with Aβ42. Frames from a 70 ns segment of a molecular dynamics simulation are shown.

charged ion of a certain size. One instance is the tweezer CLR01 (Figure 1.4B), which can reversibly bind to exposed lysine residues as found in amyloid-β[31] or Parkinson's-associated α-synuclein[32].

Recently, my group elucidated the thermodynamics of the binding of the small molecule 10074-G5 (Figure 1.4C) – also characterised for the c-Myc/Max system[28] – to Aβ42[33] (Figure 1.3). Using the kinetic aggregation assays and mathematical model described above (Figure 1.2), the compound could be shown to bind monomeric amyloid-β as opposed to the oligomeric or fibrillar form, resulting in the protein's sequestration and a corresponding reduction in aggregation. Among numerous other experimental techniques, the system was also studied in simulations using the metadynamic metainference framework, again showing a highly dynamic and entropically favourable binding process. Chemical shifts from NMR experiments were unperturbed upon binding, potentially consistent with a diffuse interaction in which the protein remains disordered. However, the kinetics on the level of the monomer could not be characterised using either simulations or experiments. I will attempt to remedy this in chapter 3 and present a kinetic analysis of the interaction.

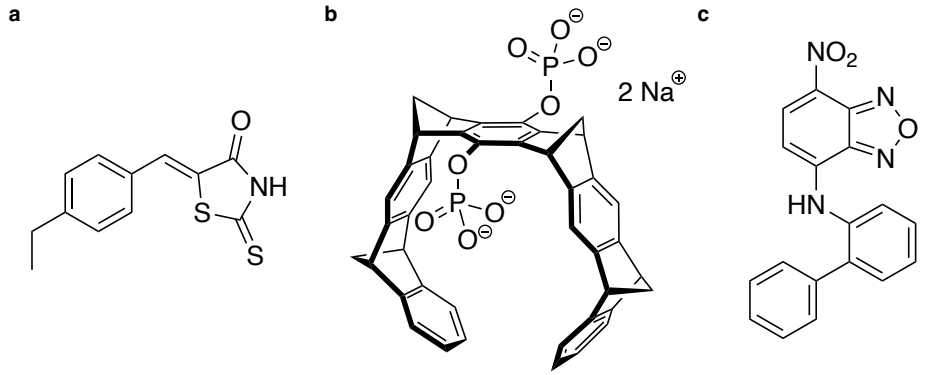


Figure 1.4: Small molecules known to bind monomeric disordered proteins. **a** 10058-F4, known to bind c-Myc[29], **b** molecular tweezer CLR01, binding exposed lysine residues in amyloid-β and α-synuclein[31, 32], **c** 10074-G5, known to bind Aβ42[33].

1.1.3 Disordered proteins as complex systems

Compared to folded proteins, disordered proteins show enormous spatial and temporal heterogeneity, not only in their structure, but also in their diverse interactions[34]. It is therefore natural to view them as complex systems, featuring non-linear and emergent behaviour. Examples of the latter include the aforementioned protein aggregation and liquid-liquid phase separation phenomena, while non-linearity can sometimes be observed in their temperature dependence: Folded proteins generally unfold with an increase in temperature, while some disordered proteins have been observed to become more ordered[35, 36]. Behaviour consistent with self-organized criticality – the tendency of some complex systems to converge to a 'tipping point'[37] – can also be observed in some proteins in the form of a power law distribution of hydrophobicities[38, 39] or their tendency express to a point close to insolubility[40]. All these phenomena point to the possibility of using tools from dynamical systems theory to effectively study and understand disordered proteins. I will introduce two related ideas from this field, Markov models and the related Koopman operator later in the chapter (Subsection 1.3.4).

1.2 Probing disorder

1.2.1 Limits of experimental methods

Over the last ~70 years, structure determination methods have evolved in massive strides. Structural biologists now have a large arsenal of tools available to them, covering various time- and length-scales (Figure 1.5). Traditionally, the vast majority of static structures have been solved using X-ray crystallography, with cryo electron microscopy (cryo-EM) quickly catching up [41].

Characterising disordered proteins and regions is infinitely more difficult. In X-ray crystallography, disordered regions are not resolved at all, whereas with cryo-EM they form regions of low electron density[42], thus yielding very little information on possible conformations. A folded protein is relatively static, with well-defined free energy minima, and as a result, a long-timescale observation we make of it will be subject to little variation. Attempting the same on a disordered protein will inherently result in a noisy measurement, and we will generally have to be content with an *ensemble average*. Thus many experimental approaches are unable to provide the resolution necessary to draw structural conclusions, or aren't applicable to disordered proteins to begin with[23]. I will briefly summarize some approaches to study the structure of proteins and their limitations with regards to disorder.

X-ray Crystallography X-ray crystallography[59] is one of the most widely used and also one of the oldest techniques to study the structure of proteins and their binding pockets. It relies on the fact that many proteins can be crystallized and then subjected to X-ray scattering, thus determining the overall structure. While most structures deposited on the protein data bank have been found through some form of X-ray crystallography[41], this approach is generally not applicable to disordered proteins as their structure is essentially heterogeneous and they are not crystallizable in their native disordered state. Furthermore, disordered regions in crystal structures are often invisible due to their dynamic nature.

Cryo electron microscopy (cryo-EM) In recent years, cryo electron microscopy has quickly advanced to yield structures of proteins and complexes of unprecedented size and potentially atomistic resolution[60]. In short, a solution of the protein of interest is applied to a fine grid and frozen within a very short time-frame before

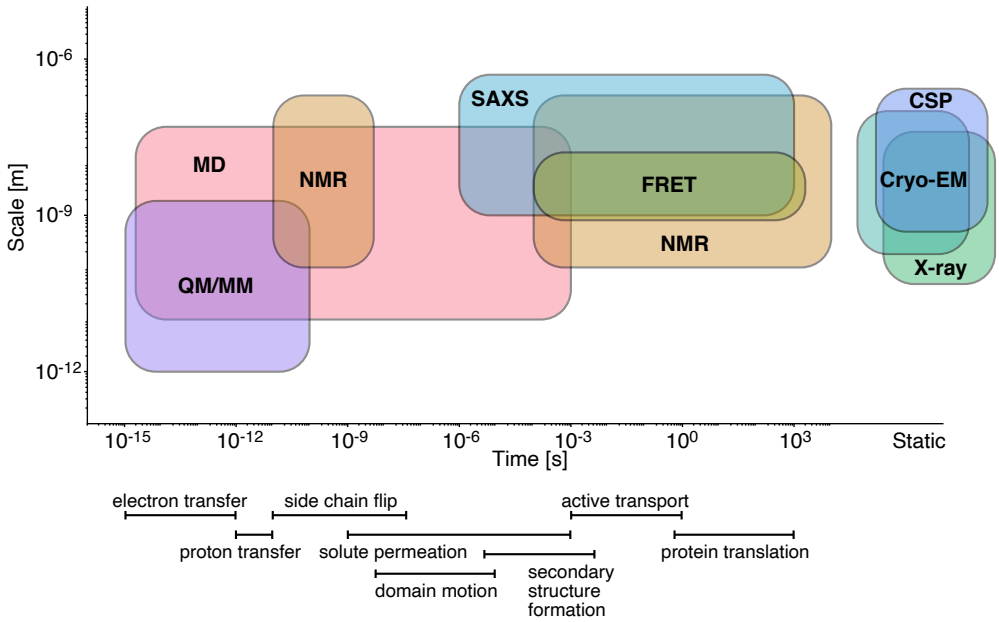


Figure 1.5: Time- and length-scales explorable with modern biophysical techniques, and the biological events that can be probed with them: molecular dynamics (MD)[43], nuclear magnetic resonance (NMR)[44–46], quantum mechanics / molecular mechanics (QM/MM)[47], small angle X-ray scattering (SAXS)[48–50], Förster resonance energy transfer (FRET)[51, 52], cryo electron microscopy (cryo-EM)[53, 54], X-ray crystallography[55, 56] and computational structure prediction (CSP)[57, 58].

being imaged using electron microscopy. The acquired two-dimensional images are then combined using computational methods to yield a detailed three-dimensional map of the electron density and thus the structure[61]. However, as disordered regions are highly dynamic and flexible, the electron density will appear to be very low and spread out, thus making structural characterizations difficult. Furthermore, not only can freezing coincide with major conformational changes[62], interactions with surfaces and other interfaces can also cause significant structural changes[63, 64].

Förster resonance energy transfer (FRET) Förster resonance energy transfer[65] can yield a fluorescence signal based on the proximity of two fluorophores, and thus

indicate a binding event. Single-molecule FRET in particular can provide information on non-equilibrium processes, rare states and events, and causal processes, such as binding and unbinding events[66]. However, the labels used are typically large and can interfere not only with the interactions being studied, but also with the conformational ensemble of the disordered protein[67].

Nuclear magnetic resonance (NMR) Nuclear magnetic resonance[68] is one of the most powerful and versatile structure determination and interaction probing methodologies available. It can be used without bulky, hydrophobic labels and in solution, two characteristics that are essential for the study of disorder. One of the main types of information gained from NMR are the chemical shifts of certain atoms of the protein, which are the resonance frequencies of the atom's nucleus with respect to a standard reference compound. They encode information on their chemical and structural environment and thus allow the determination of topology and structure of molecules[69]. This means we can still observe weakly-interacting regions, albeit in an *ensemble-averaged* way.

An issue common with disordered proteins is the low dispersion of chemical shifts due to the intrinsically uniform chemical and structural environment, making this measure less useful for structure determination. Another possible problem is the fast exchange of hydrogen at higher temperatures due to disordered proteins solvent-exposed nature, losing proton information[70, 71]. This often limits measurements to lower, non-physiological temperatures. Further structural information can be gained from NMR experiments, such as residual dipolar couplings (RDCs), which encode the relative orientation of certain bonds to a reference frame[72], ^{3}J-couplings[73], which are able to yield information on the protein backbone dihedral angle distributions, or utilising the nuclear Overhauser effect (NOE)[74], providing additional information on the proximity of certain atoms.

Small angle X-ray scattering (SAXS) Small angle X-ray scattering[75] is another scattering technique using hard X-rays, but recording significantly smaller scattering angles. This makes it possible to obtain structural *ensemble-averaged* information for systems in solution as opposed to a crystal, making it particularly useful for the study of disordered proteins. It is also well suited to integrative approaches using NMR and / or molecular simulations.

Computational structure prediction (CSP) Computational structure prediction methods based only on residue sequences and possibly multiple-sequence alignment (MSA) data have recently made headlines. Predictors using deep learning such as *AlphaFold2*[58] and *RoseTTA fold*[57] can achieve scores of 80 and higher (with 100 being a perfect match) on competition datasets such as Critical Assessment of protein Structure Prediction (CASP)[76]. They are however often limited to systems with available MSA data and few disordered regions, and while their accuracy is remarkable, the resulting structures are not generally highly-resolved enough to allow computer-aided drug design. Additionally, as these predictors have been trained on experimentally obtained structures, they will feature an inherent bias towards, for example, easily-crystallizable proteins.

Molecular dynamics (MD) Molecular dynamics is another computational technique making it possible to gain atomistic insight into the behaviour of molecules on varying timescales. While the computational expense grows with the simulated time and size of the system, recent advances in processing speed and other auxiliary algorithms have made it possible to study large systems and make predictions about their behaviour. On the other hand, their accuracy is fundamentally limited by the nature of the classical mechanics framework and the associated parameters – the *force field*. As molecular simulations are my tool of choice to investigate disordered proteins, I will dedicate the latter part of this introduction to them.

1.2.2 Kinetic aspects

So far, we have focused mostly on the structural and thermodynamic aspects of using experimental approaches to study disordered proteins. However, kinetic data – for example transitions between different states, or relaxation rates – present a further challenge for the structural biologist. In the experimental methods outlined above, structural properties are often ensemble-averaged and gaining fast-timescale kinetic information is therefore difficult.

NMR is one of the few technologies that allow the extraction of kinetic information. One such method is hydrogen-deuterium exchange (HDX)[77]. This idea relies on the fact that deuterium is *invisible* in NMR, we can thus subject our protein to a solution containing a certain percentage of D_2O. Over time, exposed regions will exchange their hydrogen with deuterium, therefore losing their signal. If we have a

system that can take on multiple states with different levels of exposure we will be able to make inferences about the transition rates.

Rates in the range of pico- to milliseconds can also be deduced from nuclear spin relaxation experiments[78], in particular relaxation dispersion[79]. The latter method can – if the differences in chemical shifts are large enough – resolve transitions between a ground and very low-population excited state. The obtained relaxation rates can thus give important clues on dihedral angle rotations and loop movement. This data has been successfully back-calculated from Markov models based on molecular dynamics[80].

Outside of the world of NMR, FRET has been successfully utilised to obtain relaxation and state lifetime estimates on the microsecond timescale, for example for Aβ42[81]. On the other end of the timescale spectrum, SAXS has been used to monitor conformational changes in real time on the minute timescale[82].

Despite these promising results, there remain large gaps in the obtainable kinetic information from experimental methods. This is especially evident in NMR, as there is a prominent lack of methods to cover the nano- to microsecond timescale range (Figure 1.5)[44]. This leaves us with molecular simulations as one of the most suitable methods to provide atomistic information on kinetics for small- to medium-sized systems. In the following I will continue with a more in-depth discussion of this method and related analysis techniques.

1.3 Molecular simulations

1.3.1 The ensemble framework

To make our lives easier when defining the precise goals of a molecular simulation and also aid in analysis, I will introduce the concept of *ensembles*. The following definitions are distinct from those of statistical mechanics by Gibbs[83]. We will restrict ourselves to the spatial coordinates only, as opposed to the classical view of phase space including velocity vectors. We define the set of all structures $\mathcal{X}^n$ for an n-atom system as:

$$\mathcal{X}^n = \prod_i^n \mathbb{R}^3 = \underbrace{\mathbb{R}^3 \times \ldots \times \mathbb{R}^3}_{n}, \tag{1.1}$$

This means we have x, y and z coordinates for every atom, giving us a $3n$-dimensional space. This space will allow us to subsequently define our ensembles.

Structural ensembles

Historically, the singular protein structure was seen as the central entity in structural studies. However, with the increased interest in disorder, we need to extend this framework to ensembles of structures. Instead of attempting to rely on a single structure $\mathbf{x} \in \mathcal{X}^n$ to explain the protein's behaviour, we consider a structural ensemble:

$$E^m_{\text{Struc}} = \left\{ (\mathbf{x}_1, ..., \mathbf{x}_m) \mid \mathbf{x}_i \in \mathcal{X}^n \right\} \tag{1.2}$$

of m conformations that can be explored by the protein. This allows us to illustrate flexible regions by providing possible alternative structures. If we decide to calculate an ensemble-averaged property $\langle A \rangle$ of our protein, we can do so by forming the arithmetic average over all m structures:

$$\langle A \rangle = \frac{1}{m} \sum_{i=1}^{m} A(\mathbf{x}_i) \tag{1.3}$$

Note that while Equation 1.3 makes no assumptions about the probability of encountering a certain structure, this probability is implied by how the conformations were generated. In molecular dynamics simulations, this is typically by sampling from the Boltzmann distribution. In the next section, we will expand this idea to explicitly account for different sampling distributions.

Thermodynamic ensembles

An inherent problem with structural ensembles is that in practice we have no information on the relative abundance of certain structures compared to others – or to put it more mathematically, we would need an *infinite* number of structures to properly encode their abundance. We would thus ideally associate a weight or probability to each single structure in the ensemble[1]. Formally speaking, this entails associating to

[1] Alternatively, one could use measure theory to more rigorously define these ensembles.

our subset of m structures a vector $\boldsymbol{w} \in \{(w_1, ..., w_m) \mid w_i \in \mathbb{R}^+\}$ with m elements. Our thermodynamic ensemble then becomes a pair[2]:

$$E^m_{\text{Thermo}} = \left(E^m_{\text{Struc}}, \boldsymbol{w}\right). \tag{1.4}$$

If we further impose a normalisation condition, i.e. $\sum_i^m w_i = 1$, then $\boldsymbol{w}$ becomes a probability distribution and p_i becomes the probability of observing a particular conformation $\mathbf{x}_i$[3]. This probability is related to the free energy F with:

$$F_i = -k_B T \log w_i \tag{1.5}$$

where k_B and T are the Boltzmann constant and the temperature, respectively. As a consequence, if we now want to calculate an ensemble-averaged property of the system, we need to use a weighted average:

$$\langle A \rangle = \frac{1}{\sum_{i=1}^m w_i} \sum_{i=1}^m w_i A(\mathbf{x}_i) \tag{1.6}$$

This weighting needs to be taken into account generally, for example when calculating probability distributions (using histograms), or when sampling a lower number of structures from the ensemble. There is still a crucial dimension missing in thermodynamic ensembles, *time*.

Kinetic ensembles

While we now know the conformations our protein can explore, and how likely those conformations are compared to others, we do not have any information on how these conformations are connected. In other words, we have no information on the timescales and pathways of interconversion between structures. This temporal information can be provided in a kinetic ensemble: we first assume that our structures have some underlying time structure, i.e. that they have been generated by some dynamical method; we will therefore now switch to using t as an index as opposed to i. We also assume we have some way of grouping similar structures together into states. We first associate a membership χ_t to each conformation $\mathbf{x}_t$. This membership can take several forms, but to be as general as possible we will introduce it as a point on

[2] This definition is essentially equivalent to the concept of fuzzy sets[84].

[3] We can alternatively view w_i as a function $w : \mathcal{X}^n \to \mathbb{R}^+; \quad w : \mathbf{x}_i \mapsto w_i$

a k-simplex $\chi_t \in \Delta^k$ defined as $\Delta^k = \{(q_1, ..., q_k) \mid q_i \in \mathbb{R}^+; \sum_i^k = 1\}$. Note that $k \ll m$. Conceptually, this membership encodes the degree of membership – or alternatively the probability – of finding our conformation $\mathbf{x}_t$ in some state. I will justify the use of this 'soft' membership compared to a more simple discrete state assignment in section 1.3.4 and chapter 2. In addition to these membership vectors, I will also introduce the idea of a transfer operator $\mathbf{P} : \Delta^k \to \Delta^k$ that can move the membership of our system by some time τ: $\mathbf{P} : \chi_t \mapsto \chi_{t+\tau}$. Our kinetic ensemble is thus defined as:

$$E^{(m,k)}_{\text{Kinetic}} = \left(E^m_{\text{Struc}}, (\chi_1, ..., \chi_m), \mathbf{P}\right) \tag{1.7}$$

We will explore this theme later in the form of *Markov models* and *Koopman operator theory* (Section 1.3.4).

1.3.2 Molecular dynamics

In molecular dynamics, we model the system as a ball-and-stick model using classical mechanics at an atomic level. In practical terms, this means first assigning spatial coordinates and velocities to each atom of the system, resulting in $6n$ degrees of freedom for n atoms. Next, we need some way of calculating the energy and thus forces of the system. These are dependent on a number of inter- and intramolecular interactions and have generally modelled after experimental[85] or quantum-chemical[86] results. This equation for the energy of a classical system is referred to as the force field, and one of the most common functional forms is as follows:

$$\begin{aligned} V(\mathbf{x}) = &\sum_{\text{bonds}} \frac{k_d}{2}(d(\mathbf{x}) - d_0)^2 + \sum_{\text{angles}} \frac{k_\theta}{2}(\theta(\mathbf{x}) - \theta_0)^2 \\ &+ \sum_{\text{dihedrals}} \frac{k_\phi}{2}(1 + \cos(n\phi(\mathbf{x}) - \phi_0)) + \sum_{\text{impropers}} \frac{k_\psi}{2}(\psi(\mathbf{x}) - \psi_0)^2 \\ &+ \sum_{\substack{\text{non-bonded} \\ \text{pairs (i,j)}}} 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}(\mathbf{x})}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}(\mathbf{x})}\right)^{6}\right] + \sum_{\substack{\text{non-bonded} \\ \text{pairs (i,j)}}} \frac{q_i \times q_j}{4\epsilon_D r_{ij}(\mathbf{x})} \end{aligned} \tag{1.8}$$

In short, it describes the energetic contributions of various motions, such as bond stretching ($d(\mathbf{x})$), angle bending ($\theta(\mathbf{x})$), rotation about dihedrals and impropers ($\phi(\mathbf{x})$ and $\psi(\mathbf{x})$), and long-range interactions such as electrostatic and van der Waals

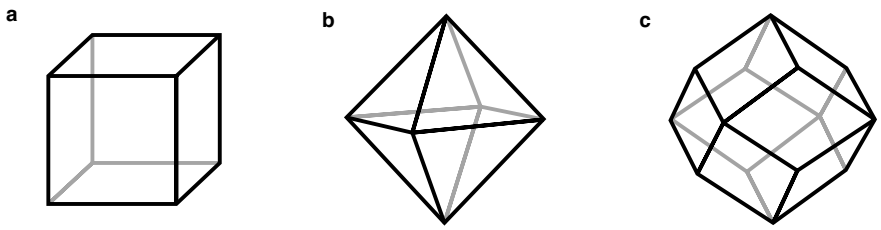


Figure 1.6: Types of boxes typically used for molecular dynamics simulations of biomolecules. **a** cubic, **b** octahedron, **c** rhombic dodecahedron.

forces (dependent on the distance $r_{ij}(\mathbf{x})$). While the bonded components are also dependent on certain force constants (k_d, k_θ, k_ϕ, k_ψ) and equilibrium values (d_0, θ_0, ϕ_0, ψ_0), the non-bonded interactions are based on Lennard-Jones parameters (σ_{ij}) and partial atomic charges (q_i).

The long range interactions require particular attention to avoid introducing artifacts at simulation boundaries. To correctly model these one makes use of periodic boundary conditions and the particle mesh Ewald method[87]. The former allows the system to interact with copies of itself, thus resolving artifacts originating from abrupt cut-offs and other truncation effects. The latter method uses a fast Fourier transform to efficiently calculate long range electrostatic interactions by discretizing point charges on to a lattice.

By choosing a suitable box geometry, one can avoid having the system experience an artificially high concentration and at the same time save computational cost by reducing the amount of solvent molecules required. A simple cubic box might be the easiest to conceptualize and implement, we however have to consider that the vast majority of (not only disordered) protein systems explore an approximately spherical volume. A suitable box choice therefore often takes the form of an *octahedron* or a *rhombic dodecahedron* (Figure 1.6).

Now that we have a way of evaluating the energy and corresponding forces $F(\mathbf{x}) = -\nabla V(\mathbf{x})$ on each atom of the system at every time-step, we can use an *integrator* to calculate new atomic positions $\mathbf{x}_i$ and velocities $\boldsymbol{v}_i$ (or equivalently momenta $\mathbf{p}_i = m_i \boldsymbol{v}_i$). We will use the concept of *Hamiltonian* dynamics to find a suitable integration scheme. A system is Hamiltonian if it satisfies the following property, in

which $\mathcal{H}(\mathbf{x}, \mathbf{p})$ (the *Hamiltonian*) is a functional representing the total energy and time-evolution of the system:

$$\frac{d\mathbf{x}}{dt} = +\frac{\partial \mathcal{H}}{\partial \mathbf{p}} \quad \text{and} \quad \frac{d\mathbf{p}}{dt} = -\frac{\partial \mathcal{H}}{\partial \mathbf{x}} \tag{1.9}$$

This solution space is known as a *symplectic* manifold, embedded in the general phase space $(\mathbf{x}, \mathbf{p})$ and defined by this special relationship between positions $\mathbf{x}$ and momenta $\mathbf{p}$. For a true Hamiltonian system, any trajectory traced on this manifold will have constant energy. We now need to find a numerical form for $\mathcal{H}(\mathbf{x}, \mathbf{p})$ preserving this property, also known as a *symplectic integrator*. Because our time step Δt is not infinitesimally small we will incur a discretization error $\mathcal{O}(\Delta t^n)$ related to the order n of the integrator. Higher order integrators effectively use fractional steps to reduce this error, but will incur a higher computational cost. One of the simplest symplectic integrators is the *symplectic Euler* integrator, a *first-order* method:

$$\boldsymbol{v}_{t+1} = \boldsymbol{v}_t + \boldsymbol{a}_t \Delta t + \mathcal{O}(\Delta t) \tag{1.10}$$

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \boldsymbol{v}_{t+1} \Delta t + \mathcal{O}(\Delta t) \tag{1.11}$$

We thus first calculate the new velocity $\boldsymbol{v}_{t+1}$ using our current estimate $\boldsymbol{v}_t$ and the acceleration (given by the force field and atomic masses $\mathbf{m} = [m_i]$ with $\boldsymbol{a}_t = \frac{F(\boldsymbol{x}_t)}{\mathbf{m}}$) before using this new velocity to compute the next atomic positions $\boldsymbol{x}_{t+1}$. This method results in an error on the order of the time step Δt. To improve our accuracy, we can use a *second-order* method, such as the *leapfrog* integrator, in which the positions and velocities are updated in an interleaved manner:

$$\boldsymbol{v}_{t+1/2} = \boldsymbol{v}_{t-1/2} + \boldsymbol{a}_t \Delta t + \mathcal{O}(\Delta t^2) \tag{1.12}$$

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \boldsymbol{v}_{t+1/2} \Delta t + \mathcal{O}(\Delta t^2) \tag{1.13}$$

This particular integrator is popular in molecular dynamics due it's symplectic nature, time-reversibility and relatively low computational cost. Unlike the Euler integrator (Equation 1.10), the leapfrog is stable for oscillatory motion such as bond vibrations. If even higher accuracy is required a higher-order method with

correspondingly lower error can be used, one example is the class of *Runge-Kutta* integrators. One however needs to verify that symplecticity and time-reversibility are preserved[88].

We can thus, using our originally assigned atomic positions, velocities, and the force field (Equation 1.8) calculate the positions and velocities after an arbitrary time t. This process is repeated until the desired simulation length has been reached. An important constraint here is that the time step Δt needs to be chosen in such a way as to fully resolve the fastest motion of the system. In the case of proteins this will typically be C-H bond vibration on the order of 10 fs, we therefore need to choose a time step of 2 fs or lower [4].

Once we have calculated a sufficient number of steps, we will be left with a trajectory of our system showing the time-evolution of the coordinates $\mathbf{x}$ and velocities $\mathbf{v}$. This leads us to the next question: How long is *sufficient*?

This question is as much a philosophical as it is a technical one. Given long enough timescales, even proteins that are considered as having an exceptionally stable fold will eventually visit an unfolded state. However, we often aren't interested in states so far away from equilibrium. We thus need to consider which specific question we are trying to answer.

To obtain data with adequate statistical significance, our (equilibrium) trajectories should be several times longer than the process of interest (Figure 1.5), so that it can be observed multiple times, or at least have sampled that process through other means, for example with enhanced sampling methods (section 1.3.3). This in turn often requires that we know the reaction coordinate (or *collective variable*) that we are trying to probe, so that we can evaluate our progress.

A necessary, but not strictly sufficient method to assess convergence in this case is by splitting our trajectory into two or more parts, and calculating a probability distribution of our coordinate of interest for each part. If both probability distributions match within a certain threshold, we can consider the simulation as fulfilling an important convergence criterion. Additionally, any kind of grouping or state definition should be stable, i.e. state populations should not be changing over time. Finally, we can also look at temporal effects: in a converged simulation we would expect state transitions to be constant and not fluctuate over chunks. The latter is a

[4] We can use longer time steps using the hydrogen mass repartitioning technique, in which the mass on the C-H bond is distributed equally on to both atoms. This has the effect of significantly slowing the vibration and thus allowing time steps of up to 5 fs with little loss in accuracy[89].

strong convergence criterion as it ensures that (possibly rare) state transitions have been sampled adequately, in contrast to the former two methods which only strictly require consistent sampling of the state itself and not the transition.

1.3.3 Increasing sampling efficiency

I will now take a brief detour into the field of *enhanced sampling*. Because we will be adding additional energy to our system, we will no longer be sampling the Boltzmann distribution and thus won't be able to make use of the time dimension for our analysis. This is because the rates depend not only on the free energy barrier between states, but also on the route taken. These enhanced sampling methods thus generally[5] sacrifice our ability to extract kinetic information from the system. However, we will be able to gain massive increases in sampling speed, provided we have some basic intuition about the behaviour of our system. The general idea behind most enhanced sampling methods is the modification of the potential energy function $V(\mathbf{x})$, either by changing parameters for non-bonded interactions (Equation 1.8), or by introducing an additional energy term to drive the system to new states.

Metadynamics

As one of the latter methods, *metadynamics*[90, 91] adds an additional time-dependent biasing potential V_{MetaD} to the force field V_{FF}:

$$V(\mathbf{x}) = V_{\text{FF}}(\mathbf{x}) + V_{\text{MetaD}}(\xi(\mathbf{x}), t) \tag{1.14}$$

The idea is to keep track of the system's location along a collective variable (CV) $\xi : \mathcal{X}^n \to \mathbb{R}$, a function of the system coordinates $\mathbf{x}$, by depositing a small gaussian 'hill' every time period τ_G. Each hill will reduce the probability of the system revisiting the same location on this CV. Once enough gaussians have been accumulated, the system can escape a local free energy minimum and explore less favourable states (Figure 1.7). The metadynamics potential $V_{\text{MetaD}}(\boldsymbol{\xi}, t)$ thus takes the following form for several CVs $\boldsymbol{\xi} = [\xi_i]$:

[5] Recent advances have allowed the recovery of kinetics from biased simulations, for example *infrequent metadynamics*.

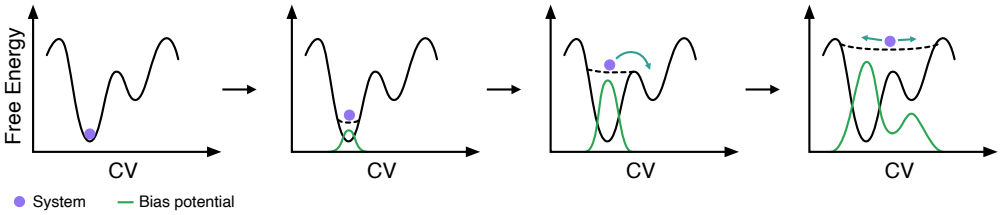


Figure 1.7: A schematic representation of the metadynamics algorithm. The system (purple dot) is subjected to a bias potential. After several iterations, the bias accumulates and allows the system to escape the minimum, exploring the full free energy surface. Adapted from Bussi et al.[91].

$$V_{\mathrm{MetaD}}(\xi, t) = \sum_{i=1}^{t/\tau_G} w_G \exp\left(-\sum_{j=1}^{N_{CV}} \frac{(\xi_j - \xi_j(\mathbf{x}_t))^2}{2\sigma_j^2}\right) \tag{1.15}$$

where w_G, σ_j are the height and width of the gaussian, and N_{CV} is the number of collective variables. The number of CVs that can be biased is limited due to the *curse of dimensionality*, the more dimensions that are added, the more gaussians need to be deposited to fill the same space. Additionally, for computational efficiency these gaussians are stored on a grid, which needs exponentially more memory for higher dimensions.

One of the countless flavours based on the original metadynamics method, *Parallel-bias metadynamics*[92], solves this dimensionality problem by instead depositing multiple one-dimensional gaussians instead of a high-dimensional single one. This allows dozens of CVs to be biased at once.

1.3.4 Kinetics from simulation

Obtaining kinetic information from simulations imposes an additional requirement: We need to be sampling from the Boltzmann distribution with no additional modifications for enhanced sampling methods or other restraints. While it is not generally required, a way to *discretize* space and time can also be immensely helpful in the model building process. This usually entails a clustering of similar structures – yielding a discretization in space – and the choice of a *lag time* τ, which represents our temporal resolution. I will first discuss the *Markov model* as a simple way to get useful kinetic information out of dynamic systems[93], and then introduce *Koop-*

man models[94], which represent a generalisation of the former with some useful properties.

Markov models

The state space of our molecular system is continuous, but with a discrete time dimension. The central assumption making this process *Markovian* is that the future state $\mathbf{x}_{t+\tau}$ of our system depends *only* on the current state $\mathbf{x}_t$, and not on any previous states[95]. While a continuous state space treatment is mathematically appropriate, discrete states are more suitable to build an actual model[93]. Practically speaking, this entails first classifying structures into discrete states, and then counting transitions between states within a lag time τ. These counts can be represented in a matrix $\mathbf{Z}(\tau)$ in which each entry z_{ij} represents the number of transitions from state i to state j. In the limit of an infinitely long trajectory, the transition matrix $\mathbf{P}(\tau)$ can be obtained by simply normalising the rows of this count matrix $\mathbf{Z}$:

$$p_{ij} = \frac{z_{ij}}{\sum_j z_{ij}} \tag{1.16}$$

We now have a transition matrix $\mathbf{P}(\tau)$ giving us the probability of observing a state transition from state i to j within a time period τ. However, with finite trajectories, there are a number of different transition matrices that can generate the same trajectory. In fact, the probability $p(\mathbf{Z}(\tau) \mid \mathbf{P}(\tau))$ that a certain transition matrix $\mathbf{P}(\tau)$ generates a particular count matrix $\mathbf{Z}(\tau)$ can be seen as the product of the individual transition probabilities. We can thus formulate this probability as follows:

$$p(\mathbf{Z}(\tau) \mid \mathbf{P}(\tau)) = \prod_{i,j} p_{ij}^{z_{ij}} \tag{1.17}$$

Equation 1.17 represents a likelihood function, and by using Bayes' theorem and imposing a suitable prior[6], we can indeed show that Equation 1.16 is the *maximum likelihood* solution to this equation[93]. A common constraint, especially useful to us when analysing equilibrium simulation data, is the *detailed balance* or *reversibility* condition:

[6] In the simplest case this can be the uniform prior with $z_{ij}^{\text{prior}} = 0$, making the posterior $p(\mathbf{P}(\tau) \mid \mathbf{Z}(\tau))$ equal to the likelihood.

$$\pi_i p_{ij} = \pi_j p_{ji} \tag{1.18}$$

This means that at equilibrium, each microscopic transition is in equilibrium with it's reverse process. Notably, Equation 1.16 does not include this constraint, it thus has to be explicitly included when determining the maximum likelihood solution to Equation 1.17.

How do we find and / or classify the states to be able to count transitions? For molecular systems we will typically need some grouping criterion. Formally, we need to find a function $\chi : \mathcal{X}^n \rightarrow \mathcal{S}$, with $\mathcal{S} = s_1, s_2, ..., s_m$ being our m-state space. Molecular simulations are often approached as follows: first, a suitable set of system coordinates is chosen, this could be suitably transformed atomic coordinates or some kind of internal, translation- and rotation-invariant coordinates such as residue-residue contact maps or backbone dihedral angles. Then, a dimensionality reduction technique is used to move from the typically high-dimensional space to a lower dimensional one more suitable for clustering algorithms. Common methods are principal component analysis (PCA), time-structure independent component analysis (TICA)[96, 97], multidimensional scaling (MDS)[98] or uniform manifold approximation and projection (UMAP)[99]. The system frames are then clustered in this low-dimensional embedding, for example using *k*-Means[100], *k*-Medoids, agglomerative[101], or density-peak clustering[102]. This procedure gives us the function χ to map from structures to states.

The next critical parameter choice we have to make is the model lag time τ. It should ideally be small enough to resolve fast processes, but also avoid introducing errors into the estimation of slower processes. The choice can be made by first building trial models using the general approach outlined above and then visualizing the dependence of the timescales t_i – a function of the eigenvalues of the transition matrix $\mathbf{P}$ – on the lag time τ. We would like these timescales to be independent of the chosen and longer lag times.

What information can we acquire from our model? The decomposition of the transition matrix $\mathbf{P}$ into the eigenvalues λ_i and eigenvectors $\mathbf{r}_i$ can give us valuable information about system properties:

$$\mathbf{P}(\tau)\mathbf{r}_i = \lambda_i \mathbf{r}_i \tag{1.19}$$

Each eigenvalue/eigenfunction pair $(\lambda_i, \mathbf{r}_i)$ corresponds to a dynamical process, the superpositions of the latter form the slow dynamics of the system. The maximum eigenvalue is $\lambda_0 = 1$, and it's associated eigenvector is the stationary (equilibrium) distribution of the system, commonly denoted by $\boldsymbol{\pi}$. If we were to observe our system for an infinitely long time, the system would tend towards this distribution. All other eigenvalues $\lambda_i < 1$ correspond to finite decay times, with higher values thus representing slower processes. These eigenvalues λ_i can be expressed as relaxation timescales t_i by relating them to the model lag time τ:

$$t_i = \frac{-\tau}{\log|\lambda_i|} \tag{1.20}$$

As we can see, with $\lambda_i \to 1$, $t_i \to \infty$, thus providing an intuitive basis for the stationary distribution $\boldsymbol{\pi}$. We can further study the eigenvalues to reveal the number of slow processes involved in our system: Apart from $\lambda_1 = 1$, a two-state system would feature a second eigenvalue $\lambda_2 \approx 1$ very close to 1, but all subsequent eigenvalues $\lambda_i \ll \lambda_2$ will be considerably smaller. This *spectral gap* can thus in some sense indicate the dimensionality of our kinetic landscape and help us find an ideal state decomposition.

Koopman models

I will now present a generalisation of Markov models that will become especially useful for disordered proteins. I will mostly be following the work of Mardt[103], Wu[104] and Klus[94].

We would like to find a linear operator $\mathbf{K}$ that can (approximately) propagate our system in time. This process can be described as follows:

$$\mathbb{E}[\boldsymbol{\chi}^{(1)}(\mathbf{x}_{t+\tau})] = \mathbf{K}^\top \mathbb{E}[\boldsymbol{\chi}^{(0)}(\mathbf{x}_t)] \tag{1.21}$$

Here, $\mathbb{E}$ denotes a time average to account for random fluctuations in the dynamics, while $\boldsymbol{\chi}^{(1)} = (\chi_1^{(1)}(\mathbf{x}), ..., \chi_k^{(1)}(\mathbf{x}))^\top$ and $\boldsymbol{\chi}^{(0)} = (\chi_1^{(0)}(\mathbf{x}), ..., \chi_k^{(0)}(\mathbf{x}))^\top$ represent transformations $\chi : \mathbb{R}^l \to \mathcal{S}$ from the original space – in which the dynamics may be non-linear – into some feature space $\mathcal{S}$ with approximately linear dynamics[7]. In molecular simulations we will also have a function $h : \mathcal{X}^n \to \mathbb{R}^l$ giving us an intermediate space, typically with $l \ll n$, for example some kind of dimensionality

[7] For practical purposes, we will consider both transformations to be identical, i.e. $\boldsymbol{\chi}^{(1)} = \boldsymbol{\chi}^{(0)} = \boldsymbol{\chi}$.

reduction. In the limit of an unlimited number of feature transformations χ we recover the exact dynamics. To reproduce our Markov model introduced above, we first need to partition our original space into sub-states $S_i \subset \mathbb{R}^l$, we can then define indicator functions yielding state membership:

$$\chi_i^{(1)}(\mathbf{x}) = \chi_i^{(0)}(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} \in X_i \\ 0 & \text{if } \mathbf{x} \notin X_i \end{cases} \tag{1.22}$$

This allows us to retain a probabilistic interpretation of the model. This means that $\chi^{(1)} = \chi^{(0)}$ become probability vectors encoding state membership, and the entries of the matrix K_{ij} become transition probabilities. Alternatively, we can define $\chi^{(1)}(\mathbf{x}) = \chi^{(0)}(\mathbf{x}) = \mathbf{x}$, this identity transformation forms the basis of approaches such as dynamic mode decomposition (DMD)[105] and TICA[96, 97] and assumes linearity in the original state.

Our goal now is to find the best possible transformations χ. From the variational approach to Markov processes (VAMP) theory[104], these should be equal to the top k left and right singular functions of the infinite-dimensional, but linear Koopman operator $\mathcal{K}$[94]. This is equivalent to solving the following minimization problem:

$$\min_{\chi^{(0)},\chi^{(1)}} \mathbb{E}_t \left[\left\| \chi^{(0)}(\mathbf{x}_{t+\tau}) - \mathbf{K}^\top \chi^{(1)}(\mathbf{x}_t) \right\|^2 \right] \tag{1.23}$$

Mardt et al. first define the following covariance matrices[103]:

$$\mathbf{C}_{00} = \mathbb{E}_t \left[\chi^{(0)}(\mathbf{x}_t) \chi^{(0)}(\mathbf{x}_t)^\top \right] \tag{1.24}$$

$$\mathbf{C}_{01} = \mathbb{E}_t \left[\chi^{(0)}(\mathbf{x}_t) \chi^{(1)}(\mathbf{x}_{t+\tau})^\top \right] \tag{1.25}$$

$$\mathbf{C}_{11} = \mathbb{E}_{t+\tau} \left[\chi^{(1)}(\mathbf{x}_{t+\tau}) \chi^{(1)}(\mathbf{x}_{t+\tau})^\top \right] \tag{1.26}$$

The solution for $\mathbf{K}$ to 1.23 then becomes:

$$\mathbf{K} = \mathbf{C}_{00}^{-1} \mathbf{C}_{01} \tag{1.27}$$

We would now like to find the best possible functions $\chi^{(0)}$ and $\chi^{(1)}$, but due to technical reasons we can not find the solution by minimizing Equation 1.23[103].

Using the VAMP variational principle[104] and the previously defined covariance matrices (Eqs. 1.24) we obtain the VAMP-2 score $\hat{\mathrm{R}}_2$:

$$\hat{\mathrm{R}}_2 \left[\chi^{(0)}, \chi^{(1)}\right] = \left\| \mathbf{C}_{00}^{-\frac{1}{2}} \mathbf{C}_{01} \mathbf{C}_{11}^{-\frac{1}{2}} \right\|_{\mathrm{F}}^2 \tag{1.28}$$

We now have a scoring function that we can utilise to evaluate the suitability of our choice of transformation functions $\chi^{(0)}$ and $\chi^{(1)}$. We can in fact use it to score Markov models built using the approach outlined in the previous section, and thus select the most appropriate clustering algorithm, the number of clusters, the type of dimensionality reduction, and so on.

Alternatively, we can use a system that can act as a universal function approximator, namely a neural network. For this, the VAMP2 score gets converted into a loss function, and two lobes of the neural network can learn the corresponding transformation functions[103]. This elegantly allows us to ignore the intricacies of any clustering and dimensionality reduction methods, because the neural network learns the most optimal transformation for the system of interest. This idea has been recently expanded to account for detailed balance – the requirement that when stationary, each process is in equilibrium with it's reverse process (see Equation 1.18) – and the avoidance of negative values in the Koopman matrix $\mathbf{K}$, making it a correct transition matrix[106].

Note that at no point have we required that our transformations $\chi^{(0)}$ and $\chi^{(1)}$ take the form of indicator functions (Equation 1.22). This means that, with certain constraints to our basis functions or neural network, we can view these transformations as probabilistic state assignments. This has profound implications for disordered proteins, which feature an inherently shallow energy landscape and low free energy barriers, as we can now classify their state as a mixture of substates instead of a single discrete one. I will, together with a novel neural network architecture, make extensive use of this methodology in chapters 2 and 3.

1.3.5 Limits of simulation

In the preceding sections, I have introduced molecular simulations and advanced analysis techniques from the area of dynamical systems as useful tools to study disordered proteins. We have seen that simulations are ideally suited to covering the 'experimentally awkward' nano- to microsecond timescale range and can in

principle provide atomistic resolution. Nevertheless, it is important to keep in mind that simulations present a simplified model of the world, and we are making many implicit assumptions about the interactions of our systems. I will outline some prominent sources of error and possible solutions.

Systematic error

The first problem we have is that of *systematic error*, i.e. the approximations we make in modelling our physical system based on the complex laws of quantum mechanics as a classical system. With the use of a force field, we ignore the electronic structure and thus chemistry of our system, making it impossible to study reactions and catalytic processes. We also generally assume that atomic charge is statically distributed in our system instead of being polarized[8]. Traditionally, force fields were optimized for folded systems, and disordered proteins thus often showed high modelling errors[108]. The increased interest in disorder, better experimental and quantum-mechanical fitting data, and improved fitting procedures means that modern force fields are now more optimized for a wider variety of systems[109]. Recently, much work has been put into machine learned potentials[110].

Methods using experimental restraints or *a posteriori* re-weighting approaches present an alternative solution to this issue. Molecular simulations can be subjected to a bias potential that can force the system to explore the region of state space most compatible with experimental data for that particular system. One such method, *metainference*[111, 112] takes into account the maximum entropy principle and potential errors in the experimental data or forward model to allow the use of many heterogeneous datasets. As an alternative, re-weighting approaches[113, 114] can provide a new weight vector $\boldsymbol{w}$ for the thermodynamic ensemble (Equation 1.4). Because with restraining methods we are no longer sampling from the Boltzmann distribution, we cannot use the resulting trajectories to build a kinetic ensemble. However, re-weighting approaches for Markov models exist, for example in the form of *augmented Markov models*[115], or a recent extension to the VAMPNet approach[116].

[8] So called *polarizable force fields*[107] can overcome this problem to a certain degree, with a corresponding steep rise in computational cost.

Even if re-weighting or restraining are not possible, we can at the very least estimate the impact of systematic error on our ensemble by backcalculating observables and comparing them to the experimentally determined values.

Statistical error

On the other side we have the issue of *statistical error*, most commonly in the form of insufficient sampling. To make accurate predictions from our simulations, it is not enough to use a physically suitable framework, but we also need to make enough observations to have our predictions carry weight. However, computational resources are limited, and accessing high microsecond timescales – required to observe many important biological phenomena – can be very demanding for large systems. Enhanced sampling approaches such as metadynamics[90] (Subsection 1.3.3), together with wise collective variable choices, can alleviate this problem somewhat and provide order-of-magnitude increases in sampling efficiency. Unfortunately we are then generally unable to recover kinetic information from our system, as we are again no longer sampling the Boltzmann distribution. A possible compromise is *adaptive sampling*, in which we can influence the sampling direction of a system by running many short trajectories and frequently resampling a set of new starting conformations according to some criterion. The system can thus be driven to new states without interfering with its equilibrium distribution. Two such ideas are termed weighted ensemble[117] and fluctuation-amplification of specific traits (FAST)[118], the latter of which I will make use of in chapters 2 and 3.

Statistical error sources are not limited to the simulation methodology itself, but also to the analysis methods used on the resulting data. One of the main points of Markov models is to discretize the system in time and space, thus introducing a corresponding discretization error that will influence any rates and relaxation constants calculated from the model. Maximum entropy approaches, like the metainference method mentioned above, rely on ensemble averages from a limited number of simulation replicas to calculate an experimental value, thus introducing statistical error[119][9].

To estimate the error incurred by limited sampling, we can perform strict convergence checks for our ensemble. For instance, monitoring the change of state populations for random subsets of our ensemble can yield important information

[9] This error is usually explicitly accounted for in the calculation of the restraint potential.

on the correctness of the thermodynamics. To evaluate the quality of our sampled state transitions, we can build multiple kinetic models on subsets of our data and compare fundamental features of the transition matrix **P** or **K**, such as the relaxation timescales.

1.4 Aims

So far I have introduced disordered proteins, their role in biology and disease, potential drug targeting approaches, and methods to study them, with a focus on computational methods. The behaviour and interactions of disordered proteins in terms of kinetics is poorly understood, due to the previous lack of experimental and computational methods able to cover the fast timescales involved. New developments in dynamical systems theory, specifically the VAMPNet approach[103, 106], now enable us to lift the veil on nano- and microsecond kinetics with high accuracy. I will use this methodology to answer some key questions about the Alzheimer's-associated disordered protein amyloid-β 42 (Aβ42):

1. Does monomeric Aβ42 feature distinct states?
2. How fast are the transitions between these states?
3. What are the structural origins of the kinetic barriers?
4. What timescales are involved in the dynamic interaction with a small molecule?

In chapter 2 I use the VAMPNet approach to construct kinetic ensembles of Aβ42 and it's methionine-oxidized variant. I found that the protein occupies distinct states with transitions on the microsecond timescale involving the folding and unfolding of transient secondary structure. The highest-population state is more extended, and transitions mostly move through this state, giving it a hub-like importance. With oxidation of methionine at residue 35, the population is shifted towards the hub state, with transitions toward it correspondingly increased. The less-folded hub state has analogies to the proposed inverted free-energy landscape[120] and the kinetic hub model of protein folding[121].

In chapter 3 I use the same methodology to study the interaction of the previously[33] thermodynamically characterised small molecule 10074-G5 with Aβ42 in terms of the kinetics, with urea as a control molecule. I discovered that this

interaction is governed by many transient nanosecond-scale π-π contacts with key residues. The global effect of this is a drastic population increase for the main hub state, with corresponding acceleration of transitions into it. In addition, while the backbone entropy of the protein is increased, side-chain dihedral autocorrelations indicate local enthalpic stabilisation. These results suggest a binding mode that is both enthalpically and entropically favourable and hint towards the possibility of obtaining drug specificity.

Disorder is of course not limited to entire sequences, but also occurs regionally. sdAbs often feature partially disordered areas in the form of CDRs, parts of the protein that determine their binding behaviour. Developing antibodies for specific targets involves carefully designing a suitable CDR to obtain high affinities, using either purely experimental techniques or computational sequence- or structure-matching approaches. I sought answers to the following two questions:

1. How does a structurally-designed sdAb antibody differ from an sequence-matching based design?

2. Is reduced binding affinity the result of lower or higher conformational entropy in the CDR?

In chapter 4 I attempt to answer these questions by performing metadynamics simulations of a designed antibody based on a sequence-matching approach targeting oligomers ('rational design') and a structurally-designed one. The rationally-designed antibody showed considerably higher flexibility in the CDR, despite the shorter sequence. This shows that the sequence of the CDR may have to be carefully chosen to obtain a certain number of CDR — scaffold contacts acting as 'anchors', thus reducing the conformational entropy and potentially improving affinity.

2 A kinetic ensemble of the Alzheimer's Aβ peptide

If it's a good idea, go ahead and do it.
It's much easier to apologize than it is to get permission.
— Grace Hopper

This chapter has been adapted from my first-author publication of the same title[122]. I designed the study and performed the analysis. Kai Kohlhoff and I ran the simulations. Kai Kohlhoff, Gabriella Heller and Carlo Camilloni assisted with analysis. Michele Vendruscolo supervised the work. Michele Vendruscolo and I wrote the manuscript with assistance from my co-authors.

2.1 Summary

The conformational and thermodynamic properties of disordered proteins are commonly described in terms of structural ensembles and free energy landscapes. To provide information on the transition rates between the different states populated by these proteins, it would be desirable to generalize this description to 'kinetic ensembles'. Approaches based on the theory of stochastic processes can be particularly suitable for this purpose. Here, we develop a Markov state model and apply it to determine a kinetic ensemble of amyloid-β 42 (Aβ42), a disordered peptide associated with Alzheimer's disease. Through the Google Compute Engine, we generated 315 μs all-atom molecular dynamics trajectories. Using a probabilistic-based definition of conformational states in a neural network approach, we found that Aβ42

is characterized by inter-state transitions on the μs timescale, exhibiting only fully unfolded or short-lived, partially-folded states. Our results illustrate how kinetic ensembles provide effective information about the structure, thermodynamics, and kinetics of disordered proteins.

2.2 Introduction

Proteins that are fully or partially disordered make up approximately one third of the human proteome, perform a variety of biological functions and are closely involved with many major human disorders[22, 123]. The existence of disordered proteins is making it necessary to extend the structure-function relationship that has driven major advances in protein science in the last 50 years to a structure-dynamics-function relationship, in order to account for the essential role of structural disorder in determining the normal and aberrant behaviours of these proteins[22, 123, 124].

Because of their conformational heterogeneity, it is typically insufficient to characterize disordered proteins using one or a few specific structures, as it is standard for folded proteins[41]. Instead, it has become common to describe these proteins in terms of structural ensembles, which in turn are often represented through free energy landscapes[125–127], when the statistical weights of the states in the ensemble are available, for example, through the use of enhanced sampling techniques[125]. This is a powerful description, which concisely captures information about the structure and thermodynamics of disordered proteins. Here, we refer to these ensembles here as 'thermodynamic ensembles.' It has also been recently observed, however, that a more complete description should include information about the kinetics, which can be achieved by adding the transition rates between different conformational states[42]. We refer to these ensembles here as 'kinetic ensembles.' This task requires a characterisation of the kinetic properties of disordered proteins, which remains a challenging task, both experimentally and computationally.

In this work, we describe an approach to generate kinetic ensembles, which contain information about the molecular structures of proteins, the populations of their metastable states, and the transition rates between these different metastable states (Figure 2.1). While acquiring structural and population information is already possible with molecular dynamics simulations alone, gaining interpretable kinetic information can be more effectively achieved by exploiting the theory of stochas-

tic processes[128] during the analysis, and it is typically done using Markov state models[95, 129].

In contrast to thermodynamic ensembles, constructing kinetic ensembles requires a state decomposition. To define the states in these models, one can choose a set of features from a trajectory, such as backbone dihedral angles or the root-mean-square deviation to some reference structure, and then use a clustering algorithm to obtain a state assignment for each frame. One can then count the transitions between states and normalize these counts to obtain a transition matrix[95, 130]. An additional coarse-graining step can then be performed to obtain a more interpretable model with fewer states. Alternatively, clustering can be preceded with an additional dimensionality reduction step. One such example particularly relevant for building Markov models is time-structure independent component analysis (TICA)[96, 97] which projects conformations into a space where distances have a kinetic meaning. Clustering in this space thus has the potential of naturally preserving kinetic separation.

This particular model-building approach presents unique challenges for disordered proteins, due to the heterogeneous nature of their conformations. In disordered proteins, transitions between the short-lived states are typically fast and not necessarily characterized by large variations in the overall shape of the protein. This is in contrast to folded proteins, where states can often readily be classified on global structural properties alone[93, 131, 132], such as the open and closed states of a G-protein binding receptor[131]. For disordered proteins, even with suitable structural measures, because these transitions are fast, dividing this space into discrete areas is exceedingly difficult (Figure 2.1b), and many clustering algorithms may fail to consistently separate states. Consequently, it would be ideal to obtain a type of clustering, which we can describe as 'soft,' such that state assignments can have a probabilistic nature. In this approach, every frame is assigned a discrete probability distribution encoding its membership in a certain state. This information can be used to build a soft Markov state model, which can be interpreted in much the same way as one built from individual transition counts[94]. To optimize for kinetic separation and allow for probabilistic state assignments, we employed the VAMPNet technique with physical constraints (Figure 2.1c)[103, 106] to construct a soft Markov state model (or Koopman model, see Methods). The resulting stochastic matrix allows

us to compute the equilibrium distribution of the ensemble as well as the transition rates between metastable states.

We illustrate this approach to determine a kinetic ensemble of amyloid-β 42 (Aβ42), using 315 μs of explicit solvent molecular dynamics simulations. We chose Aβ42 because of its association with Alzheimer's disease, which is among the leading causes of death in the developed world, with no disease-modifying treatments available[133]. One of the characteristics of Alzheimer's disease is the formation of neurotoxic aggregates of Aβ42[17, 19, 134]. This complex process can be divided in to a set of non-linearly coupled microscopic steps, for all of which the monomeric state of the protein plays a key role[135]. It is therefore of great importance to better understand the structural and kinetic properties of the monomeric state of Aβ42, as the dynamic properties in particular are key to determine the effects of small molecule drug candidates on the aggregation behaviour of this peptide[33, 136]. Because of the ensemble-averaged nature of most experimental measurements of Aβ42, molecular simulations can provide a uniquely precise, atomistic description of its kinetic ensemble. Previous computational studies to characterize Aβ42[81, 120, 137–140] have focused mainly on the thermodynamic properties of this peptide, but less so on the kinetics of state transitions. It is therefore still debated as to whether this peptide ever adopts long-lived (i.e. more than hundreds of μs) states associated with its aggregation behaviour. We validate the kinetic ensemble using independent experimental data obtained from nuclear magnetic resonance (NMR) spectroscopy. The resulting kinetic ensemble provides the structural properties and the populations of the states of Aβ42 and the transition rates between them. It also allows us to make predictions of involved timescales and is especially informative in terms of the kinetics of secondary structure formation. We believe that this approach offers a unique perspective into the structure, thermodynamics, and kinetics of disordered proteins towards an increased understanding of their dynamic behaviour.

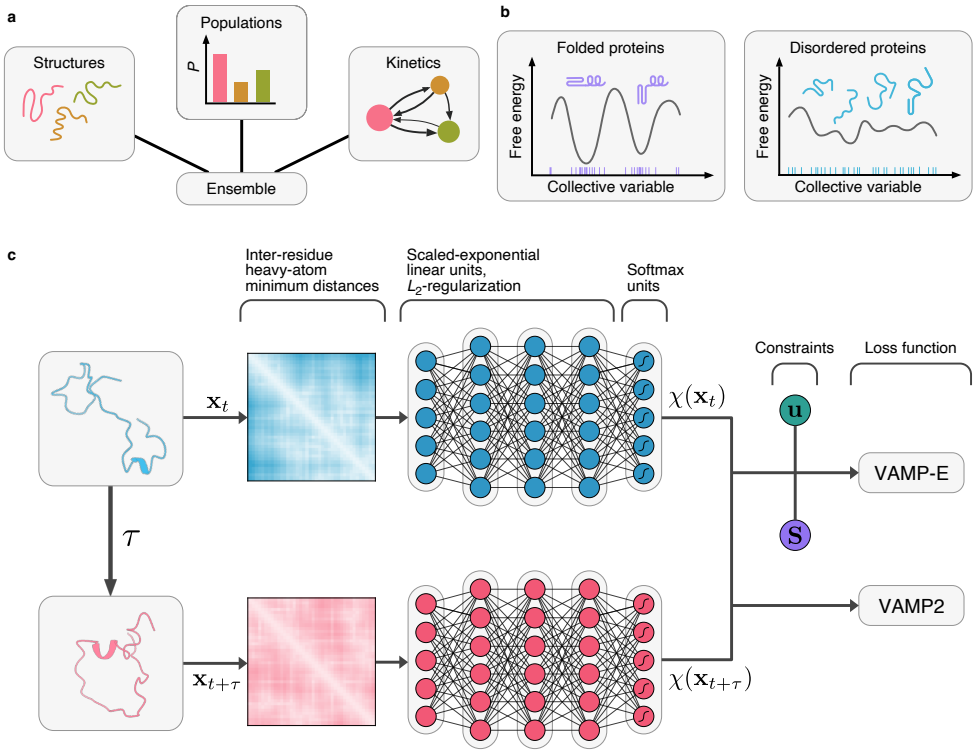


Figure 2.1: Illustration of a kinetic ensemble of a protein and the training methodology. **a**, A kinetic ensemble consists of three components: (1) structures, (2) their statistical weights (populations) and (3) interconversion rates between groups of related structures (states). Structural and thermodynamic ensembles have only the first or first two components, respectively. **b**, The free energy landscape of disordered proteins is flatter and more heterogeneous than that of folded proteins, making a state decomposition difficult, and thus the determination of a kinetic ensemble challenging. **c**, Architecture and function of the constrained VAMPnet (Methods).

2.3 Results

Molecular dynamics simulations on the Google Compute Engine. We performed explicit solvent molecular dynamics simulations in 5 rounds with 1024 trajectories each, using the fluctuation-amplification of specific traits (FAST) approach[118] to accelerate the exploration of conformational space. Cloud architectures, such as the Google Compute Engine are especially well suited to generate Markov state models, as the individual simulations can be run independently from one another, with no communication between machines required[141]. We chose the CHARMM22 $\times$ over the CHARMM36m force field due to the better agreement with macroscopic observables such as radius of gyration for the closely related Aβ40 peptide[109, 142].

Standard Markov model techniques. We first attempted to build discrete Markov state models using existing state-of-the-art techniques. We used TICA[96, 97] as a preliminary dimensionality reduction step with inter-residue nearest-neighbour heavy atom distances, followed by clustering with various algorithms, different numbers of input dimensions, and varying amounts of clusters. We used the cross-validated generalized matrix Rayleigh coefficient technique[143] to score the individual models (see Methods, Supplementary Figure 7.1). Despite an extensive hyperparameter search, we were unable to construct a model with converging timescales (Extended Data Figure 2.5 a). Nevertheless, we used the best-performing hyperparameters, specifically the minibatch *k*-Means clustering algorithm using 16 dimensional input and 200 clusters to build discrete coarse-grained models using the hidden Markov state model[144] (Extended Data Figure 2.5 b-d) and Perron cluster-cluster analysis approaches[145] (Extended Data Figure 2.5 e-f). These models, however, also suffered from non-converging or overly fast timescales, suggesting the presence of significant problems in this particular approach for state-space discretization. We therefore decided to pursue an approach based on a probabilistic state description.

Soft Markov state models. We next attempted to build a neural network for the disordered ensemble of Aβ42 with soft state assignments. To this end we employed the VAMPNet technique with physical constraints[103, 106]. VAMPNet is a two-lobed, unsupervised neural network, only taking as input two frames separated by a lag time τ, and yielding a soft state assignment vector $\chi(\mathbf{x}_t)$ for each frame $\mathbf{x}_t$ (Figure 2.1 b). The loss function is given by the variational approach to Markov

processes (VAMP) score[104], allowing the learning of a state decomposition without explicit state labelling. The dimensionality reduction and clustering steps are thus performed in a single procedure, allowing for highly non-linear state membership functions to be learned. A recent addition[106] to this approach allows the use of constraints to keep the elements in the learned transition matrix positive, and the model reversible. This is accomplished by training two auxiliary weights **u** and **S** and using the VAMP-E loss function. The transition matrix can then be learned by first training the unconstrained VAMPNet, followed by the constraint vectors and finally all trainable parameters together. We adopted this constrained framework but employed a self-normalising neural network architecture[146] to improve the hyperparameter search (see Methods). As input, we used inter-residue nearest-neighbour heavy atom distances, resulting in an input dimension of 780, with a network lag time of 5 ns. We used 2 to 6 output states, finding that the 4-state model struck the best balance between interpretability, detail, and state assignment errors (Extended Data Figures 2.6 and 2.7). Outputs with increased numbers of states resulted in larger errors and a lack of interpretability.

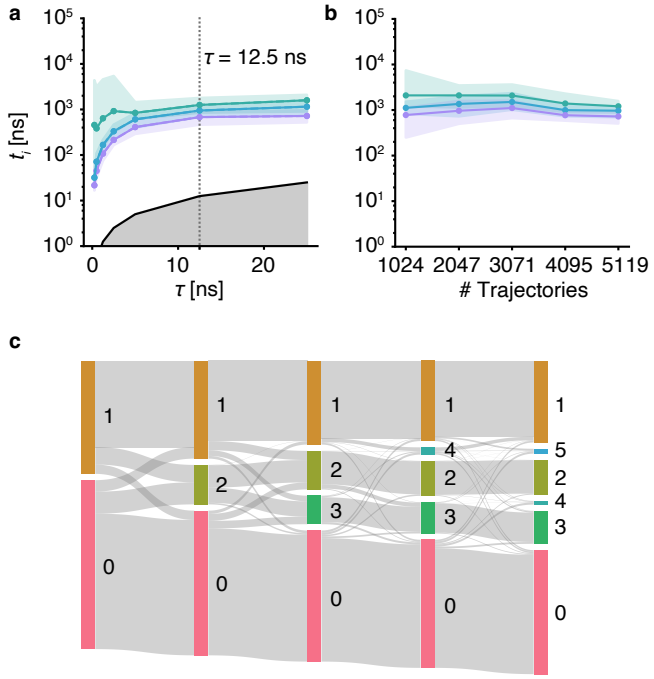


Figure 2.2: Determination of the states in the kinetic ensemble of Aβ42. **a**, Dependence of the three longest relaxation timescales (green, cyan and purple, respectively) on the lag time τ. These timescales are derived from the eigenvalues of the transition matrix (Methods). The gray shading indicates the timescales for which the Koopman model can no longer resolve the relaxation timescales. Shaded colored areas indicate 95th percentiles of the bootstrap sample of the mean over all 20 models. **b**, Dependence of the relaxation timescales on the number of trajectories used to build the kinetic ensemble as a four-state model. Shaded areas indicate 95th percentiles of the bootstrap sample of the mean over all 20 models. **c**, Shift of equilibrium distributions and state assignments over different output sizes. Gray lines indicate state decomposition across multiple models with varying numbers of states, showing consistency across different output sizes.

Computational validation of the Aβ42 kinetic ensemble. To construct a kinetic model, a suitable lag time τ between successive conformations along a trajectory needs to be chosen. This lag time should be small enough for the model to effectively resolve fast degrees of freedom, but large enough to not introduce a significant

statistical error in the prediction of longer timescales. To choose this parameter and evaluate the general quality of the model it is helpful to visualize the longer timescales — a kinetic property of the model — as a function of the lag time itself. To this end, we assessed the dependence of the relaxation timescales, t_i, (also called implied timescales in this approach, see Methods) on the model lag time τ (Figure 2.2 a), while keeping the network lag time constant at $\tau = 5$ ns. We observed that a model lag time $\tau = 12.5$ ns can resolve longer timescales well. To further validate this choice for the model lag time, we used the Chapman-Kolmogorov test (Supplementary Figures 7.2 to 7.4), a stringent measure of the predictive abilities of our model. To evaluate the quality of sampling in the context of the final model, we estimated Koopman operators with a limited number of trajectories (Figure 2.2 b) from the existing full model. We would expect the relaxation timescales to accelerate and converge with the number of trajectories utilized as we improve the sampling of each state transition. The timescales indeed converge to within the error of the model. To understand how the choice of the number of states impacts the corresponding state assignments of individual frames, we illustrate the state decomposition as a tree (Figure 2.2 c). States can be seen to be mostly consistent across different output sizes.

Experimental validation of the Aβ42 kinetic ensemble. We compared experimental NMR chemical shifts[33] to values back-calculated from the simulations (Extended Data Figure 2.8 a). We found that the root-mean-square deviation between experiments and simulations is within the error of the forward model[147]. Additionally, we directly compared the distributions of the radius of gyration between the kinetic ensemble derived here and of a thermodynamic ensemble determined previously using metadynamic metainference[111, 148] simulations[33], using the same force field and experimental restraints in the form of chemical shifts (Extended Data Figure 2.8 b). Previous chemical shift, ^{3}J-coupling, and nuclear Overhauser effect (NOE) measurements[149] imply fast relaxations faster than 10 µs consistent with our results (Figure 2.2 a). However, acquiring more precise data on state transitions of disordered proteins using NMR spectroscopy is limited by the vast structural variance of those proteins on short timescales and ensemble-averaged nature of the experiments[150].

Unconstrained model comparison. We compared the transition matrices, relaxation timescales and state lifetimes generated by constrained and unconstrained VAMPNets (Supplementary Figure 7.5). We found that all timescales and transition probabilities are consistently slower in the unconstrained case. This is to be expected, as the estimation of the eigenfunctions is impeded by the constraints on non-negativity and reversibility[94]. However, as we are performing equilibrium simulations, we should expect detailed balance to hold. We therefore consider the constrained model to be better suited to this analysis.

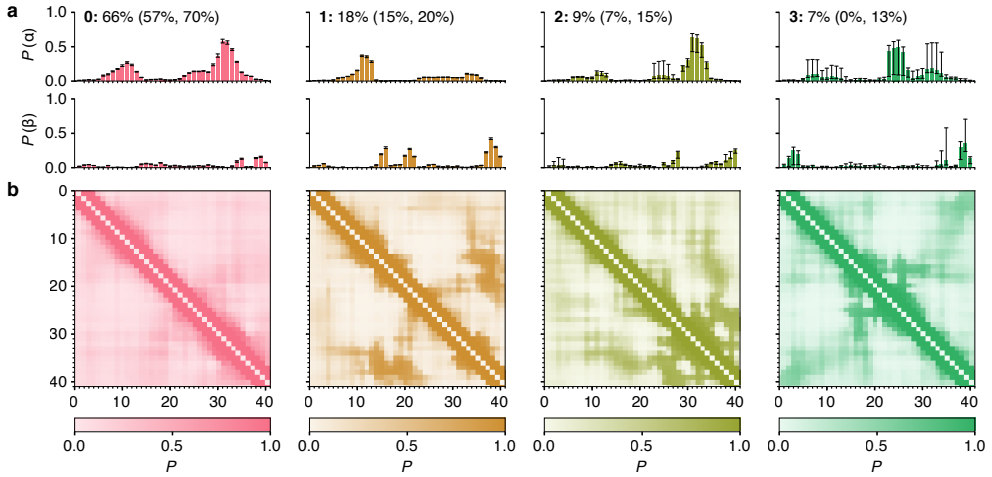


Figure 2.3: Structural properties of Aβ42 in the kinetic ensemble. **a**, Populations of α-helical and β-sheet content per residue in each of the four states in the kinetic ensemble, as calculated using the secondary structure predictor DSSP. The equilibrium percentage of each state is given above, with the 95th percentile in parentheses. **b**, Contact probability maps of each of the four states with a cutoff of 0.8 nm. Error bars indicate 95th percentiles of the bootstrap sample of the mean over all 20 models.

Structural diversity of the Aβ42 states. To characterize the local structural features and long-ranged self-interactions of the states in the kinetic ensemble, we calculated the α-helical and β-sheet content per residue (Figure 2.3 a, Supplementary Figures 7.6 and 7.7 a) and inter-residue contact probabilities (Figure 2.3 b, Supplementary Figures 7.6 and 7.7 b) in each state. We observed unique structural properties in each state and found that the C-terminal region of the peptide is generally more structured

than the N-terminus. Specifically, each state is characterized by varying degrees of α-helix formation in localized areas, while the region between residues 10 and 20 generally remains completely disordered. The β-sheet content is also concentrated towards the C-terminal region with state 3 being a notable exception, partially forming an end-to-end contact. Moreover, many conformations identified by the predictor Dictionary of Protein Secondary Structure (DSSP) as being rich in β-sheet content are characterized more by strong backbone interactions and less by explicit β-like structures, as can be seen in the contact probability maps. Using a cut-off of 0.45 nm we find an ensemble-wide proportion of 3.68 ± 0.14 % of end-to-end contacts, comparable to the ensemble reported by Meng et al[81].

μs transitions between Aβ42 states. The overall kinetic properties of a protein are primarily given by the inter-state transition probabilities (Figure 2.4 a), the slowest relaxation timescales (Figure 2.4 b), and mean state lifetimes (Figure 2.4 c). We observe mean first-passage times (MFPTs) between ~1 and ~12 μs, with no observation of longer-lived folded states. Instead, we observe the formation of a disordered hub-like state 0, to which other structured states transition relatively quickly. On the other hand, transitions into structured states, i.e. direct formation of secondary structural motifs, are significantly slower. Transitions bypassing the hub-state are rare, such as transitions between states 1 to 2. It may appear counter-intuitive that most MFPTs between states are slower than the equilibrium relaxations (Figure 2.4 b). This apparent discrepancy was previously investigated for the fast-folding proteins Trp-cage[151] and NTL9[152] by investigating the kinetics of their respective unfolded states. It was found that for these states, the MFPTs are significantly slower than the relaxations to unfolded equilibrium. Specifically, the MFPTs to a particular state can be seen to be approximately equal to the relaxation to equilibrium divided by the population of that state, implying slower MFPTs for finer discretizations. This observation is consistent with our ensemble with MFPTs between 1 and 12 μs (Figure 2.4 a) and equilibrium relaxations on the order of 2 μs (Figure 2.4 b). With coarser and finer discretizations we see acceleration and deceleration of MFPTs respectively (Supplementary Figure 7.13).

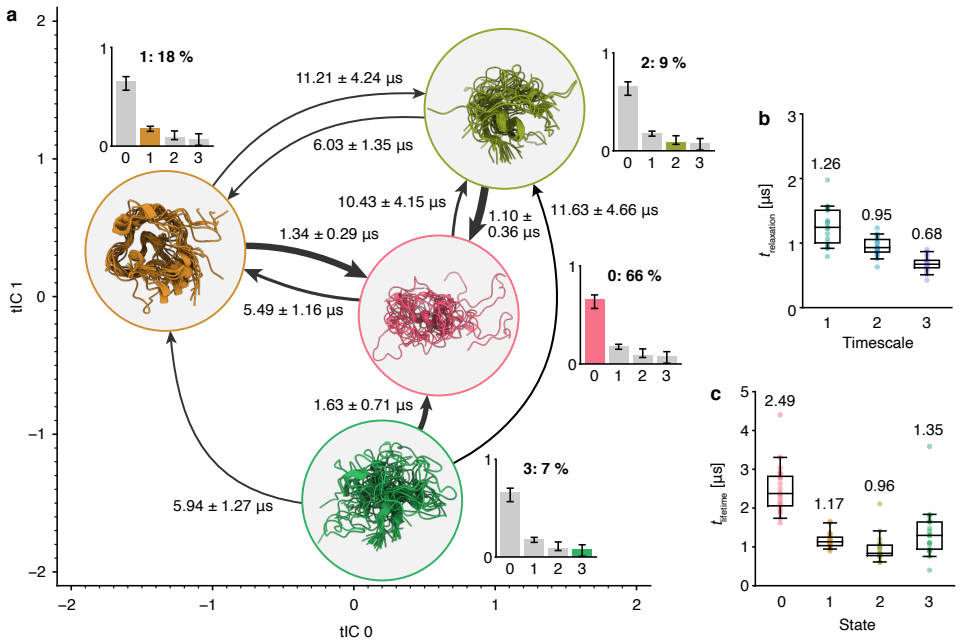


Figure 2.4: Populations and mean first-passage times in the kinetic ensemble of Aβ42. **a**, Mean first-passage times and their standard deviations between states in the kinetic ensemble. Thicker arrows correspond to higher transition probabilities. The state location is projected on to the two slowest time-structure independent coordinates (TICs) and the structures shown are 20 high-weight conformations from all 20 models aligned on the most prominent secondary structure motifs (Figure 2.3). Transitions with mean first-passage times slower than 20 µs are not shown (Supplementary Figure 7.8 a, c). **b**, The slowest relaxation timescales of the four-state model. **c**, Mean lifetime of each state in the four-state model. The whiskers, boxes and horizontal lines indicate 95th percentiles, quartiles and the median values over all 20 models, respectively. The labels show the mean values over all 20 models.

Methionine-oxidized Aβ42 kinetic ensemble. To contextualize our findings and demonstrate the transferability of this approach, we also performed additional simulations of the methionine-oxidized form of Aβ42 (Aβ42-MetSO). These were carried out in three rounds of 1024 trajectories each, using the same FAST procedure reported above. We validated our model using the same approach, i.e. plotting the dependence of the relaxation timescales on the model lag time and number of trajectories (Supplementary Figure 7.9) and using the Chapman-Kolmogorov test (Supplementary Figure 7.10). We found that the state decomposition remains largely identical (Supplementary Figure 7.11), with a population shift away from the ordered states 1, 2, and 3 towards the more extended state 0. This shift is also evident in the relaxation timescales and state transitions towards state 0, which are accelerated in the Aβ42-MetSO ensemble as compared to that of Aβ42 (Supplementary Figure 7.12 a, b). Additionally, the lifetime of state 0 is also vastly prolonged, from ~1 μs in the Aβ42 ensemble to ~6 μs for Aβ42-MetSO (Supplementary Figure 7.12 c). These findings are consistent with nuclear magnetic resonance studies on both forms, in which Aβ42-MetSO was shown to exhibit higher backbone mobility than Aβ42[153] and overall lower β-sheet content[154]. These results coincide with higher solubility and reduced aggregation propensity[155].

2.4 Discussion

In this work we determined the kinetic ensemble of Aβ42 using a neural network approach. We observed that the choice of the number and quality of inputs is crucial. In particular, we observed a large increase in the VAMP2 scores upon switching from inter-residue Cα distance matrices to nearest-neighbour heavy atom distances. These results suggest that using more sophisticated and higher resolution features could further improve the state decomposition. One such method could be the use of convolutional layers[156–158], either acting on distance matrices or on the 3D protein structure itself.

Our study also identifies the importance of developing robust mathematical tools to handle the Koopman matrix and corresponding error estimations. There have been numerous developments in the Markov model literature for these problems, such as the use of Bayesian methods to estimate both state discretization and statistical sampling uncertainties[159] or the use of additional experimental data to improve

the transition matrix estimation[115]. These tools are, however, all based on the availability of discrete transition counts, and there are no direct analogs for Koopman models so far.

Aβ42 has been studied extensively using molecular simulations in the context of thermodynamic ensembles, commonly using enhanced sampling techniques such as replica-exchange molecular dynamics or variants thereof[81, 120, 137–140]. These and related studies suggest that this peptide is extremely sensitive to the choice of force field and water model[160] as well as the amount of sampling[109]. These studies used higher simulation temperatures, making a quantitative comparison with our study difficult. The work by Lin et al.[137] is especially notable as the authors used a Markov state model to analyze their 200 μs simulation. They, however, did not report on the kinetic properties of the system and focused on ensemble-averaged structural features instead. Meng et al. also identified a large disordered population of 72 % and an end-to-end contact population of 3 % in their analysis[81], consistent with our analysis. Rosenman et al. found many highly diverse clusters with populations of ~5 %, comprising both extended and locally structured conformations[138]. Sgourakis et al. used a spectral clustering technique based on contact maps to identify many locally structured conformations[139]. Overall, given that it is challenging to obtain atomic-level information about the kinetic ensemble of Aβ42 directly from experimental data due to the disordered nature of this peptide, computational approaches such as the one described here offer important structural, thermodynamic and kinetic insights into possible drug discovery approaches for Alzheimer's disease aimed at stabilizing the native state of Aβ42[33].

Our results are particularly relevant when viewed in context of the kinetic hub model of protein folding[121]. In this model, the native state of a folded protein is the most populated state, quickly reachable from other partially or completely unfolded states. This model can be seen as the kinetic complement of the thermodynamic funnel concept of protein folding[161]. We may therefore propose an analogous model for disordered proteins, i.e. a kinetic counterpart to the recently proposed inverted funnel concept[120]. Aβ42 clearly adopts a highly populated disordered state, with slow transitions into less populated, partially folded states. We can thus view this kinetic geometry as an 'inverted kinetic hub'.

Overall, the ability to capture accurately the structural and kinetic differences between the Aβ42 and Aβ42-MetSO peptides, which only differ by a single atom, in

a way that is consistent with independent experimental data demonstrates the ability of the soft Markov state model method for studying the complexity of disordered proteins.

2.5 Methods

Simulation details. All simulations were carried out with GROMACS 2018.1[162] using 1,024 individual Virtual Machine instances on Google Compute Engine. All instances used the compute-optimized `n1-highcpu-8` machine type, each configured with 8 Intel Haswell CPU cores, 7.2 GB of RAM, and 100 GB of disk space. 1,024 Aβ42 starting conformations were chosen from a previous metadynamics-biased molecular dynamics simulation by sample-weighted k-Means clustering on the space of backbone dihedral angles and picking a random conformation from each of the 1024 clusters. Each conformation was solvated separately in a rhombic dodecahedron box with a volume of 358 nm^3 using between 11,711 and 11,751 water molecules. The system was minimized using the steepest descent algorithm to a target force of less than 1000 kJ / (mol / nm). Equilibration was performed over a time range of 500 ps in the NVT ensemble using the Bussi thermostat[163] and 500 ps in the NPT ensemble using Berendsen pressure coupling[164] while applying a position restraint on all heavy atoms. Production simulations were carried out at 278 K using the CHARMM22* force field[165] and the TIP3P water model[166] using a 2 fs timestep in the NVT ensemble. Electrostatic interactions were modelled using the Particle-Mesh-Ewald approach[87] with a cut-off for the short-range interactions of 1.2 nm. Constraints were applied on all bonds with the Linear Constraint Solver (LINCS) algorithm[167] using a matrix expansion on the order of 4 and 1 iteration per step. Simulations were carried out with the initial 1,024 starting conformations, the resulting trajectories were then used with the fluctuation-amplification of specific traits (FAST) approach[118] to choose 1,024 new starting structures. The clustering for FAST was conducted by first performing time-structure independent component analysis (TICA)[96, 97] with a lag time of 5 ns using Cα-distance matrices as input and the *k*-Means algorithm in this reduced space to create 128 clusters. We chose to both maximize the deviation to the mean Cα-distance matrix for each cluster and maximize the sampling of existing clusters using a balance parameter of $\alpha = 1.0$. All amino acids were weighted equally. This procedure was repeated three times to

yield a total of 5,119 trajectories, with an aggregated simulation time of 315 µs. The shortest and longest trajectory lengths were 9.75 and 90.5 ns respectively. The trajectories were subsampled to 250 ps time steps to yield 1,259,172 frames. Simulations of methionine-oxidized Aβ42 were performed analogously in three rounds using the same FAST scheme and simulation parameters yielding 3071 trajectories with an aggregated simulation time of 317 µs and 1,268,139 frames. Oxidized methionine parameters were constructed based on the parameters for dimethyl sulfoxide from the CHARMM generalized force field[168]. Initial conformations were created by sampling 1024 structures from the Aβ42 ensemble and mutating the methionine residues to the oxidized form.

Neural network. State decomposition was performed using the VAMPNet approach with physical constraints[103, 106]. The two-lobed neural network takes as input a pair of frames separated by some lag time and yields a state assignment vector as output. We chose inter-residue nearest-neighbour heavy atom distance matrices with first- and second-degree off-diagonals removed as input (780 dimensions) and used between 2 and 8 output nodes with softmax activation. The architecture of the χ layers of the neural network follows the self-normalizing setup described by Klambauer et al.[146], using scaled-exponential linear units, normal LeCun weight initialization, and alpha dropout. The constrained part is composed of two additional layers **u** and **S** which are trained with and without the χ layers, successively. The training procedure is based on the VAMP2 score:

$$\mathrm{R}\left[\chi_t, \chi_{t+\tau}\right] = \left\| \mathbf{C}_{\chi\chi}^{-\frac{1}{2}} \mathbf{C}_{\chi\tau} \mathbf{C}_{\tau\tau}^{-\frac{1}{2}} \right\|_{\mathrm{F}}^2 \tag{2.1}$$

where

$$\mathbf{C}_{\chi\chi} = \mathbb{E}[\chi(\mathbf{x}_t)\chi(\mathbf{x}_t)^\top] \tag{2.2}$$

$$\mathbf{C}_{\chi\tau} = \mathbb{E}[\chi(\mathbf{x}_t)\chi(\mathbf{x}_{t+\tau})^\top] \tag{2.3}$$

$$\mathbf{C}_{\tau\tau} = \mathbb{E}[\chi(\mathbf{x}_{t+\tau})\chi(\mathbf{x}_{t+\tau})^\top] \tag{2.4}$$

The maximum value of Equation 2.1 is reached when the subspaces spanned by the left and right singular functions of the Koopman operator **K** are identical to the ones spanned by χ_t and $\chi_{t+\tau}$. Because the first singular functions are always equal

to the constant function 1, we can simply add 1 to the score for it to become more suitable for training the $\chi(\mathbf{x}_t)$ network:

$$R\left[\begin{pmatrix}1\\ \chi_t\end{pmatrix},\begin{pmatrix}1\\ \chi_{t+\tau}\end{pmatrix}\right] = \left\|\mathbf{C}_{\chi\chi}^{-\frac{1}{2}}\mathbf{C}_{\chi\tau}\mathbf{C}_{\tau\tau}^{-\frac{1}{2}}\right\|_F^2 + 1 \tag{2.5}$$

The constrained network is trained using the VAMP-e score[106] R:

$$R = \mathrm{tr}(\mathbf{S}^\top\mathbf{C}_{\chi\chi}\mathbf{S}\mathbf{C}_{\gamma\gamma} - 2\mathbf{S}^\top\mathbf{C}_{\chi\gamma}) \tag{2.6}$$

where

$$\mathbf{C}_{\chi\gamma} = \mathbb{E}[\chi(\mathbf{x}_t)\gamma(\mathbf{x}_{t+\tau})^\top] \tag{2.7}$$
$$\mathbf{C}_{\gamma\gamma} = \mathbb{E}[\gamma(\mathbf{x}_{t+\tau})\gamma(\mathbf{x}_{t+\tau})^\top] \tag{2.8}$$
$$\gamma(\mathbf{x}) = \chi(\mathbf{x})\chi(\mathbf{x})^\top\mathbf{u} \tag{2.9}$$

and $\mathbf{S}$ and $\mathbf{u}$ are the trainable reversibility and non-negativity constraints, respectively. Error estimates were obtained through bootstrap aggregation (bagging), i.e. by training 20 independent neural networks with independently randomized and shuffled 9:1 train-validation splits, and, to prevent overfitting, stopping training when the validation score no longer improved. The model was implemented using Keras 2.2.4[169] with the Tensorflow 2.1.0[170] backend. The neural network parameters were chosen through two successive random grid searches with scikit-optimize 0.5.2[171], first using a coarse grid spanning a large parameter space, then a finer grid over a local space around the optimum parameters. We found the best parameters to be a network lag time of 5 ns, a layer width of 256 nodes, a depth of 5 layers, an L_2 regularization strength of 10^{-8} and no dropout. Training was performed on a single Google Compute Engine instance with an NVidia Tesla V100 GPU, 12 Haswell cores, and 78 GB of RAM. Training of the χ model was performed using batch sizes of 10000 frame pairs and the Adam minimizer with a learning rate of 0.05, $\beta_2 = 0.99$ and $\epsilon = 0.0001$. Overfitting was addressed through early stopping, i.e. training was stopped when the VAMP validation set score did not increase by at least 0.001 over the previous 5 epochs. The constraint layers $\mathbf{u}$ and $\mathbf{S}$ as well as the full network including the χ layers were trained with the Adam minimizer with a learning rate of

0.0005. The constraint layers were trained with batches the size of the full dataset, with early stopping with no improvement in the loss after 10 epochs.

Analysis. The output of a single neural network is a state assignment vector $\chi(\mathbf{x}_t)$ of each frame $\mathbf{x}_t$ of the simulation. The ensemble averaged value of an observable $A(\mathbf{x}_t)$ for a state i is therefore an average weighted by the state assignment for T time steps:

$$\langle A_i \rangle = \left(\sum_{t=1}^{T} \chi_i(\mathbf{x}_t) \right)^{-1} \sum_{t=1}^{T} \chi_i(\mathbf{x}_t) A(\mathbf{x}_t) \tag{2.10}$$

Ensemble averaged quantities can be calculated by first computing a weight w_t for each frame of the simulation based on the state assignment $\chi(\mathbf{x}_t)$ at time t and the equilibrium distribution π:

$$w_t = \frac{\langle \chi(\mathbf{x}_t) | \pi \rangle}{\sum_{t=1}^{T} \langle \chi(\mathbf{x}_t) | \pi \rangle} \tag{2.11}$$

The ensemble averaged observable $\langle A \rangle$ can then be calculated as the weighted average:

$$\langle A \rangle = \sum_{t=1}^{T} w_t A(\mathbf{x}_t) \tag{2.12}$$

To estimate the error, we sort each state assignment vector, as multiple trained neural networks do not necessarily conserve the order of identified states. To do so, we calculate the mean inter-residue nearest-neighbour heavy atom distance matrix for each identified state and sort the states based on the lowest root-mean-square deviation between these matrices. We also ensure that the sorting is unique, i.e. we are never duplicating state assignments. The Koopman matrix $\mathbf{K}(\tau)$ is then estimated directly from the constrained neural network. Internal model validation is performed with the Chapman-Kolmogorov test (Supplementary Figures 7.2 to 7.4 and 7.10):

$$\mathbf{K}(n\tau) \approx \mathbf{K}^n(\tau) \tag{2.13}$$

i.e. we expect a model estimated at some lag time τ to behave the same way as one estimated at a multiple $n\tau$ of it. The correct lag time τ for the model is estimated

by plotting the relaxation (also known as implied) timescales t_i (Figure 2.4 a and Supplementary Figure 7.9 a):

$$t_i = \frac{-\tau}{\log|\lambda_i|} \tag{2.14}$$

where λ_i is the ith eigenvalue of the Koopman matrix $\mathbf{K}(\tau)$:

$$\mathbf{K}(\tau)\mathbf{r}_i = \lambda_i \mathbf{r}_i \tag{2.15}$$

As the Koopman matrix is row-stochastic, the largest eigenvalue is 1, and its associated eigenvector is the stationary distribution π of the system. The state lifetimes $\bar{t}_i$ are given by[172]

$$\bar{t}_i = \frac{-\tau}{\log\left(\mathbf{K}(\tau)_{ii}\right)} \tag{2.16}$$

where $\mathbf{K}(\tau)_{ii}$ are the diagonal entries of the Koopman matrix[1].

Chemical shifts were back calculated with CamShift[147] as implemented in PLUMED 2.4.1[173, 174] using the full trajectories, averaged as described above and compared to previously recorded NMR data[33]. Time-structure independent component analysis was performed with inter-residue nearest-neighbour heavy atom distances as input, using a lag time of 5 ns and using kinetic maps[175].

Conventional Markov state model. Markov state models based on discrete states were constructed as follows: time-structure independent component analysis was performed with inter-residue nearest-neighbour heavy atom distances as input, with a lag time of 5 ns. Various clustering methodologies were then evaluated using the 5-fold cross-validated generalized matrix Rayleigh quotient score using every 50th frame of the full trajectory (25183 frames) and a lag time of 12.5 ns[143]. Evaluation was performed over the number of time-lagged independent components, clusters, and algorithm (minibatch *k*-Means[100], minibatch *k*-Medoids, Gaussian mixture model) using MSMBuilder[176] (Supplementary Figure 7.1). 100 Markov models with the highest scoring parameters were then sampled and the timescales evaluated for the conventional Bayesian (Extended Data Figure 2.5 a)[95, 177], hidden Markov

[1] This relation can be derived from the probability $p_i(t) = K_{ii}^t$ of observing a state i for a duration t in continuous time by computing the expectation: $\bar{t} = \tau \int_0^\infty \mathbf{K}(\tau)_{ii}^t dt = \frac{-\tau}{\log\left(\mathbf{K}(\tau)_{ii}\right)}$

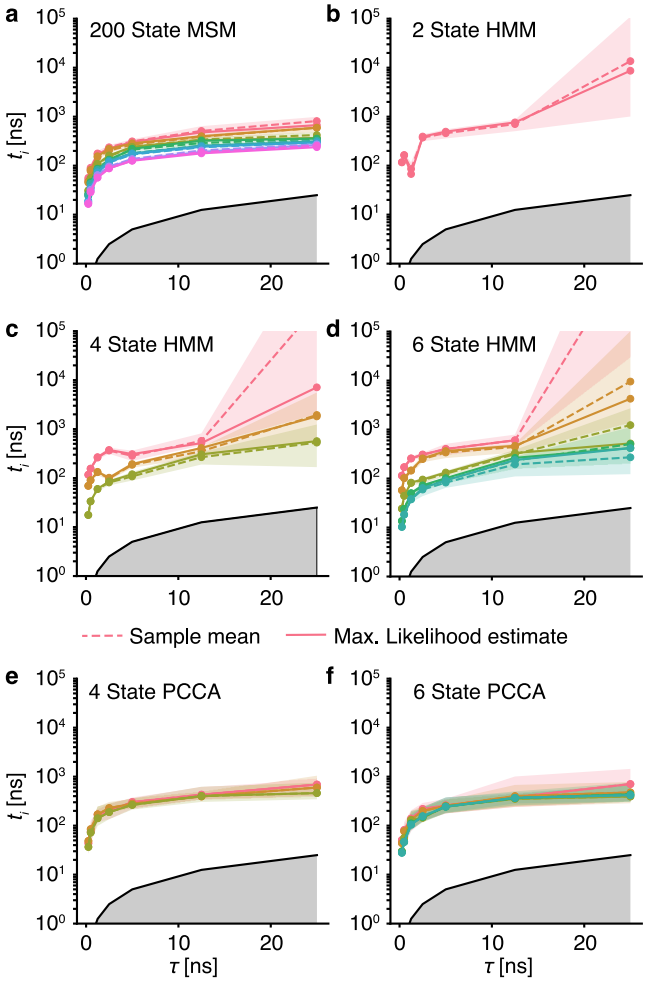
model (Extended Data Figure 2.5 b-d)[144] and Perron cluster-cluster analysis[145] (Extended Data Figure 2.5 e-f) cases.

Errors. Observable errors for each state were calculated by taking a trajectory sample based on the training data for a neural network sample. All statistics were then calculated on the bootstrap sample, and the errors reported in the figures show the 95th percentile of this sample unless noted otherwise.

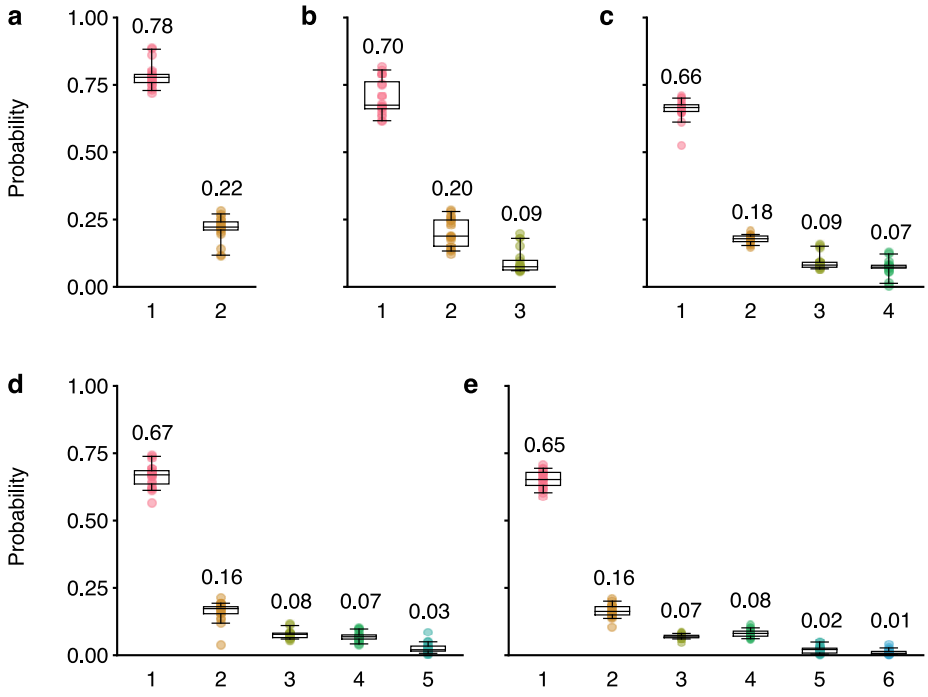
Data availability. Subsampled trajectory and intermediate data, as well as the trained neural network weights, analysis notebooks, and source data for figures 2-4 and extended data figures 1-4 are available from Zenodo.

Code availability. Analysis notebooks, code, and example data are available from GitHub and Zenodo.

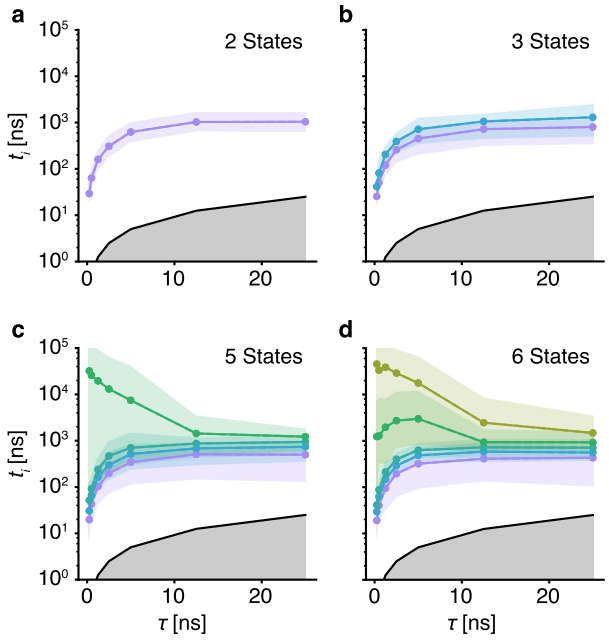
2.6 Extended data



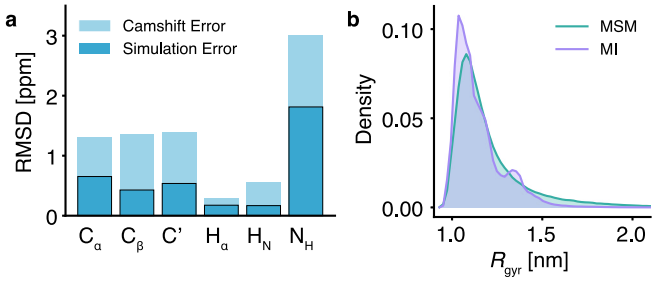
Extended Data Figure 2.5: Relaxation (implied) timescales for conventional discrete-state Markov state models, showing inability to construct a model with converging timescales. **a**, Relaxation timescales for a 200-microstate model as a function of model lag time. **b-d**, Relaxation timescales for hidden Markov state models using 2, 4, and 6 output states respectively, as a function of model lag time. **e-f**, Relaxation timescales for 4 and 6-state Markov state models built using Perron cluster-cluster analysis (PCCA) from the 200-microstate model as a function of model lag time. The gray shading indicates the timescales for which the Koopman model can no longer resolve the relaxation timescales. Shaded areas indicate 95 % confidence intervals of the sample mean of the 20 models.



Extended Data Figure 2.6: Equilibrium distributions of the models. **a**, 2, (**b**) 3, (**c**) 4, (**d**) 5, and (**e**) 6-state model equilibrium distributions. The whiskers, boxes and horizontal lines indicate 95th percentiles, quartiles, and the median values over all 20 models, respectively, the labels show the mean model values.



Extended Data Figure 2.7: Relaxation (implied) timescales as a function of model lag time. **a**, 2, (**b**) 3, (**c**) 5, and (**d**) 6-state model timescales. The gray shading indicates the timescales for which the Koopman model can no longer resolve the relaxation timescales. Shaded areas indicate 95th percentiles of the sample mean over 20 models.



Extended Data Figure 2.8: Experimental validation and comparison to an existing ensemble. **a**, Root-mean-square deviations between experimentally determined NMR chemical shifts and those back calculated using CamShift[147]; the deviations are smaller than the intrinsic CamShift errors. **b**, Comparison of the probability distributions of the radius of gyration computed for the current Markov state model (MSM, green) and the previously performed metadynamic metainference simulations (MI, purple)[33].

3 A small molecule stabilises the disordered native state of the Alzheimer's Aβ peptide

Time is a drug. Too much of it kills you.
— Terry Pratchett, in *Hogfather*

This chapter has been adapted from a manuscript currently in preparation for submission to a peer-reviewed journal. I designed the study, performed the analysis and wrote the manuscript. Kai Kohlhoff and I ran the simulations. Kai Kohlhoff, Gabriella Heller and Carlo Camilloni assisted with analysis. Gabriella Heller and I parameterized the small molecule. Michele Vendruscolo supervised the work.

3.1 Summary

The stabilisation of native states of proteins is a powerful drug discovery strategy. It is still unclear, however, whether this approach can be applied to intrinsically disordered proteins. Here we report a small molecule that stabilises the native state of the Aβ42 peptide, an intrinsically disordered protein fragment associated with Alzheimer's disease. We show that this stabilisation takes place by a highly dynamic binding mechanism, in which both the small molecule and the Aβ42 peptide remain disordered. We then characterise the forces responsible for this binding mechanism, revealing nanosecond lifetime π-stacking interactions as major contributors to the stabilisation of the bound state. These results indicate that the disordered binding

of small molecules and proteins requires transient non-specific interactions that provide enthalpic gain while simultaneously increasing the conformational entropy of the protein.

3.2 Introduction

Drug development for Alzheimer's disease has been distinguished by countless failures and setbacks in past decades[178]. Advances have primarily focused on the treatment of symptoms rather than the underlying mechanism. The illness is characterised by the formation of protein aggregates, such as fibrillar forms of amyloid-β 42 (Aβ42)[17, 18, 179]. This protein is intrinsically disordered, i.e. it does not form a single stable folded structure as a monomer, but instead exists in a dynamic equilibrium of states with transient local structure and fast transitions[33, 81, 122, 138–140, 149, 180, 181]. Many drug development efforts focused on aggregation-prone proteins such as Aβ42 attempt to target the already-formed fibril, the structurally elusive oligomeric species[21, 182, 183], or attempt to bind the monomer into a single stable conformation[31]. In contrast, there have been few efforts to design drugs directly targeting monomeric disordered proteins[22, 23]. The lack of success in this area is partially due to the difficulty in characterizing the binding mode on an atomistic level. While some experimental methods such as nuclear magnetic resonance can provide sparse information, it is often not sufficient to clearly understand the interactions and kinetics underlying the binding[23].

Molecular dynamics is one of the few tools that can provide the necessary spatial and temporal resolution to study the interaction between disordered proteins and small molecules[23]. Together with Bayesian restraints from experimental data, simulation has been used to thermodynamically characterize these binding modes in the case of the oncoprotein c-Myc[30] and Aβ42[33]. In the former study, urea was used as a control molecule to assess the sequence-specificity of the drug. In the latter case of Aβ42, we studied the interaction with the small molecule 10074-G5, and showed it was able to inhibit Aβ42 aggregation in various models. In both cases the binding mode was found to be highly dynamic, a quantitative study of the kinetics was however not possible. The microscopic kinetics in form of contact lifetimes and autocorrelations can be especially instructive to fully understand the origin of entropic and enthalpic stabilisation (Figure 3.1)[22]. The binding of small

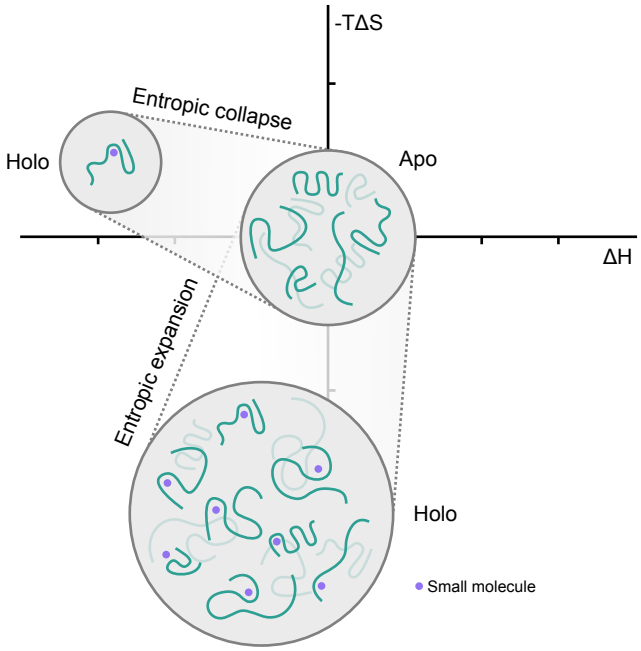


Figure 3.1: Native state stabilisation of disordered proteins. The interaction with a small molecule can result in a reduction or increase of conformational space of the protein, thus resulting in a positive or negative entropic contribution to the binding free energy. A loss of entropic native state stabilisation will often be compensated with a stronger enthalpic binding affinity, while an increase in conformational entropy often requires more dynamic and thus weaker binding.

molecules to monomeric disordered proteins was also explored for the case of the Parkinson's disease related α-synuclein using long-timescale molecular dynamics simulations[184]. The authors found that the interactions were predominantly driven by π-π stacking, in a process they refer to as 'dynamic shuttling'.

A quantitative study of the kinetics of these interactions may allow a more targeted approach to the design of both drugs and better experiments to probe their binding modalities. However, even with atomistic computational approaches, gaining insight into the kinetics, i.e. transition rates, relaxation constants, autocorrelations, and state lifetimes can be challenging. This is because in contrast to folded systems, the definition of states for disordered proteins is not always clear: due to the generally shallow free energy landscape state transitions may be fast, but not always

distinct[122]. New developments in the theory of dynamical systems now allow an optimal state decomposition and transition operator to be 'learned' using deep neural networks, for example using the VAMPNet framework[103, 106]. To acquire kinetic information for a system one would traditionally use a Markov state model[129, 130]: One first finds a suitable low-dimensional embedding of the system coordinates, followed by using a clustering algorithm to define microstates. Transitions between these can then be counted to build up statistics and thus construct a transition matrix. This matrix can then be coarse-grained to obtain macroscopic kinetics[144, 185].

Koopman-operator[94, 104] based models such as VAMPNet combine these two steps into a single function that can be approximated by a neural network and also yield a probabilistic state assignment in lieu of a discrete one[103, 106]. The former feature has the advantage of both simplifying the model construction process, as the hyperparameter search over various dimensionality reduction and clustering techniques is replaced by a simplified search over neural network parameters, and allowing a more accurate model due to the use of a single arbitrarily non-linear function compared to two steps that are heavily restricted in terms of search space. Probabilistic state assignments are inherently well suited to disordered proteins, as the typically shallow free energy basins and low barriers can be encoded with some ambiguity. This constrained VAMPNet approach was recently utilized by us to determine the kinetic ensemble of the disordered Aβ42 monomer (Chapter 3).

Here, we use this technique to build kinetic ensembles of Aβ42 with 10074-G5 and urea as a control molecule to expand on our previous thermodynamic ensembles[33]. We compare the transition rates, lifetimes, and state populations with the previous kinetic ensemble of the Aβ42 monomer (Chapter 3), and further characterize the atomic-level protein-drug interactions.

3.3 Results

Molecular dynamics simulations and soft Markov state models. We performed two explicit-solvent molecular dynamics simulations of Aβ42 with one molecule of urea and one molecule of 10074-G5, respectively. Both simulations were performed in multiple rounds of 1,024 trajectories on the Google Compute Engine as described previously[122]. As before we used a soft Markov state model approach using the constrained VAMPNet framework[106] to construct kinetic ensembles. The major

advantages of this method, compared to regular discrete Markov state models, are the soft state definitions and the use of a single function mapping directly from arbitrary system coordinates to a state assignment probability, allowing for more optimal models. To aid our analysis, we added our previous simulation of Aβ42 with no additional molecules to our dataset, we refer to it as the *apo* ensemble[122]. We compared all ensembles using a decomposition into two states. This allows for an easier evaluation and comparison of the slowest timescales in contrast to higher state-count models.

Computational and experimental validation. Constructing a kinetic ensemble using the constrained VAMPNet approach requires choosing the number of states and the model lag time. The latter is a critical parameter that needs to be chosen such that the model can accurately resolve both long and short timescales. This can be done by plotting the dependence of the slowest relaxation timescales on the lag time (Extended Data Figure 3.5). A stricter measure is the Chapman-Kolmogorov test, comparing multiple applications of the Koopman operator estimated at a certain lag time τ with a Koopman operator estimated at a multiple of this lag time $n\tau$ (Extended Data Figure 3.6, Methods)[93]. To evaluate sampling convergence, we visualized the dependence of the mean relaxation timescales on the number of trajectories used to evaluate these timescales (Extended Data Figure 3.9). With sufficient sampling of kinetics, we would expect the global timescales to be unchanged within error. Experimental validation was performed by comparing back-calculated chemical shifts to ones from experiment[147]. Because the small molecule 10074-G5 has no effect on the chemical shifts of Aβ42[30], we used the chemical shift dataset from the apo ensemble as a point of comparison (Extended Data Figure 3.7).

10074-G5 has minor impact on ensemble-averaged structural properties of Aβ42. To evaluate the influence of 10074-G5 and urea on the structural conformations of Aβ42, we calculated state-averaged contact maps and secondary structure content for each state of all ensembles (Extended Data Figure 3.8A-C). In all cases we find a state decomposition into a more extended state with few inter-residue contacts, and a slightly more compact form with a higher number of local backbone interactions. We will refer to these as the compact and extended states, respectively. The addition of a small molecule has little effect on the formation of contacts and other structural motifs. This finding is consistent with our recent experimental thermodynamic and

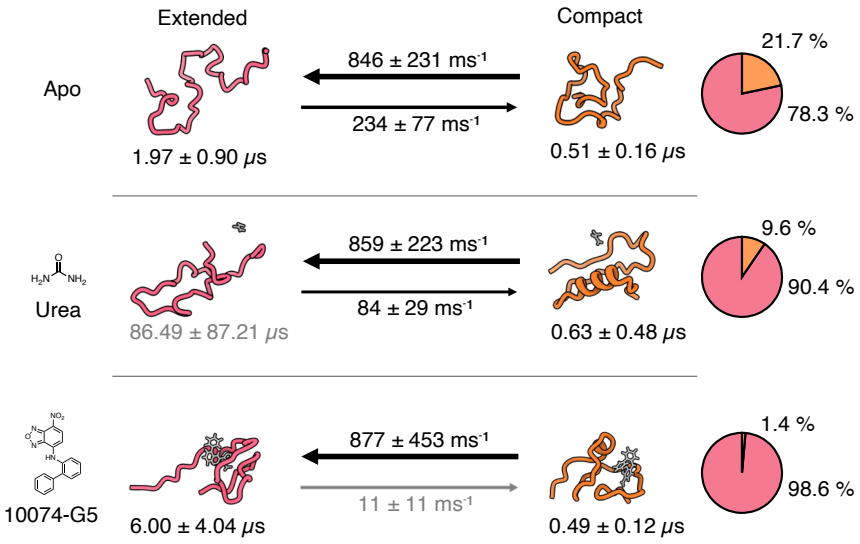


Figure 3.2: Impact of small molecules on the state transition rates, state lifetimes, and populations. The arrows indicate the mean state transition rates, the number below the representative structures is the mean state lifetime, the pie charts show the mean state populations. Errors are the standard deviations of the bootstrap sample of the mean over all 20 models.

kinetic characterization of this interaction, and the absence of strong chemical shift perturbations in the holo ensemble[33].

Small molecules decelerate the formation of more compact states. Compared to the previously published apo ensemble, the kinetic ensembles with both urea and 10074-G5 show a deceleration of more compact state formation (Figure 3.2). Notably, the transition from the more compact form to the more extended state is unaffected. This change is also mirrored in the state populations, which exhibit a strong shift towards the extended state. We note that even though there are strong changes in the state populations, the ensemble-averaged contact maps are very similar (Extended Data Figure 3.8A-C). This is likely due to the high sensitivity of the VAMPNet method to minor changes in free energy barrier regions. These will have a significant effect on the kinetics and thus state populations, but not on the ensemble averaged structure due to the relatively low thermodynamic weight[186]. While the lifetimes of the extended states increase, the ones for the more compact form are unchanged within

model error. We can thus conclude that within our model, the small molecule has a strong effect on the contact-formation rates, but no influence on the reverse.

Small molecules shift the system to more entropically stable states by short-lived local interactions. To evaluate the impact of 10074-G5 on the conformational space of Aβ42, we calculated the Ramachandran and state entropy for all ensembles, as well as the autocorrelation of sidechain χ_1 dihedral angles (Figure 3.3). The Ramachandran entropy can indicate relative flexibility of the backbone, thus revealing potential regions of dynamic changes as a result of interactions between the peptide and small molecule[33]. Resolving this change in the entropy over residues (Figure 3.3A) indicates strong increases in the relatively hydrophobic C-terminal region of Aβ42. This conformational entropy increase is confirmed globally by the sum of the entropies over all residues (Figure 3.3B). As an alternative metric, we also calculated the entropy in the state assignments (Figure 3.3C), this can be thought of as indicating the overall ambiguity in the state definition. Again, we find a relatively strong increase in the conformational entropy of Aβ42 for the ensemble with 10074-G5, and only minor increases for urea. These results are in strong agreement with our previous observations from simulations of the equilibrium (non-dynamic) ensembles in that the presence of 10074-G5 increases the conformations available to Aβ42, via the 'entropic expansion' mechanism[33].

To better understand the impact of the small molecule on local kinetics we calculated the autocorrelation of the sidechain χ_1 dihedral angles (Figure 3.3). We see an increase in the autocorrelation, specifically for aromatic residues and MET35, indicating a slowing of side chain rotations. This suggests that despite an increase in the backbone entropy, the peptide is able to visit many locally stable states, resulting in local enthalpic stabilisation.

Interactions of 10074-G5 with Aβ42 are dominated by π-stacking and other electrostatic effects. To better understand the origin of the global and local effects of the small molecule on the ensemble we analysed the interactions on a residue and atomistic level (Figure 3.4). While the probability of forming a contact between the small molecule and a residue shows certain mild preferences (Figure 3.4A), these become more evident when looking at the lifetimes of these contacts (Figure 3.4B). Here, the longest contacts are formed by π-stacking with certain aromatic residues (F4, Y10, F19, F20) and by interactions with MET35. This result also explains the

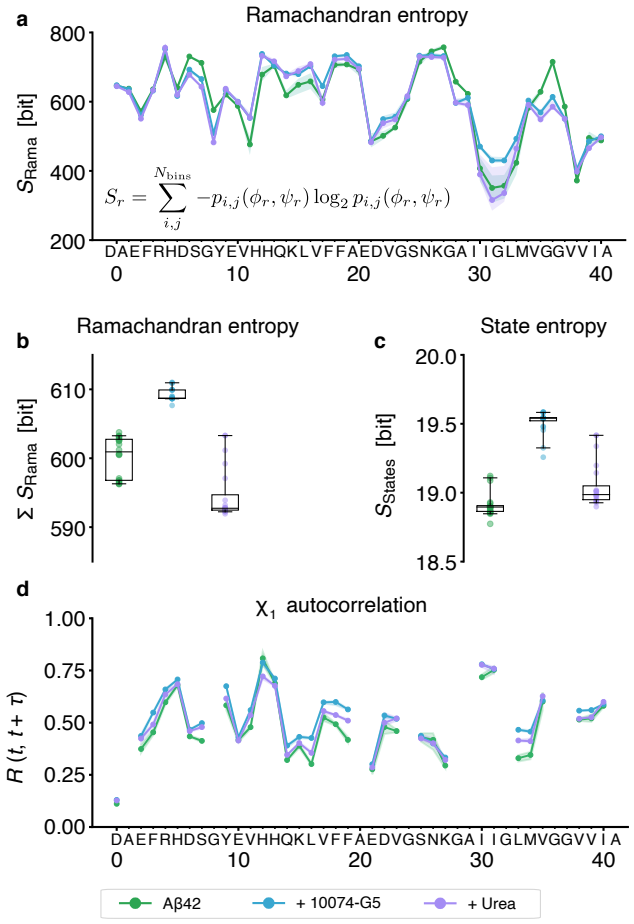


Figure 3.3: Effect of small molecules on conformational and state entropy of Aβ42, showing that 10074-G5 increases the conformational entropy of the peptide. **a** Ramachandran entropy, i.e. information entropy over the distribution of ϕ and ψ backbone dihedral angle conformations, using 100 bins. **b** Sum of the Ramachandran entropies over all residues for all ensembles. **c** State entropy, i.e., the population-weighted mean of the information entropy of each set of state assignments. More ambiguity in the state assignments leads to a correspondingly higher state entropy. **d** Autocorrelation of all sidechain χ_1 dihedral angles with a lag time of $\tau = 5$ ns. Shaded areas in **a** and **d** indicate the 95th percentiles of the bootstrap sample of the mean over all 20 models. Whiskers and boxes in **b** and **c** indicate the 95th percentiles and quartiles of the bootstrap sample of the mean over all 20 models, respectively.

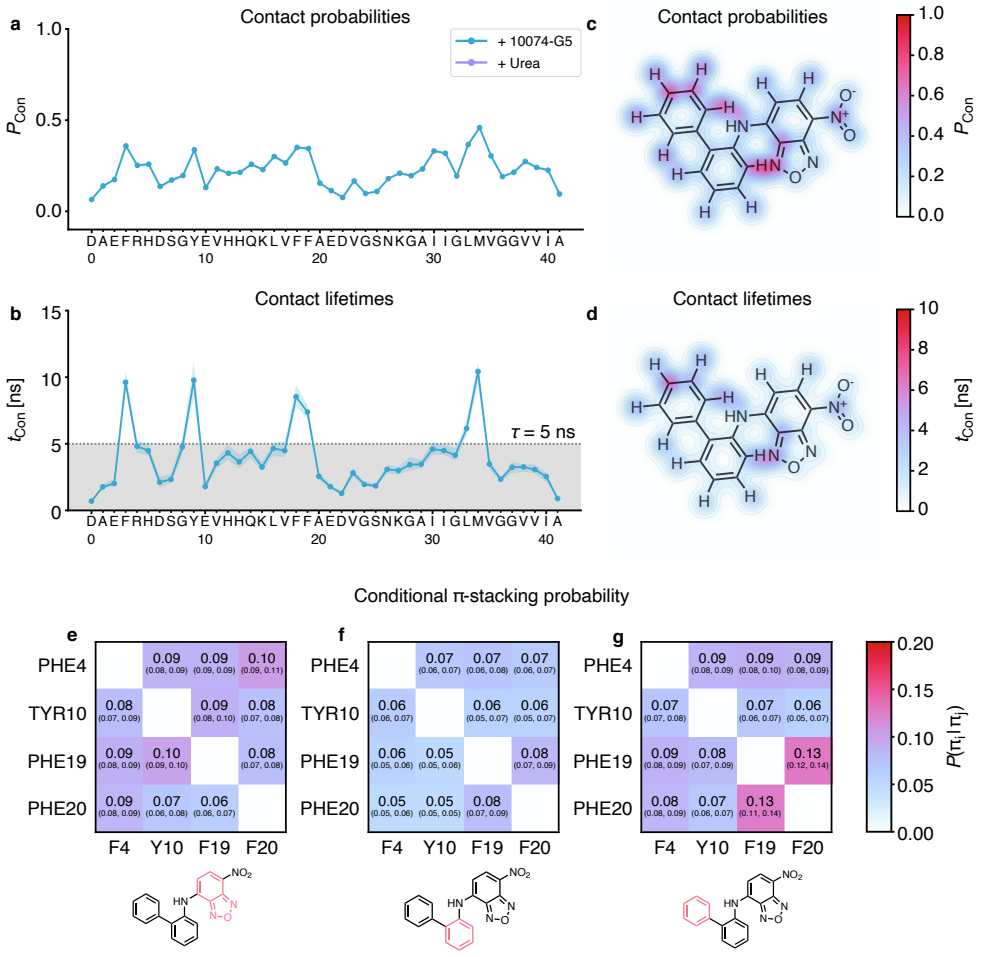


Figure 3.4: Residue- and atomic level interactions of 10074-G5 with Aβ42 showing regions on the small molecule responsible for binding. **a** Contact probabilities of 10074-G5 and Aβ42 with a cut-off of 0.45 nm. **b** Lifetimes of these contacts, estimated using a Markov state model for each contact formation with a lag time of $\tau = 5$ ns, indicated with grey shading. Coloured shaded areas in **a** and **b** indicate the 95th percentiles of the bootstrap sample of the mean over all 20 models. **c**, **d** Contact probabilities and lifetimes of each atom of 10074-G5 with any residue of Aβ42. **e**, **f**, **g** Conditional probability of forming a π-stacking interaction, given an existing π-stacking interaction for all aromatic groups in the small molecule. Tuples indicate the 95th percentiles of the bootstrap sample of the mean over all 20 models.

reduction in side-chain rotations for these residues (Figure 3.3D). On an atomistic level the π-π interactions exhibit some anisotropy (Extended Data Figure 3.10), however fully resolving these interactions is beyond the capabilities of the force field. The importance of the nitro- and benzofurazan fragments is also highlighted. Finally, we also investigated the conditionality of π-π interactions, i.e., if we see an interaction between the molecule and residue i, what is the probability of also observing an interaction with residue j (Figure 3.4E-G)? The probabilities here are uniformly low but indicate a slight preference (13 %) for a triple π-stack involving the terminal aromatic ring of 10074-G5 and residues F19 and F20 of Aβ42. The importance of π-π stacking interactions was also noted in a computational study on the interactions of small molecules with α-synuclein[184].

These results indicate that this disordered binding mechanism operates on two levels: local enthalpically favourable interactions coupled with global entropically advantageous effects. The local interactions are predominantly of electrostatic nature and result in a reduction of sidechain rotations on specific residues. At the same time, these interactions also allow the exploration of more backbone conformations, thus resulting in a net conformational entropy increase for Aβ42. This influence expands into the global kinetics of the system, significantly slowing the formation of local structure.

3.4 Discussion

The results outlined above present a possible example of the previously proposed entropic expansion mechanism for the binding of small molecules to disordered proteins[22, 187]. This idea stands in contrast to the heavily explored theme of entropic collapse or folding-on-binding mechanism[13, 188]. The concept of disordered binding is notoriously difficult to probe, as the tools suitable to detecting small changes in the conformational ensemble of disordered proteins are lacking[23]. Nuclear magnetic resonance experiments can potentially provide sparse information, but it must usually be interpreted in a structural framework, necessitating molecular simulations with ensemble-averaged restraints[125], or re-weighting approaches[114]. This constraint causes issues whenever we're also interested in kinetics, as we are no longer sampling the Boltzmann distribution. Nevertheless, an approach to incorporate ensemble-averaged experimental measurements into Koopman models

has recently been proposed[116]. Neither is it generally possible to use enhanced sampling methods to study kinetics without having some *a priori* knowledge of the system states. A framework allowing the incorporation of experimental data into a kinetic model and also allowing the use of enhanced sampling methods such as metadynamics[90], without prior knowledge of states, would make the study of these systems easier and more accurate.

As we have shown, a kinetic model is crucial to fully explain the nature of these binding interactions. This is in part due to the ability to use the slowest timescales of the system to reliably define metastable states, something that is notoriously difficult for disordered proteins without access to the time dimension. This clustering alone is already sensitive enough to reveal differences between systems that are nearly invisible when comparing ensemble-averaged results and more conventional clustering methods[33]. Increases in local autocorrelation and global state transitions might be seen as indicators of both local enthalpic stabilisation and global entropic expansion. The former result hints at the possibility of designing small molecules that exhibit high specificity, as the global entropic stabilisation effect may be due to transient, local, enthalpically-favourable interactions[22]. The two level global conformational entropy – local enthalpy effect becomes especially visible when looking at the timescales: The protein's slowest state transitions are on the order of microseconds, while the local, enthalpically-favourable π-π interaction lifetimes are no longer than tens of nanoseconds.

The observed binding mechanism also throws a spotlight on π-π stacking interactions as a major driving force. Similar effects have been observed for small molecules and the disordered α-synuclein protein[184]. π-π stacking also plays a major role in liquid-liquid phase separation[189], suggesting a possible link between the effect of these small molecules and the hypothesized state of some proteins in a crowded environment. For molecular simulations, the force field may present a barrier in studying π-π interactions in more detail. This is because these interactions are not explicitly part of the potential, but only approximated with a combination of electrostatic and hydrophobic terms[190]. Polarizable force fields may offer a computationally affordable alternative that could more accurately model this type of binding[107].

3.5 Methods

Simulation details. All simulations were performed on the Google Compute Engine with `n1-highcpu-8` virtual machine instances, equipped with eight Intel Haswell CPU cores and 7.2 GB of RAM. Molecular dynamics simulations were performed with GROMACS 2018.1[162], with 1024 starting structures sampled from the previously performed apo simulation[122] using the Koopman model weights. Each conformation was placed in the center of a rhombic dodecahedron box with a volume of 358 nm^3, and the corresponding small molecule was placed in the corner of the box. The force field parameters for urea were taken from the CHARMM22* force field[165] and the ones for 10074-G5 were computed using the Force Field Toolkit (FFTK)[191] and Gaussian 09[192], as described previously[33]. The systems were then solvated using between 11698 (11707) and 11740 (11749) water molecules. Both systems were minimized using the steepest descent method to a maximum target force of 1,000 kJ/mol/nm. Both systems were subsequently equilibrated, first over a time range of 500 ps in the NVT ensemble using the Bussi thermostat[163] and then over another 500 ps in the NPT ensemble using Berendsen pressure coupling[164]. In both equilibrations position restraints were placed on all heavy atoms. All production simulations were performed using 2 fs time steps in the NVT ensemble using the CHARMM22*[165] force field and TIP3P water model[166] at 278 K and LINCS constraints[167] on all bonds. Electrostatic interactions were modelled using the Particle-Mesh-Ewald approach[87] with a short-range cutoff of 1.2 nm. We again used the fluctuation-amplification of specific traits (FAST) approach[118] to adaptive sampling, with clustering performed through time-structure independent component analysis (TICA)[96, 97] using a lag time of 5 ns and Cα distances fed to the *k*-means clustering algorithm[100] to yield 128 clusters. 1,024 new structures were then sampled from these clusters based on maximizing the deviation to the mean Cα distance matrix for each cluster and maximizing the sampling of the existing clusters, using a balance parameter of $\alpha = 1.0$, with all amino acids weighted equally. This approach was performed once for each ensemble, however we also chose to perform 32 additional long-trajectory simulations for the 10074-G5 ensemble, yielding a total of 2,079 trajectories for the latter, and 2048 trajectories for the urea ensemble. The total simulated times were 306 µs and 279 µs for the 10074-G5 and urea ensembles, respectively. The shortest and longest trajectories for 10074-G5 (urea) were 21 (24)

ns and 1134 (196) ns. All trajectories were subsampled to 250 ps timesteps for further analysis.

Neural network. State decomposition and kinetic model construction was performed using the constrained VAMPNet approach[103, 106], using the same method described previously[122]. We again chose flattened inter-residue nearest-neighbour heavy-atom distance matrices as inputs, resulting in 780 input dimensions. We used the self-normalizing neural network architecture[146] with scaled-exponential linear units, normal LeCun weight initialization and alpha dropout. We chose an output dimension of 2, thus yielding a soft two-state assignment. The datasets were prepared by first creating a test dataset by randomly sampling 10 % of the frames. In the case of 10074-G5 we excluded all frames in which the closest distance between the small molecule and peptide was higher than 0.5 nm. We then created 20 randomized 9:1 train-validation splits to allow a model error estimate. Training was performed by using three trials for each train-validation split and picking the best performing model based on the VAMP2 score[104] of the test set. We implemented the model using Keras 2.2.4[169] with the Tensorflow 2.1.0[170] backend. We chose the following model hyperparameters based on two successive coarser and finer grid searches: A network lag time of 5 ns, a layer width of 512 nodes, a depth of 2 layers, an L_2 regularization strength of 10^{-7} and a dropout of 0.05. Training was performed in 10000 frame pairs using the Adam minimizer with a learning rate of 0.05, $\beta_2 = 0.99$ and epsilon of 10^{-4}, and an early stopping criterion of a minimum validation score improvement of 10^{-3} over the last five epochs. For the constrained part of the model, we reduced the learning rate by a factor of 0.02. We used a single Google Compute Engine instance with 12 Intel Haswell cores, 78 GB of RAM, and an NVidia Tesla V100 GPU.

Analysis. After training, VAMPNet yields a state assignment vector $\chi(\mathbf{x}_t)$ for each frame $\mathbf{x}_t$ of the ensemble. Based on this vector, we can calculate state averages $\langle A_i \rangle$ for any observable $A(\mathbf{x}_t)$:

$$\langle A_i \rangle = \left(\sum_{t=1}^{T} \chi_i(\mathbf{x}_t) \right)^{-1} \sum_{t=1}^{T} \chi_i(\mathbf{x}_t) A(\mathbf{x}_t) \tag{3.1}$$

Here, i is the corresponding state and the sum runs over all time steps. To calculate an ensemble average $\langle A \rangle$, one first calculates a weight w_t for each frame using the model equilibrium distribution π:

$$w_t = \frac{\langle \chi(\mathbf{x}_t) | \pi \rangle}{\sum_{t=1}^{T} \langle \chi(\mathbf{x}_t) | \pi \rangle}, \tag{3.2}$$

which leads to the ensemble average

$$\langle A \rangle = \sum_{t=1}^{T} w_t A(\mathbf{x}_t). \tag{3.3}$$

Because each trained model will classify the states in an arbitrary order, we need to sort the state assignment vectors based on state similarity. We did this by comparing the state-averaged contact maps using root-mean-square deviation as a metric, and grouping states based on the lowest value. Any deviations are thus accounted for in the overall model error.

Model validation. The Koopman matrix $\mathbf{K}(\tau)$ is given directly by the neural network model, along with the equilibrium distribution π. We validated our models using the Chapman-Kolmogorov test:

$$\mathbf{K}(n\tau) \approx \mathbf{K}^n(\tau) \tag{3.4}$$

where τ is the model lag time, and $n\tau$ is a low integer-multiple of the lag time. The model should therefore behave the same way whether we estimate it at a longer lag time or repeatedly apply the transfer operator. We first estimate a suitable lag time τ by plotting the relaxation timescales over the chosen lag time. The lag time τ should be chosen to be as small as possible, but large enough to not have any impact on the longer relaxation timescales, which represent the slowest motions of the system. The temporal resolution of the model is thus given by this lag time. The relaxation timescales t_i are calculated from the eigenvalues λ_i of the Koopman matrix $\mathbf{K}(\tau)$ as follows:

$$t_i = \frac{-\tau}{\log |\lambda_i|} \tag{3.5}$$

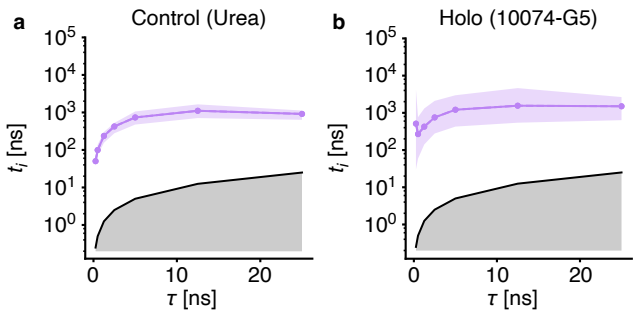
We can similarly compute the state lifetimes $\bar{t}_i$ from the diagonal elements of the Koopman matrix $\mathbf{K}(\tau)_{ii}$ using[172]:

$$\bar{t}_i = \frac{-\tau}{\log(\mathbf{K}(\tau)_{ii})} \tag{3.6}$$

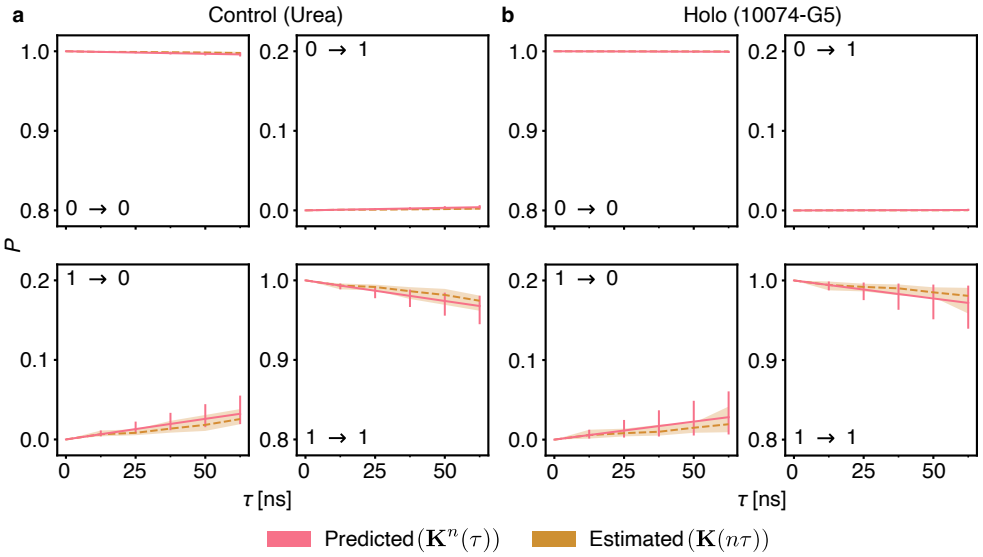
Experimental validation. We backcalculated the nuclear magnetic resonance chemical shifts using the CamShift algorithm[147] as implemented in PLUMED 2.4.1[173, 174]. We again used the same ensemble averaging procedure described above.

Errors. Errors are calculated over all trained neural network models. To obtain a more meaningful estimate, we only consider frames that were part of the bootstrap training sample of the corresponding model, i.e., one of the 20 models described above. The reported averages are the mean, and the errors the 95th percentiles over all 20 models, unless reported otherwise.

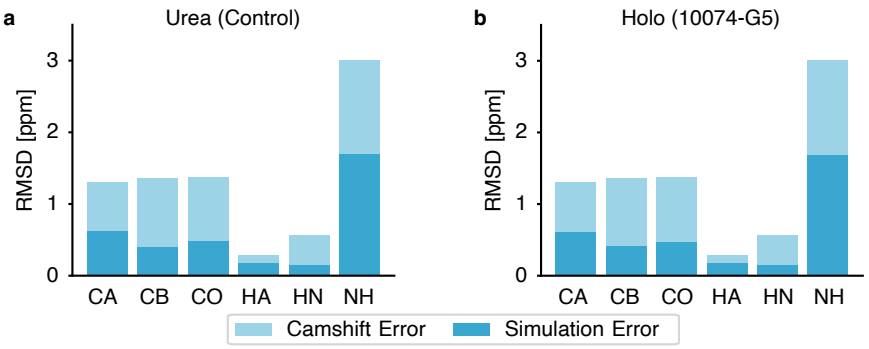
3.6 Extended data



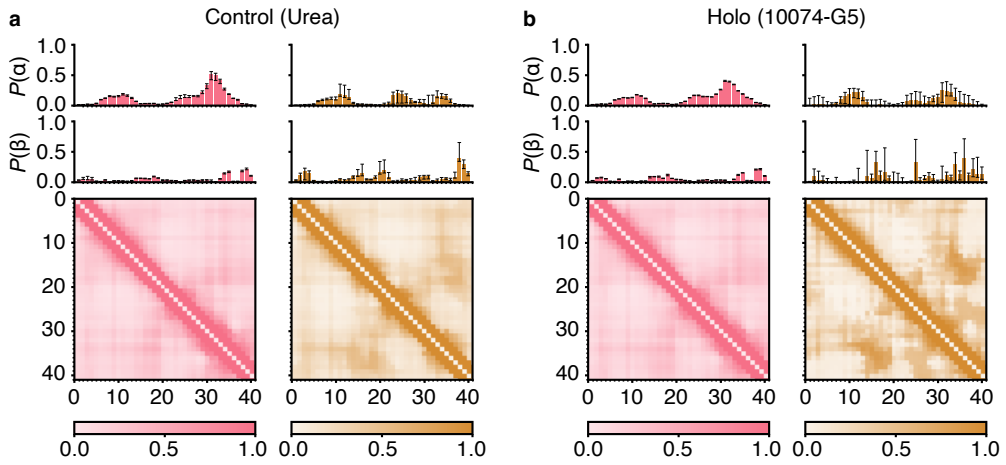
Extended Data Figure 3.5: Relaxation timescale as a function of model lag time for **a** control (urea) and **b** holo (10074-G5) ensembles. Gray shaded areas indicate timescales the Koopman model can no longer resolve. Coloured shaded areas indicate 95th percentiles of the sample mean over all 20 models.



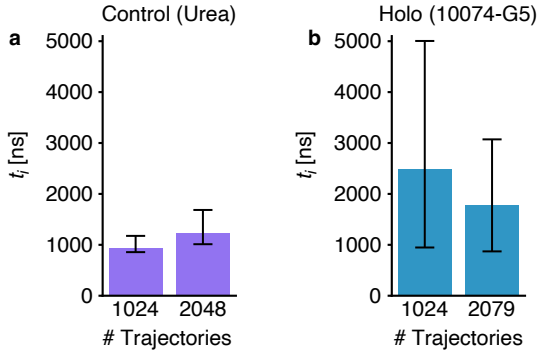
Extended Data Figure 3.6: Chapman-Kolmogorov test for **a** control (urea) and **b** holo (10074-G5) ensembles. Each panel indicates the transition probability for one matrix entry for successive applications (predicted, red) and estimations (estimated, orange) of the Koopman matrix. Shaded areas and error bars indicate 95th percentiles of the mean over all 20 models.



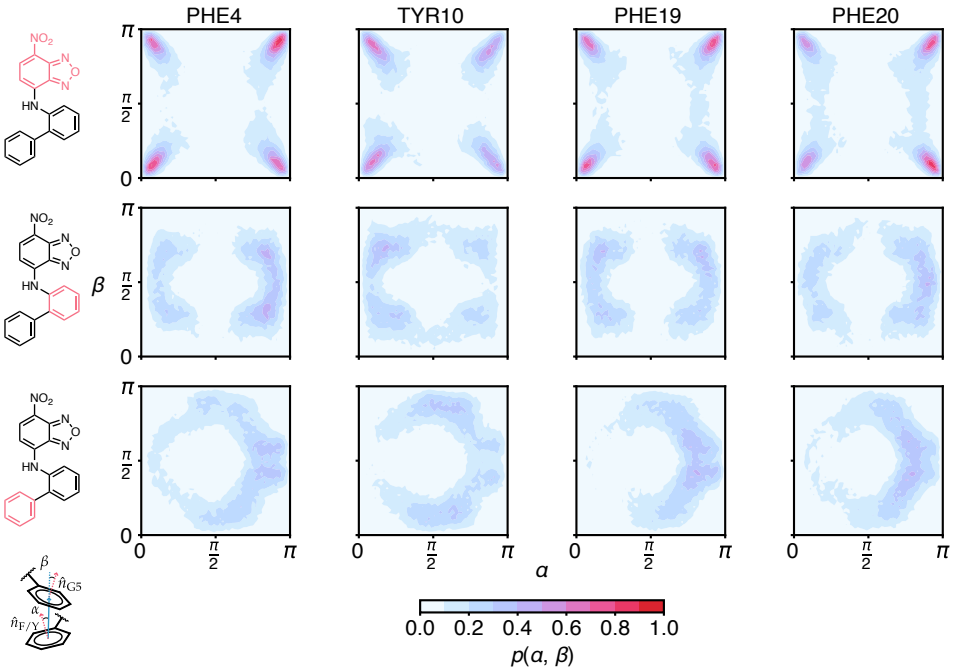
Extended Data Figure 3.7: Root-mean-square deviations between experimentally determined NMR chemical shifts[33] and those back-calculated using CamShift[147] for **a** control (urea) and **b** holo (10074-G5) ensembles. Light shaded areas indicate the intrinsic error of the CamShift predictor.



Extended Data Figure 3.8: Structural properties of the **a** control (urea) and **b** holo (10074-G5) ensembles. Top panels indicate α-helical and β-sheet contents over all residues as calculated using DSSP[193]. Bottom panels show heavy-atom contact probability maps with a cut-off of 0.8 nm. Error bars indicate 95th percentiles of the bootstrap sample of the mean over all 20 models.



Extended Data Figure 3.9: Dependence of the relaxation timescales on the number of trajectories used to build the model. Errorbars indicate 95th percentiles of the bootstrap sample of the mean over the first 5 models.



Extended Data Figure 3.10: Anisotropy of π-π molecule — aromatic side chain stacking interactions for the holo (10074-G5) ensemble. α is the stacking angle between the inter-aromatic distance vector and the aromatic side chain normal vector, while β is the angle inter-aromatic distance vector and the normal vector of the small molecule aromatic system[184]. Distributions show the density under the condition that the distance between both groups is below 0.6 nm.

4 Conformational entropy in a designed antibody

You fail 'cause you [take the] average
— Del the Funky Homosapien, *Press Rewind*

This chapter details ongoing work of understanding the behaviour of binding regions in antibodies designed using computational approaches. Pietro Sormanni and I designed the study. I carried out the simulations and performed the analysis. Pietro Sormanni provided the initial single-domain antibody structures. Michele Vendruscolo supervised the work.

4.1 Summary

In recent years, *in silico* antibody design has become a viable alternative to more traditional *in vivo* and *in vitro* approaches. However, the structural features of the complementarity determining region, specifically the role of rigidity and conformational entropy, are still unclear. We used enhanced-sampling molecular dynamics simulations to compare the free energy landscapes of single-domain antibodies designed using structure-based (VHH) and sequence-based approaches (DesAbO). Our results indicate that the CDRs of both VHH and DesAbO explore similar states, but that DesAbO is more conformationally heterogeneous. This difference underlines the challenges in the rational design of antibodies by revealing the presence of substates likely to have different binding properties and implicate a forced loss of entropic stabilisation upon binding.

4.2 Introduction

Designed antibodies have become essential tools in the fields of biological chemistry, medical diagnostics and therapeutics. Their ease of development has spawned numerous applications, for example in cancer treatment[194] and diagnostics[195], medication for autoimmune diseases[196] or lateral flow tests for the detection of SARS-CoV2[197].

Antibody production methodologies can be grouped into three general techniques. *In vivo* approaches utilise the immune system of a living animal to generate them; this idea has been extended to allow the use of *transgenic mice* to generate human antibodies[198, 199]. *In vitro* techniques typically use screening libraries to identify antibodies binding the desired target sequence with high affinity. However, isolating these identified antibodies is difficult, and their biophysical properties can be lacking compared to ones found with *in vivo* methods. One example of this approach is the use *phage display*[200] techniques.

These *in vivo* and *in vitro* methods become difficult if the target is only weakly immunogenic, as is often the case for disordered proteins and regions. Today, *in silico* approaches to antibody design can provide an attractive alternative[201] and circumvent this issue. Moving the time-consuming work of developing the correct sequence *in silico* can significantly accelerate the development time and allow a more efficient search of sequence space.

One such method was developed by Sormanni *et al.*[202]. The general idea is to identify a peptide with complementarity to an epitope on the target antigen. This peptide can then be grafted on to a suitable antibody scaffold as a CDR. The complementarity is achieved by mining the Protein Data Bank (PDB)[41] for β-strand conformations and identifying suitable fragments with part of the epitope sequence. By cascading along the sequence and identifying further fragments, the complementary peptide sequence can be constructed. This approach has been successfully used to design single-domain antibodies (sdAbs) targeting the elusive and mostly disordered oligomers of amyloid-β found in Alzheimer's disease[203]. Different epitopes were selected, and the correspondingly designed antibodies caused both a reduction in primary and secondary nucleation aggregation events, depending on the location of the epitope.

Despite these successful uses, the structural and dynamical features of designed sdAbs are still mostly unknown. Gaining a deeper insight into the behaviour of the

CDR in solution could enable further improvements in the computational design stage, yielding higher affinities and improved biophysical properties. In particular, it is unknown what role the conformational entropy plays in sdAbs ability to bind a potentially disordered target with high affinity. To better understand the dynamics of the CDR, I performed molecular dynamics simulations with enhanced sampling of an sdAb targeting amyloid-β oligomers (DesAbO)[203] – generated using the rational design approach outlined above – and one targeting human serum albumin (VHH) – using a the D3 scaffold, with the CDR created using a novel fragment-based structural design approach[204]. Both sdAbs thus represent examples of orthogonal design approaches, with potentially different conformational properties.

4.3 Results

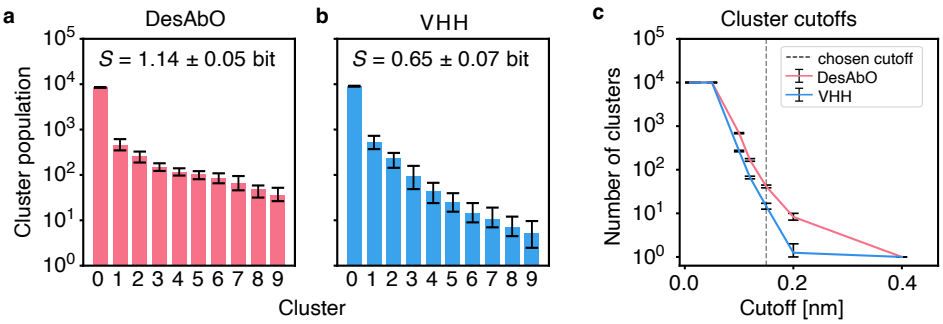


Figure 4.1: Results of the GROMOS clustering algorithm for different cut-off values. Populations for the top 10 clusters for DesAbO **a** and VHH **b** with the mean information entropy over the normalised populations indicated. **c** Number of found clusters for varying cut-off values. Error bars indicate the 95th percentile of the bootstrap sample-of-the-mean over all 20 samples consisting of 10000 frames sampled from the ensemble based on the metadynamics weights.

Metadynamics simulations of sdAbs. We performed all-atom, explicit water, parallel-bias metadynamics[92] simulations of DesAbO, designed using the rational sequence-based method[202], and VHH, built using a novel structural approach[204]. After 9.3 μs both simulations were found to be largely converged for cluster populations larger than 10 (out of 10000) frames (Figure 4.4).

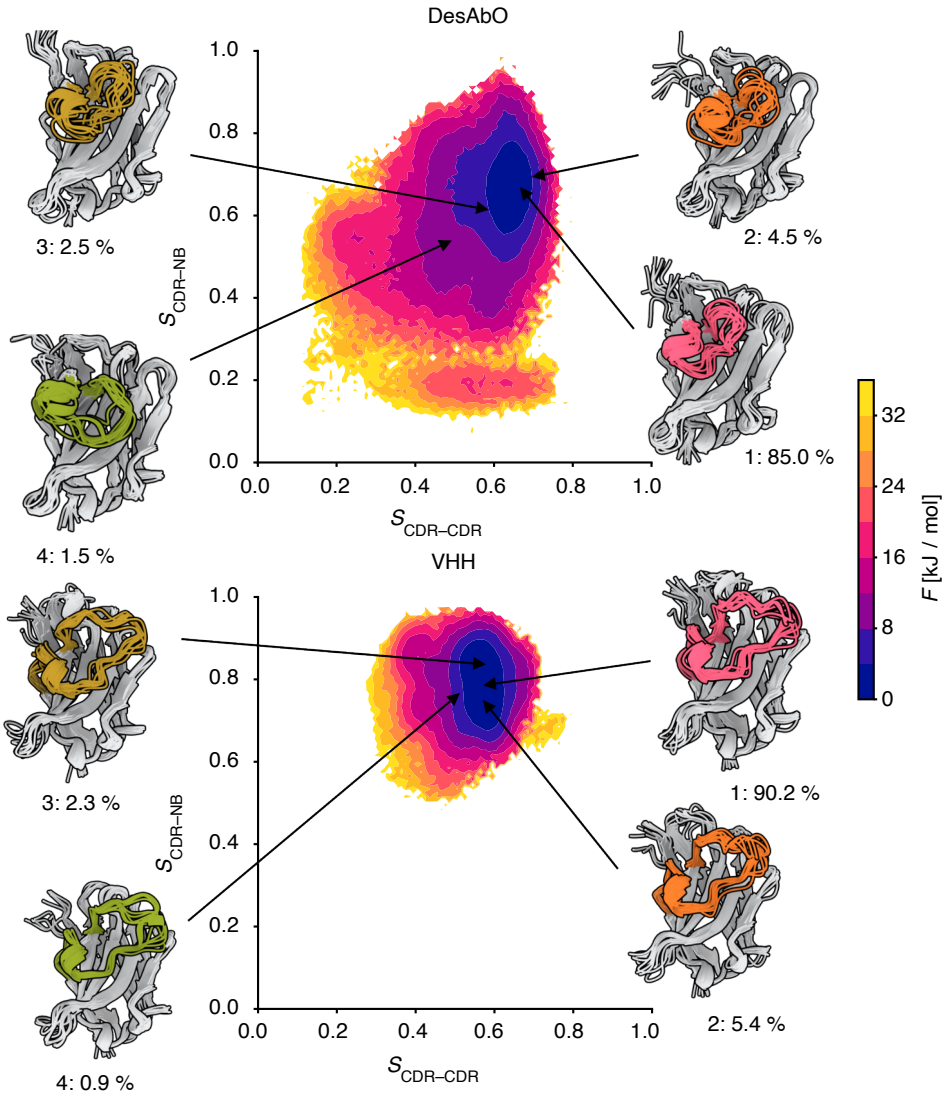


Figure 4.2: Free energy landscapes. The x- and y-axes represent the relative proportion of intra-CDR (CDR–CDR) and CDR-scaffold (CDR–NB) contacts respectively. 10 randomly sampled structures are shown from the top 4 clusters, with the CDR loop coloured and the cluster percentage of the ensemble indicated below.

Higher conformational entropy in a designed antibody. To evaluate conformational flexibility of the CDR, we performed a clustering analysis using the GROMOS algorithm[205] (Figure 4.1). We evaluated several different cut-off values for the Cα

root-mean-square deviation (RMSD) and found that the number of identified clusters was consistently higher in the DesAbO ensemble (Figure 4.1C). To obtain a reasonable number of clusters we chose a cut-off of 0.15 nm for both systems. The per-cluster population decays faster for the VHH ensemble (Figure 4.1B), indicating lower structural heterogeneity and a more compact conformational landscape in this naturally occurring antibody. Based on the normalized cluster populations we calculated the information entropy over all clusters (Figure 4.1A, B), again indicating higher conformational flexibility in the DesAbO ensemble.

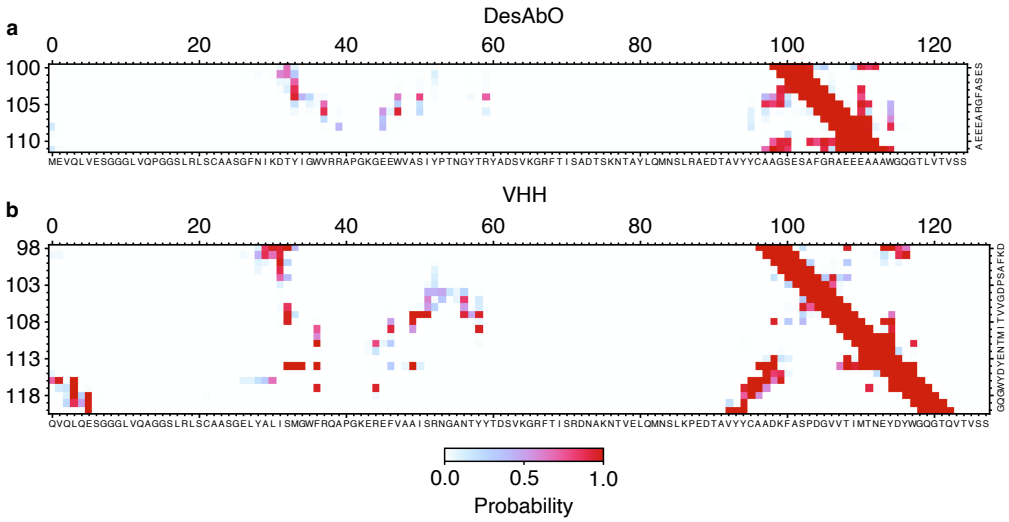


Figure 4.3: Structural properties of the sdAbs. Maps for **a** DesAbO and **b** VHH for CDR – scaffold contacts, indicating the probability of forming a contact between two residues.

Conformational flexibility in the complementarity determining region. To better understand the specific conformational changes of the rationally-designed sdAb (DesAbO) compared to the structurally-designed one (VHH), we projected the free energy on to collective variables encoding the number of intra-CDR and CDR-scaffold contacts (Figure 4.2). While the VHH ensemble is compact and maintains a large number of CDR-scaffold contacts, the DesAbO ensemble shows a loss of contacts not only with the scaffold, but also internally within the CDR. Even in the lowest energy state, the DesAbO system forms relatively fewer contacts with the scaffold than VHH. The origin of this higher flexibility in the sequence-based design

can be seen in a loss of contacts with residues 45-60 of the scaffold (Figure 4.3). We further see a reduction in the contact probability between the first part of the CDR and residues 30-35 in the scaffold, thus allowing for higher conformational heterogeneity.

4.4 Discussion

The simulations indicate clear structural differences between both antibodies on an ensemble level. Notably, despite the increased length of the VHH CDR theoretically allowing more flexibility, this CDR is in fact more structured, with stronger CDR-scaffold contacts. As the design process for this sdAb specifically optimises for structure, this is not necessarily surprising. However, both antibodies share many of the same contacts, with DesAbO only missing interactions between the first half of the CDR and residues 50 to 60 in the scaffold. The lack of these particular contacts may be sufficient to decrease the rigidity of the CDR, and potentially impact binding affinities. On the other hand, higher conformational entropy in the loop might be beneficial in binding disordered targets such as oligomers. In that case, the necessary structural rearrangements to form a β-sheet structure with the epitope might present a significant entropic barrier.

The effect of CDR-scaffold interactions might have to be taken into account in the sequence-based antibody design procedure. These interactions might even be tunable to make the formation of β-sheets easier, for example by arranging the residues of the CDR appropriately, or creating anchor points on either side of the CDR to force a particular arrangement. However, the general role of rigidity in antibody — antigen binding remains unclear, with some results indicating only a slight reduction in antibodies produced through affinity maturation compared to naïve antibodies[206], and others suggesting an increase in rigidity together with an increase in affinity[207]. Other studies hint at the role of water in the binding process and the entropically favourable formation of salt bridges[208, 209].

Comparing my results to an ensemble of a naturally occurring antibody could be especially instructive to understand the precise benefit of CDR-scaffold interactions (or a lack thereof) in general. Finally, experimental data on the interaction of the CDR loop with the scaffold, as well as detailed information on the binding mode would be

instructive. Unfortunately, elucidating the binding of multiple disordered proteins / regions – as would be required for an sdAb such as DesAbO – is notoriously difficult.

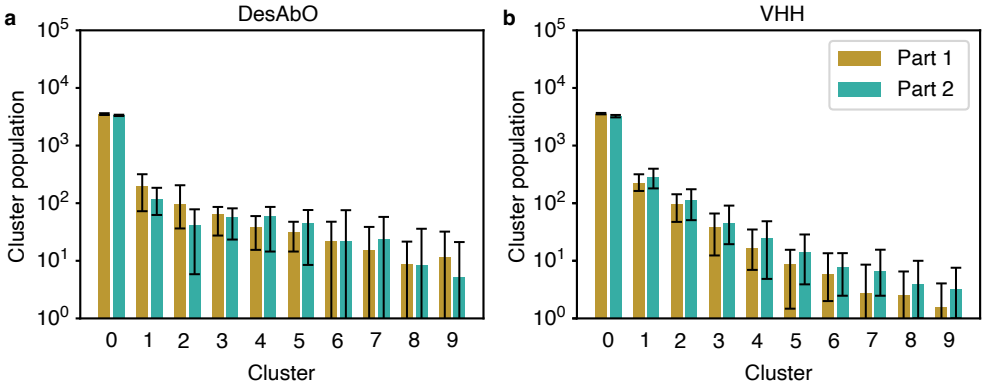
4.5 Methods

Simulation details. Both simulations were performed using GROMACS 2019.3 [162] and a development version of PLUMED 2.6[173, 174] (git commit 8859093). We chose CHARMM36m[210] as the force field, together with the TIP3P[166] water model. Starting conformations were created using MODELLER[211]. The structures were solvated in a rhombic dodecahedron box with a volume of 214 nm^3 (VHH: 197 nm^3) using 6726 (VHH: 6103) water molecules. Each system was minimized using the steepest descent algorithm to a target force of 1000 kJ / (mol / nm) and equilibrated over a time of 500 ps in the NVT ensemble with the Bussi thermostat[163], and over 5 ns in the NPT ensemble using Berendsen pressure coupling[164], while applying a position restraint on all heavy atoms, at a temperature of 300 K. The systems were then each simulated at 400 K for 5 ns in the NVT ensemble and 32 new starting structures were then sampled from the respective trajectories at random, to produce a set of diverse CDR conformations. Each conformation was then equilibrated using the same procedure outlined above, at a temperature of 300 K. Production simulations were performed at 300 K in the NPT ensemble using Parrinello-Rahman pressure coupling[212] with a timestep of 2 fs. Constraints were applied using the LINCS algorithm[167] with a matrix expansion on the order of 4 and 1 iteration per timestep. Modelling of electrostatic interactions was performed using the Particle-Mesh-Ewald[87] approach with a cut-off for short-range interactions at 1.2 nm. Both simulations were performed using Parallel-Bias metadynamics[92], using the well-tempered[213] and multiple-walkers[214] protocols with 32 replicas. Chosen collective variables were 1) the sum of all backbone dihedral angles (ALPHABETA action in PLUMED) in the CDR, 2) the sum of side chain χ_1 dihedral angles in the CDR, 3) the root-mean-square deviation of consecutive residues in the CDR to an ideal α-helix (ALPHARMSD in PLUMED), and 4) the root-mean-square deviation of consecutive residues in the CDR to an antiparallel β-sheet (ANTIBETARMSD in PLUMED). Gaussians were deposited every 500 steps (1 ps) using an initial height of 1.2 kJ / mol, a bias-factor of 24 and widths of 1) 1.5 rad, 2) 0.5 rad, 3) 2.0 and 4)

1.5. For both systems simulations were run for 293 ns per replica, totalling 9.3 µs of cumulative sampling.

Analysis. The individual replica trajectories were concatenated and the statistical weight for each frame was calculated using the approach by Torrie and Valleau[215]. All observables were calculated as weighted ensemble averages. Convergence was assessed by clustering 20 bootstrap samples of each trajectory and comparing the populations of each cluster in the case of the first and second halves of the respective simulations (Figure 4.4). We chose the GROMOS clustering algorithm[205] based on Cα RMSDs, as implemented in GROMACS 2019.3 with a cut-off of 0.15 nm based on the evaluation of several different values (Figure 4.1C). Convergence was assessed by clustering the whole trajectory using the same method, discarding the first 10 % of frames, and comparing the cluster populations between the remaining first and second halves of the simulation (Figure 4.4).

4.6 Extended data



Extended Data Figure 4.4: Simulation convergence for **a** the designed antibody DesAbO and **b** the VHH antibody. After discarding the first 20 % of frames, simulations were split into equal blocks and the cluster populations compared. Error bars indicate the 95th percentile of the bootstrap sample-of-the-mean over all 20 samples consisting of 10000 frames sampled from each block based on the metadynamics weights.

5 Conclusion

I have come to believe that the whole world is an enigma,
a harmless enigma that is made terrible by our own mad attempt
to interpret it as though it had an underlying truth.
— Umberto Eco, in *Foucault's Pendulum*

In the preceding chapters, I have outlined how one can use methods from dynamical systems theory to make sense of simulations of disordered proteins. In chapter 2 I performed long-timescale molecular dynamics simulations of amyloid-β 42 and constructed a kinetic ensemble using a probabilistic state definition, given by a neural network. This approach revealed state interconversion times on the range of microseconds, with the network of states forming an 'inverted kinetic hub'.

In chapter 3 I utilised the same methods to study the interaction of Aβ42 with a small molecule known to bind the monomeric form of the protein. While data on the thermodynamic behaviour of the interaction is already available – showing only minor changes to the ensemble upon binding[33] – detailed data on the atomistic kinetics was not. I discovered the presence of transient, highly localized π-π stacking interactions lasting only nanoseconds. The interaction however seems to increase the conformational entropy of the protein, suggesting both a favourable enthalpic as well as entropic effect.

Finally, in chapter 4 I departed from Aβ42 and utilized traditional enhanced sampling techniques to study the behaviour of complementarity determining regions (CDRs) – the binding motifs on (designed) single-domain antibodies (sdAbs). I compared an sdAb designed using a sequence-matching method with one utilising a

structural matching approach. The latter sdAb features a more compact free energy surface, with a less structurally heterogeneous CDR compared to the former sdAb.

What is the bottom line? The behaviour of disordered proteins is complex and presents new angles of attack for target identification, experiment design, and drug development. However, studying them at a highly resolved level presents significant difficulties to structural biology, and challenges assumptions underlying many modelling techniques. I will first summarize some of the limitations of my computational models before moving on to the context of their behaviour.

5.1 New models

An obvious limitation of the approach presented in chapters 2 and 3 is the use of pure molecular dynamics simulations to generate the data necessary to build kinetic ensembles. While approaches using ensemble-restrained simulations have become important tools in integrative structural biology[148, 216–218], their use is sadly not possible when one is also interested in kinetics. However, there have been developments to allow re-weighting methods to be used with kinetic models, notably augmented Markov models (AMMs)[115] and and an extension to VAMPNets[116]. Especially interesting is the option of using experimental data directly encoding kinetic properties such as relaxation times[80]. Beyond experimental data we have the option of moving to more resolving physical models. While quantum-mechanical approaches are an unrealistic option due to their large resource use, polarizable force fields[107] may present an attractive option to study the often subtle electrostatic interactions of disordered proteins.

There are also statistical considerations when using Markov models, our temporal and spatial resolution is limited by the lag time and state discretization. The former has recently been addressed in the form of history-augmented Markov state models (haMSMs)[186], allowing the use of very short lag times to build accurate models. The issue of limited state space can be argued to be beneficial, as this could somewhat aid interpretability. On the other hand, an ideal state definition is often difficult to achieve, and limited by the clustering procedure, the input data, and in the case of VAMPNet, the neural network architecture. The problem of how to effectively represent molecules and three-dimensional protein structures for neural networks has seen significant interest in the past few years, with graph-convolutional networks

and other methods showing great promise[158, 219, 220]. Generally, providing a more information-dense molecular representation to the network may substantially improve performance, while potentially also accelerating training times.

5.2 Universality

Chapter 2 raises the question of universality. Is an 'inverted kinetic hub' a general phenomenon seen in disordered proteins, or is Aβ42 an exception? α-synuclein, implicated in Parkinson's disease, has also enjoyed computational interest. A major problem in terms of simulation is the increased length of 140 residues, not only vastly increasing the required box size, but also the potential space of transient folds and other intramolecular interactions. Aβ42 in some way represents an excellent middle ground as a model system, as it is large enough to exhibit highly complex behaviour, but not too large to be unrealistic to simulate effectively. Another obvious step is studying the intermolecular behaviour of these aggregation-prone proteins. This poses the same problems as α-synuclein, as the increase in number of monomers interacting is similar to an increase in sequence length. The combinatorial complexity of these interactions can quickly get out of hand and require extremely long sampling times.

In chapter 3 I discussed the interaction of Aβ42 with a small molecule. We can once again ask if this interaction is unique to this particular combination, or if some more universal theme is at play. Simulations on α-synuclein suggest comparable mechanisms[184] with differences observed between small molecules, potentially allowing a structure-activity relationship to be established. The authors however also observed chemical shift perturbations, hinting at strong ensemble modulation – this is in contrast to our previous findings on Aβ42[33] and small molecules. Generally, previous findings on small molecule binding to disordered proteins[30, 33, 221, 222] hint at the possibility of achieving somewhat specific binding, and thus an attractive avenue to novel drugs.

Finally, in chapter 4 I studied the behaviour of complementarity determining regions (CDRs) in designed antibodies. Again, the question of universality is a given, but it is also as of yet unclear whether some flexibility could be beneficial to binding certain targets, or if high affinities can only be reached with a rigid binding motif and correspondingly strong enthalpic interactions. Exploring other types of sdAbs

would help to shed light on this phenomenon and could eventually aid in the design of novel therapeutics.

6 Bibliography

I do things like get in a taxi and say, "The library, and step on it."
— David Foster Wallace, in *Infinite Jest*

(1) Chouard, T. "Structural Biology: Breaking the Protein Rules". *Nature* **2011**, *471*, 151–153.

(2) Taylor, S. S.; Radzio-Andzelm, E. "Three Protein Kinase Structures Define a Common Motif". *Structure* **1994**, *2*, 345–355.

(3) Wang, B.; Yang, W.; McKittrick, J.; Meyers, M. A. "Keratin: Structure, Mechanical Properties, Occurrence in Biological Organisms, and Efforts at Bioinspiration". *Progress in Materials Science* **2016**, *76*, 229–318.

(4) Rimon, O.; Suss, O.; Goldenberg, M.; Fassler, R.; Yogev, O.; Amartely, H.; Propper, G.; Friedler, A.; Reichmann, D. "A Role of Metastable Regions and Their Connectivity in the Inactivation of a Redox-Regulated Chaperone and Its Inter-Chaperone Crosstalk". *Antioxidants & Redox Signaling* **2017**, *27*, 1252–1267.

(5) Pentony, M. M.; Jones, D. T. "Modularity of Intrinsic Disorder in the Human Proteome". *Proteins: Structure, Function, and Bioinformatics* **2010**, *78*, 212–221.

(6) Schad, E.; Tompa, P.; Hegyi, H. "The Relationship between Proteome Size, Structural Disorder and Organism Complexity". *Genome Biology* **2011**, *12*, R120.

(7) Babu, M. M.; van der Lee, R.; de Groot, N. S.; Gsponer, J. "Intrinsically Disordered Proteins: Regulation and Disease". *Current Opinion in Structural Biology* **2011**, *21*, 432–440.

(8) Metskas, L. A.; Rhoades, E. "Single-Molecule FRET of Intrinsically Disordered Proteins". *Annual Review of Physical Chemistry* **2020**, *71*, 391–414.

(9) Neuman, K. C.; Nagy, A. "Single-Molecule Force Spectroscopy: Optical Tweezers, Magnetic Tweezers and Atomic Force Microscopy". *Nature Methods* **2008**, *5*, 491–505.

(10) Wahl, M. C.; Will, C. L.; Lührmann, R. "The Spliceosome: Design Principles of a Dynamic RNP Machine". *Cell* **2009**, *136*, 701–718.

(11) Mackereth, C. D.; Madl, T.; Bonnal, S.; Simon, B.; Zanier, K.; Gasch, A.; Rybin, V.; Valcárcel, J.; Sattler, M. "Multi-Domain Conformational Selection Underlies Pre-mRNA Splicing Regulation by U2AF". *Nature* **2011**, *475*, 408–411.

(12) Borgia, A.; Borgia, M. B.; Bugge, K.; Kissling, V. M.; Heidarsson, P. O.; Fernandes, C. B.; Sottini, A.; Soranno, A.; Buholzer, K. J.; Nettels, D.; Kragelund, B. B.; Best, R. B.; Schuler, B. "Extreme Disorder in an Ultrahigh-Affinity Protein Complex". *Nature* **2018**, *555*, 61–66.

(13) Robustelli, P.; Piana, S.; Shaw, D. E. "Mechanism of Coupled Folding-upon-Binding of an Intrinsically Disordered Protein". *Journal of the American Chemical Society* **2020**, *142*, 11092–11101.

(14) Hyman, A. A.; Weber, C. A.; Jülicher, F. "Liquid-Liquid Phase Separation in Biology". *Annual Review of Cell and Developmental Biology* **2014**, *30*, 39–58.

(15) Krainer, G. et al. "Reentrant Liquid Condensate Phase of Proteins Is Stabilized by Hydrophobic and Non-Ionic Interactions". *Nature Communications* **2021**, *12*, 1085.

(16) Hardenberg, M.; Horvath, A.; Ambrus, V.; Fuxreiter, M.; Vendruscolo, M. "Widespread Occurrence of the Droplet State of Proteins in the Human Proteome". *Proceedings of the National Academy of Sciences* **2020**, *117*, 33254–33262.

(17) Hardy, J. A.; Higgins, G. A. "Alzheimer's Disease: The Amyloid Cascade Hypothesis". *Science* **1992**, *256*, 184–185.

(18) Hardy, J.; Allsop, D. "Amyloid Deposition as the Central Event in the Aetiology of Alzheimer's Disease". *Trends in Pharmacological Sciences* **1991**, *12*, 383–388.

(19) Knowles, T. P. J.; Vendruscolo, M.; Dobson, C. M. "The Amyloid State and Its Association with Protein Misfolding Diseases". *Nature Reviews Molecular Cell Biology* **2014**, *15*, 384–396.

(20) Knowles, T. P. J.; Waudby, C. A.; Devlin, G. L.; Cohen, S. I. A.; Aguzzi, A.; Vendruscolo, M.; Terentjev, E. M.; Welland, M. E.; Dobson, C. M. "An Analytical Solution to the Kinetics of Breakable Filament Assembly". *Science* **2009**, *326*, 1533–1537.

(21) Habchi, J.; Chia, S.; Limbocker, R.; Mannini, B.; Ahn, M.; Perni, M.; Hansson, O.; Arosio, P.; Kumita, J. R.; Challa, P. K.; Cohen, S. I. A.; Linse, S.; Dobson, C. M.; Knowles, T. P. J.; Vendruscolo, M. "Systematic Development of Small Molecules to Inhibit Specific Microscopic Steps of Aβ42 Aggregation in Alzheimer's Disease". *Proceedings of the National Academy of Sciences* **2017**, *114*, E200–E208.

(22) Heller, G. T.; Sormanni, P.; Vendruscolo, M. "Targeting Disordered Proteins with Small Molecules Using Entropy". *Trends in Biochemical Sciences* **2015**, *40*, 491–496.

(23) Heller, G. T.; Aprile, F. A.; Vendruscolo, M. "Methods of Probing the Interactions between Small Molecules and Disordered Proteins". *Cellular and Molecular Life Sciences* **2017**, *74*, 3225–3243.

(24) Iakoucheva, L. M.; Brown, C. J.; Lawson, J. D.; Obradović, Z.; Dunker, A. K. "Intrinsic Disorder in Cell-signaling and Cancer-associated Proteins". *Journal of Molecular Biology* **2002**, *323*, 573–584.

(25) Xu, Y.; Shi, J.; Yamamoto, N.; Moss, J. A.; Vogt, P. K.; Janda, K. D. "A Credit-Card Library Approach for Disrupting Protein–Protein Interactions". *Bioorganic & Medicinal Chemistry* **2006**, *14*, 2660–2673.

(26) Shi, J.; Stover, J. S.; Whitby, L. R.; Vogt, P. K.; Boger, D. L. "Small Molecule Inhibitors of Myc/Max Dimerization and Myc-induced Cell Transformation". *Bioorganic & Medicinal Chemistry Letters* **2009**, *19*, 6038–6041.

(27) Berg, T.; Cohen, S. B.; Desharnais, J.; Sonderegger, C.; Maslyar, D. J.; Goldberg, J.; Boger, D. L.; Vogt, P. K. "Small-Molecule Antagonists of Myc/Max Dimerization Inhibit Myc-induced Transformation of Chicken Embryo Fibroblasts". *Proceedings of the National Academy of Sciences of the United States of America* **2002**, *99*, 3830–3835.

(28) Yin, X.; Giap, C.; Lazo, J. S.; Prochownik, E. V. "Low Molecular Weight Inhibitors of Myc–Max Interaction and Function". *Oncogene* **2003**, *22*, 6151–6159.

(29) Hammoudeh, D. I.; Follis, A. V.; Prochownik, E. V.; Metallo, S. J. "Multiple Independent Binding Sites for Small-Molecule Inhibitors on the Oncoprotein c-Myc". *Journal of the American Chemical Society* **2009**, *131*, 7390–7401.

(30) Heller, G. T.; Aprile, F. A.; Bonomi, M.; Camilloni, C.; De Simone, A.; Vendruscolo, M. "Sequence Specificity in the Entropy-Driven Binding of a Small Molecule and a Disordered Peptide". *Journal of Molecular Biology* **2017**, *429*, 2772–2779.

(31) Sinha, S. et al. "Lysine-Specific Molecular Tweezers Are Broad-Spectrum Inhibitors of Assembly and Toxicity of Amyloid Proteins". *Journal of the American Chemical Society* **2011**, *133*, 16958–16969.

(32) Richter, F.; Subramaniam, S. R.; Magen, I.; Lee, P.; Hayes, J.; Attar, A.; Zhu, C.; Franich, N. R.; Bove, N.; De La Rosa, K.; Kwong, J.; Klärner, F.-G.; Schrader, T.; Chesselet, M.-F.; Bitan, G. "A Molecular Tweezer Ameliorates Motor Deficits in Mice Overexpressing α-Synuclein". *Neurotherapeutics* **2017**, *14*, 1107–1119.

(33) Heller, G. T. et al. "Small-Molecule Sequestration of Amyloid-β as a Drug Discovery Strategy for Alzheimer's Disease". *Science Advances* **2020**, *6*, eabb5924.

(34) Uversky, V. N. "Intrinsically Disordered Proteins and Their "Mysterious" (Meta)Physics". *Frontiers in Physics* **2019**, *7*, 10.

(35) Camilloni, C.; Bonetti, D.; Morrone, A.; Giri, R.; Dobson, C. M.; Brunori, M.; Gianni, S.; Vendruscolo, M. "Towards a Structural Biology of the Hydrophobic Effect in Protein Folding". *Scientific Reports* **2016**, *6*, 28285.

(36) Kim, T. D.; Ryu, H. J.; Cho, H. I.; Yang, C.-H.; Kim, J. "Thermal Behavior of Proteins: Heat-Resistant Proteins and Their Heat-Induced Secondary Structural Changes". *Biochemistry* **2000**, *39*, 14839–14846.

(37) Bak, P.; Tang, C.; Wiesenfeld, K. "Self-Organized Criticality: An Explanation of the $1/f$ Noise". *Physical Review Letters* **1987**, *59*, 381–384.

(38) Moret, M. A.; Zebende, G. F. "Amino Acid Hydrophobicity and Accessible Surface Area". *Physical Review E* **2007**, *75*, 011920.

(39) Phillips, J. C. "Scaling and Self-Organized Criticality in Proteins: Lysozyme c". *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics* **2009**, *80*, 051916.

(40) Vecchi, G.; Sormanni, P.; Mannini, B.; Vandelli, A.; Tartaglia, G. G.; Dobson, C. M.; Hartl, F. U.; Vendruscolo, M. "Proteome-Wide Observation of the Phenomenon of Life on the Edge of Solubility". *Proceedings of the National Academy of Sciences* **2020**, *117*, 1015–1020.

(41) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. "The Protein Data Bank". *Nucleic Acids Research* **2000**, *28*, 235–242.

(42) Bonomi, M.; Vendruscolo, M. "Determination of Protein Structural Ensembles Using Cryo-Electron Microscopy". *Current Opinion in Structural Biology* **2019**, *56*, 37–45.

(43) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wriggers, W. "Atomic-Level Characterization of the Structural Dynamics of Proteins". *Science* **2010**.

(44) Ban, D.; Smith, C. A.; de Groot, B. L.; Griesinger, C.; Lee, D. "Recent Advances in Measuring the Kinetics of Biomolecules by NMR Relaxation Dispersion Spectroscopy". *Archives of Biochemistry and Biophysics* **2017**, *628*, 81–91.

(45) Gauto, D. F. et al. "Integrated NMR and Cryo-EM Atomic-Resolution Structure Determination of a Half-Megadalton Enzyme Complex". *Nature Communications* **2019**, *10*, 2697.

(46) Lu, M.; Russell, R. W.; Bryer, A. J.; Quinn, C. M.; Hou, G.; Zhang, H.; Schwieters, C. D.; Perilla, J. R.; Gronenborn, A. M.; Polenova, T. "Atomic-Resolution Structure of HIV-1 Capsid Tubes by Magic-Angle Spinning NMR". *Nature Structural & Molecular Biology* **2020**, *27*, 863–869.

(47) Saura, P.; Röpke, M.; Gamiz-Hernandez, A. P.; Kaila, V. R. I. In *Biomolecular Simulations: Methods and Protocols*, Bonomi, M., Camilloni, C., Eds.; Methods in Molecular Biology; Springer: New York, NY, 2019, pp 75–104.

(48) Brosey, C. A.; Tainer, J. A. "Evolving SAXS Versatility: Solution X-ray Scattering for Macromolecular Architecture, Functional Landscapes, and Integrative Structural Biology". *Current Opinion in Structural Biology* **2019**, *58*, 197–213.

(49) Putnam, C. D.; Hammel, M.; Hura, G. L.; Tainer, J. A. "X-Ray Solution Scattering (SAXS) Combined with Crystallography and Computation: Defining Accurate Macromolecular Structures, Conformations and Assemblies in Solution". *Quarterly Reviews of Biophysics* **2007**, *40*, 191–285.

(50) Welborn, S. S.; Detsi, E. "Small-Angle X-ray Scattering of Nanoporous Materials". *Nanoscale Horizons* **2019**, *5*, 12–24.

(51) Chirio-Lebrun, M.-C.; Prats, M. "Fluorescence Resonance Energy Transfer (FRET): Theory and Experiments". *Biochemical Education* **1998**, *26*, 320–323.

(52) LeBlanc, S. J.; Kulkarni, P.; Weninger, K. R. "Single Molecule FRET: A Powerful Tool to Study Intrinsically Disordered Proteins". *Biomolecules* **2018**, *8*, 140.

(53) Merk, A.; Bartesaghi, A.; Banerjee, S.; Falconieri, V.; Rao, P.; Davis, M. I.; Pragani, R.; Boxer, M. B.; Earl, L. A.; Milne, J. L. S.; Subramaniam, S. "Breaking Cryo-EM Resolution Barriers to Facilitate Drug Discovery". *Cell* **2016**, *165*, 1698–1707.

(54) Zhao, G.; Perilla, J. R.; Yufenyuy, E. L.; Meng, X.; Chen, B.; Ning, J.; Ahn, J.; Gronenborn, A. M.; Schulten, K.; Aiken, C.; Zhang, P. "Mature HIV-1 Capsid Structure by Cryo-Electron Microscopy and All-Atom Molecular Dynamics". *Nature* **2013**, *497*, 643–646.

(55) Hirano, Y.; Takeda, K.; Miki, K. "Charge-Density Analysis of an Iron–Sulfur Protein at an Ultra-High Resolution of 0.48 Å". *Nature* **2016**, *534*, 281–284.

(56) Makino, D. L.; Larson, S. B.; McPherson, A. "The Crystallographic Structure of Panicum Mosaic Virus (PMV)". *Journal of Structural Biology* **2013**, *181*, 37–52.

(57) Baek, M. et al. "Accurate Prediction of Protein Structures and Interactions Using a Three-Track Neural Network". *Science* **2021**, *373*, 871–876.

(58) Jumper, J. et al. "Highly Accurate Protein Structure Prediction with AlphaFold". *Nature* **2021**, *596*, 583–589.

(59) Drenth, J., *Principles of Protein X-Ray Crystallography*, 3rd ed.; Springer-Verlag: New York, 2007.

(60) Yip, K. M.; Fischer, N.; Paknia, E.; Chari, A.; Stark, H. "Atomic-Resolution Protein Structure Determination by Cryo-EM". *Nature* **2020**, *587*, 157–161.

(61) Zivanov, J.; Nakane, T.; Forsberg, B. O.; Kimanius, D.; Hagen, W. J.; Lindahl, E.; Scheres, S. H. "New Tools for Automated High-Resolution Cryo-EM Structure Determination in RELION-3". *eLife* **2018**, *7*, ed. by Egelman, E. H.; Kuriyan, J., e42166.

(62) Bock, L. V.; Grubmüller, H. "Effects of Cryo-EM Freezing on the Structural Ensemble". *Biophysical Journal* **2018**, *114*, 160a.

(63) Carragher, B.; Cheng, Y.; Frost, A.; Glaeser, R.; Lander, G.; Nogales, E.; Wang, H.-W. "Current Outcomes When Optimizing 'Standard' Sample Preparation for Single-Particle Cryo-EM". *Journal of Microscopy* **2019**, *276*, 39–45.

(64) Glaeser, R. M.; Han, B.-G. "Opinion: Hazards Faced by Macromolecules When Confined to Thin Aqueous Films". *Biophysics Reports* **2017**, *3*, 1–7.

(65) Hovius, R.; Vallotton, P.; Wohland, T.; Vogel, H.; Hovius, R.; Vallotton, P.; Wohland, T.; Vogel, H. "Fluorescence Techniques: Shedding Light on Ligand–Receptor Interactions". *Trends in Pharmacological Sciences* **2000**, *21*, 266–273.

(66) Brucale, M.; Schuler, B.; Samorì, B. "Single-Molecule Studies of Intrinsically Disordered Proteins". *Chemical Reviews* **2014**, *114*, 3281–3317.

(67) Riback, J. A.; Bowman, M. A.; Zmyslowski, A. M.; Plaxco, K. W.; Clark, P. L.; Sosnick, T. R. "Commonly Used FRET Fluorophores Promote Collapse of an Otherwise Disordered Protein". *Proceedings of the National Academy of Sciences* **2019**, *116*, 8889–8894.

(68) Dyson, H. J.; Wright, P. E. "NMR Illuminates Intrinsic Disorder". *Current Opinion in Structural Biology* **2021**, *70*, 44–52.

(69) Rule, G. S.; Hitchens, T. K., *Fundamentals of Protein NMR Spectroscopy*; Focus on Structural Biology v. 5; Springer: Dordrecht, 2006; 530 pp.

(70) Bermel, W.; Bertini, I.; Felli, I. C.; Gonnelli, L.; Koźmiński, W.; Piai, A.; Pierattelli, R.; Stanek, J. "Speeding up Sequence Specific Assignment of IDPs". *Journal of Biomolecular NMR* **2012**, *53*, 293–301.

(71) Hsu, S.-T. D.; Bertoncini, C. W.; Dobson, C. M. "Use of Protonless NMR Spectroscopy To Alleviate the Loss of Information Resulting from Exchange-Broadening". *Journal of the American Chemical Society* **2009**, *131*, 7222–7223.

(72) Tolman, J. R.; Al-Hashimi, H. M.; Kay, L. E.; Prestegard, J. H. "Structural and Dynamic Analysis of Residual Dipolar Coupling Data for Proteins". *Journal of the American Chemical Society* **2001**, *123*, 1416–1424.

(73) Karplus, M. "Contact Electron□Spin Coupling of Nuclear Magnetic Moments". *The Journal of Chemical Physics* **1959**, *30*, 11–15.

(74) Neuhaus, D. In *eMagRes*; American Cancer Society: 2011.

(75) Kachala, M.; Valentini, E.; Svergun, D. I. "Application of SAXS for the Structural Characterization of IDPs". *Advances in Experimental Medicine and Biology* **2015**, *870*, 261–289.

(76) Moult, J.; Pedersen, J. T.; Judson, R.; Fidelis, K. "A Large-Scale Experiment to Assess Protein Structure Prediction Methods". *Proteins: Structure, Function, and Bioinformatics* **1995**, *23*, ii–iv.

(77) Kleckner, I. R.; Foster, M. P. "An Introduction to NMR-based Approaches for Measuring Protein Dynamics". *Biochimica et biophysica acta* **2011**, *1814*, 942–968.

(78) Igumenova, T. I.; Frederick, K. K.; Wand, A. J. "Characterization of the Fast Dynamics of Protein Amino Acid Side Chains Using NMR Relaxation in Solution". *Chemical Reviews* **2006**, *106*, 1672–1699.

(79) Neudecker, P.; Lundström, P.; Kay, L. E. "Relaxation Dispersion NMR Spectroscopy as a Tool for Detailed Studies of Protein Folding". *Biophysical Journal* **2009**, *96*, 2045–2054.

(80) Olsson, S.; Noé, F. "Mechanistic Models of Chemical Exchange Induced Relaxation in Protein NMR". *Journal of the American Chemical Society* **2017**, *139*, 200–210.

(81) Meng, F.; Bellaiche, M. M.; Kim, J.-Y.; Zerze, G. H.; Best, R. B.; Chung, H. S. "Highly Disordered Amyloid-β Monomer Probed by Single-Molecule FRET and MD Simulation". *Biophysical Journal* **2018**, *114*, 870–884.

(82) Röllen, K.; Granzin, J.; Batra-Safferling, R.; Stadler, A. M. "Small-Angle X-ray Scattering Study of the Kinetics of Light-Dark Transition in a LOV Protein". *PLOS ONE* **2018**, *13*, e0200746.

(83) Josiah Willard Gibbs, *Elementary Principles in Statistical Mechanics: Developed with Especial Reference to the Rational Foundation of Thermodynamics*; C. Scribner's Sons: 1902; 239 pp.

(84) Zadeh, L. A. "Fuzzy Sets". *Information and Control* **1965**, *8*, 338–353.

(85) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. "Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids". *Journal of the American Chemical Society* **1996**, *118*, 11225–11236.

(86) 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: Structure, Function, and Bioinformatics* **2006**, *65*, 712–725.

(87) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. "A Smooth Particle Mesh Ewald Method". *The Journal of Chemical Physics* **1995**, *103*, 8577–8593.

(88) Yoshida, H. "Recent Progress in the Theory and Application of Symplectic Integrators". *Celestial Mechanics and Dynamical Astronomy* **1993**, *56*, 27–43.

(89) Hopkins, C. W.; Le Grand, S.; Walker, R. C.; Roitberg, A. E. "Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning". *Journal of Chemical Theory and Computation* **2015**, *11*, 1864–1874.

(90) Laio, A.; Parrinello, M. "Escaping Free-Energy Minima". *Proceedings of the National Academy of Sciences* **2002**, *99*, 12562–12566.

(91) Bussi, G.; Branduardi, D. In *Reviews in Computational Chemistry*, Parrill, A. L., Lipkowitz, K. B., Eds.; John Wiley & Sons, Inc: Hoboken, NJ, USA, 2015, pp 1–49.

(92) Pfaendtner, J.; Bonomi, M. "Efficient Sampling of High-Dimensional Free-Energy Landscapes with Parallel Bias Metadynamics". *Journal of Chemical Theory and Computation* **2015**, *11*, 5062–5067.

(93) *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*; Bowman, G. R., Pande, V. S., Noé, F., Eds.; Advances in Experimental Medicine and Biology, Vol. 797; Springer Netherlands: Dordrecht, 2014.

(94) Klus, S.; Nüske, F.; Koltai, P.; Wu, H.; Kevrekidis, I.; Schütte, C.; Noé, F. "Data-Driven Model Reduction and Transfer Operator Approximation". *Journal of Nonlinear Science* **2018**, *28*, 985–1010.

(95) Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schütte, C.; Noé, F. "Markov Models of Molecular Kinetics: Generation and Validation". *The Journal of Chemical Physics* **2011**, *134*, 174105.

(96) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. "Identification of Slow Molecular Order Parameters for Markov Model Construction". *The Journal of Chemical Physics* **2013**, *139*, 015102.

(97) Schwantes, C. R.; Pande, V. S. "Improvements in Markov State Model Construction Reveal Many Non-Native Interactions in the Folding of NTL9". *Journal of Chemical Theory and Computation* **2013**, *9*, 2000–2009.

(98) Cox, M. A. A.; Cox, T. F. In *Handbook of Data Visualization*, Chen, C.-h., Härdle, W., Unwin, A., Eds.; Springer Handbooks Comp.Statistics; Springer: Berlin, Heidelberg, 2008, pp 315–347.

(99) McInnes, L.; Healy, J.; Melville, J. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*, 2018.

(100) Arthur, D.; Vassilvitskii, S. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, New Orleans, Louisiana, Society for Industrial and Applied Mathematics: Philadelphia, PA, USA, 2007, pp 1027–1035.

(101) Defays, D. "An Efficient Algorithm for a Complete Link Method". *The Computer Journal* **1977**, *20*, 364–366.

(102) Rodriguez, A.; Laio, A. "Clustering by Fast Search and Find of Density Peaks". *Science* **2014**, *344*, 1492–1496.

(103) Mardt, A.; Pasquali, L.; Wu, H.; Noé, F. "VAMPnets for Deep Learning of Molecular Kinetics". *Nature Communications* **2018**, *9*, 5.

(104) Wu, H.; Noé, F. "Variational Approach for Learning Markov Processes from Time Series Data". *Journal of Nonlinear Science* **2020**, *30*, 23–66.

(105) Tu, J. H.; Rowley, C. W.; Luchtenburg, D. M.; Brunton, S. L.; Kutz, J. N. "On Dynamic Mode Decomposition: Theory and Applications". *Journal of Computational Dynamics* **2014**, *1*, 391–421.

(106) Mardt, A.; Pasquali, L.; Noé, F.; Wu, H. In ed. by Lu, J.; Ward, R., PMLR: Princeton University, Princeton, NJ, USA, 2020; Vol. 107, pp 451–475.

(107) Baker, C. M. "Polarizable Force Fields for Molecular Dynamics Simulations of Biomolecules". *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2015**, *5*, 241–254.

(108) Palazzesi, F.; Prakash, M. K.; Bonomi, M.; Barducci, A. "Accuracy of Current All-Atom Force-Fields in Modeling Protein Disordered States". *Journal of Chemical Theory and Computation* **2015**, *11*, 2–7.

(109) Robustelli, P.; Piana, S.; Shaw, D. E. "Developing a Molecular Dynamics Force Field for Both Folded and Disordered Protein States". *Proceedings of the National Academy of Sciences* **2018**, *115*, E4758–E4766.

(110) Unke, O. T.; Chmiela, S.; Sauceda, H. E.; Gastegger, M.; Poltavsky, I.; Schütt, K. T.; Tkatchenko, A.; Müller, K.-R. "Machine Learning Force Fields". *Chemical Reviews* **2021**, *121*, 10142–10186.

(111) Bonomi, M.; Camilloni, C.; Vendruscolo, M. "Metadynamic Metainference: Enhanced Sampling of the Metainference Ensemble Using Metadynamics". *Scientific Reports* **2016**, *6*, 31232.

(112) Bonomi, M.; Camilloni, C.; Cavalli, A.; Vendruscolo, M. "Metainference: A Bayesian Inference Method for Heterogeneous Systems". *Science Advances* **2016**, *2*, e1501177.

(113) Crehuet, R.; Buigues, P. J.; Salvatella, X.; Lindorff-Larsen, K. "Bayesian-Maximum-Entropy Reweighting of IDP Ensembles Based on NMR Chemical Shifts". *Entropy* **2019**, *21*, 898.

(114) Rangan, R.; Bonomi, M.; Heller, G. T.; Cesari, A.; Bussi, G.; Vendruscolo, M. "Determination of Structural Ensembles of Proteins: Restraining vs Reweighting". *Journal of Chemical Theory and Computation* **2018**, *14*, 6632–6641.

(115) Olsson, S.; Wu, H.; Paul, F.; Clementi, C.; Noé, F. "Combining Experimental and Simulation Data of Molecular Processes via Augmented Markov Models". *Proceedings of the National Academy of Sciences* **2017**, *114*, 8265–8270.

(116) Mardt, A.; Noé, F. *Progress in Deep Markov State Modeling: Coarse Graining and Experimental Data Restraints*, 2021.

(117) Zuckerman, D. M.; Chong, L. T. "Weighted Ensemble Simulation: Review of Methodology, Applications, and Software". *Annual Review of Biophysics* **2017**, *46*, 43–57.

(118) Zimmerman, M. I.; Bowman, G. R. "FAST Conformational Searches by Balancing Exploration/Exploitation Trade-Offs". *Journal of Chemical Theory and Computation* **2015**, *11*, 5747–5757.

(119) Boomsma, W.; Ferkinghoff-Borg, J.; Lindorff-Larsen, K. "Combining Experiments and Simulations Using the Maximum Entropy Principle". *PLoS Computational Biology* **2014**, *10*, ed. by Levitt, M., e1003406.

(120) Granata, D.; Baftizadeh, F.; Habchi, J.; Galvagnion, C.; De Simone, A.; Camilloni, C.; Laio, A.; Vendruscolo, M. "The Inverted Free Energy Landscape of an Intrinsically Disordered Peptide by Simulations and Experiments". *Scientific Reports* **2015**, *5*, 15449.

(121) Bowman, G. R.; Pande, V. S. "Protein Folded States Are Kinetic Hubs". *Proceedings of the National Academy of Sciences* **2010**, *107*, 10890–10895.

(122) Löhr, T.; Kohlhoff, K.; Heller, G. T.; Camilloni, C.; Vendruscolo, M. "A Kinetic Ensemble of the Alzheimer's Aβ Peptide". *Nature Computational Science* **2021**, *1*, 71–78.

(123) Babu, M. M.; van der Lee, R.; de Groot, N. S.; Gsponer, J. "Intrinsically Disordered Proteins: Regulation and Disease". *Current Opinion in Structural Biology* **2011**, *21*, 432–440.

(124) Sormanni, P. et al. "Simultaneous Quantification of Protein Order and Disorder". *Nature Chemical Biology* **2017**, *13*, 339–342.

(125) Bonomi, M.; Heller, G. T.; Camilloni, C.; Vendruscolo, M. "Principles of Protein Structural Ensemble Determination". *Current Opinion in Structural Biology* **2017**, *42*, 106–116.

(126) Fraser, J. S.; van den Bedem, H.; Samelson, A. J.; Lang, P. T.; Holton, J. M.; Echols, N.; Alber, T. "Accessing Protein Conformational Ensembles Using Room-Temperature X-ray Crystallography". *Proceedings of the National Academy of Sciences* **2011**, *108*, 16247–16252.

(127) Lindorff-Larsen, K.; Best, R. B.; DePristo, M. A.; Dobson, C. M.; Vendruscolo, M. "Simultaneous Determination of Protein Structure and Dynamics". *Nature* **2005**, *433*, 128–132.

(128) Van Kampen, N. G., *Stochastic Processes in Physics and Chemistry*; Elsevier: 2007.

(129) Chodera, J. D.; Noé, F. "Markov State Models of Biomolecular Conformational Dynamics". *Current Opinion in Structural Biology* **2014**, *25*, 135–144.

(130) Husic, B. E.; Pande, V. S. "Markov State Models: From an Art to a Science". *Journal of the American Chemical Society* **2018**, *140*, 2386–2396.

(131) Kohlhoff, K. J.; Shukla, D.; Lawrenz, M.; Bowman, G. R.; Konerding, D. E.; Belov, D.; Altman, R. B.; Pande, V. S. "Cloud-Based Simulations on Google Exacycle Reveal Ligand Modulation of GPCR Activation Pathways". *Nature Chemistry* **2014**, *6*, 15–21.

(132) Voelz, V. A.; Bowman, G. R.; Beauchamp, K.; Pande, V. S. "Molecular Simulation of Ab Initio Protein Folding for a Millisecond Folder NTL9(1–39)". *Journal of the American Chemical Society* **2010**, *132*, 1526–1528.

(133) Cummings, J.; Lee, G.; Ritter, A.; Sabbagh, M.; Zhong, K. "Alzheimer's Disease Drug Development Pipeline: 2019". *Alzheimer's & Dementia: Translational Research & Clinical Interventions* **2019**, *5*, 272–293.

(134) Jack, C. R. et al. "NIA-AA Research Framework: Toward a Biological Definition of Alzheimer's Disease". *Alzheimer's & Dementia* **2018**, *14*, 535–562.

(135) Cohen, S. I.; Vendruscolo, M.; Dobson, C. M.; Knowles, T. P. "From Macroscopic Measurements to Microscopic Mechanisms of Protein Aggregation". *Journal of Molecular Biology* **2012**, *421*, 160–171.

(136) Michaels, T. C. T.; Šarić, A.; Meisl, G.; Heller, G. T.; Curk, S.; Arosio, P.; Linse, S.; Dobson, C. M.; Vendruscolo, M.; Knowles, T. P. J. "Thermodynamic and Kinetic Design Principles for Amyloid-Aggregation Inhibitors". *Proceedings of the National Academy of Sciences* **2020**, *117*, 24251–24257.

(137) Lin, Y.-S.; Bowman, G. R.; Beauchamp, K. A.; Pande, V. S. "Investigating How Peptide Length and a Pathogenic Mutation Modify the Structural Ensemble of Amyloid Beta Monomer". *Biophysical Journal* **2012**, *102*, 315–324.

(138) Rosenman, D. J.; Connors, C. R.; Chen, W.; Wang, C.; García, A. E. "Aβ Monomers Transiently Sample Oligomer and Fibril-Like Configurations: Ensemble Characterization Using a Combined MD/NMR Approach". *Journal of Molecular Biology* **2013**, *425*, 3338–3359.

(139) Sgourakis, N. G.; Merced-Serrano, M.; Boutsidis, C.; Drineas, P.; Du, Z.; Wang, C.; Garcia, A. E. "Atomic-Level Characterization of the Ensemble of the Aβ(1–42) Monomer in Water Using Unbiased Molecular Dynamics Simulations and Spectral Algorithms". *Journal of Molecular Biology* **2011**, *405*, 570–583.

(140) Nasica-Labouze, J. et al. "Amyloid β Protein and Alzheimer's Disease: When Computer Simulations Complement Experimental Studies". *Chemical Reviews* **2015**, *115*, 3518–3563.

(141) Hellerstein, J. L.; Kohlhoff, K. J.; Konerding, D. E. "Science in the Cloud: Accelerating Discovery in the 21st Century". *IEEE Internet Computing* **2012**, *16*, 64–68.

(142) Rahman, M. U.; Rehman, A. U.; Liu, H.; Chen, H.-F. "Comparison and Evaluation of Force Fields for Intrinsically Disordered Proteins". *Journal of Chemical Information and Modeling* **2020**, DOI: 10.1021/acs.jcim.0c00762.

(143) McGibbon, R. T.; Pande, V. S. "Variational Cross-Validation of Slow Dynamical Modes in Molecular Kinetics". *The Journal of Chemical Physics* **2015**, *142*, 124105.

(144) Noé, F.; Wu, H.; Prinz, J.-H.; Plattner, N. "Projected and Hidden Markov Models for Calculating Kinetics and Metastable States of Complex Molecules". *The Journal of Chemical Physics* **2013**, *139*, 184114.

(145) Röblitz, S.; Weber, M. "Fuzzy Spectral Clustering by PCCA+: Application to Markov State Models and Data Classification". *Advances in Data Analysis and Classification* **2013**, *7*, 147–179.

(146) Klambauer, G.; Unterthiner, T.; Mayr, A.; Hochreiter, S. In *Advances in Neural Information Processing Systems 30*, Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: 2017, pp 971–980.

(147) Kohlhoff, K. J.; Robustelli, P.; Cavalli, A.; Salvatella, X.; Vendruscolo, M. "Fast and Accurate Predictions of Protein NMR Chemical Shifts from Interatomic Distances". *Journal of the American Chemical Society* **2009**, *131*, 13894–13895.

(148) Löhr, T.; Jussupow, A.; Camilloni, C. "Metadynamic Metainference: Convergence towards Force Field Independent Structural Ensembles of a Disordered Peptide". *The Journal of Chemical Physics* **2017**, *146*, 165102.

(149) Roche, J.; Shen, Y.; Lee, J. H.; Ying, J.; Bax, A. "Monomeric Aβ 1–40 and Aβ 1–42 Peptides in Solution Adopt Very Similar Ramachandran Map Distributions That Closely Resemble Random Coil". *Biochemistry* **2016**, *55*, 762–775.

(150) Konrat, R. "NMR Contributions to Structural Dynamics Studies of Intrinsically Disordered Proteins". *Journal of Magnetic Resonance* **2014**, *241*, 74–85.

(151) Levy, R. M.; Dai, W.; Deng, N.-J.; Makarov, D. E. "How Long Does It Take to Equilibrate the Unfolded State of a Protein?" *Protein Science* **2013**, *22*, 1459–1465.

(152) Dai, W.; Sengupta, A. M.; Levy, R. M. "First Passage Times, Lifetimes, and Relaxation Times of Unfolded Proteins". *Physical Review Letters* **2015**, *115*, 048101.

(153) Yan, Y.; McCallum, S. A.; Wang, C. "M35 Oxidation Induces Aβ40-like Structural and Dynamical Changes in Aβ42". *Journal of the American Chemical Society* **2008**, *130*, 5394–5395.

(154) Hou, L.; Shao, H.; Zhang, Y.; Li, H.; Menon, N. K.; Neuhaus, E. B.; Brewer, J. M.; Byeon, I.-J. L.; Ray, D. G.; Vitek, M. P.; Iwashita, T.; Makula, R. A.; Przybyla, A. B.; Zagorski, M. G. "Solution NMR Studies of the Aβ(1–40) and Aβ(1–42) Peptides Establish That the Met35 Oxidation State Affects the Mechanism of Amyloid Formation". *Journal of the American Chemical Society* **2004**, *126*, 1992–2005.

(155) Hou, L.; Kang, I.; Marchant, R. E.; Zagorski, M. G. "Methionine 35 Oxidation Reduces Fibril Assembly of the Amyloid Aβ-(1–42) Peptide of Alzheimer's Disease". *Journal of Biological Chemistry* **2002**, *277*, 40173–40176.

(156) Boomsma, W.; Frellsen, J. In *Advances in Neural Information Processing Systems 30*, Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: 2017, pp 3433–3443.

(157) Fout, A.; Byrd, J.; Shariat, B.; Ben-Hur, A. "Protein Interface Prediction Using Graph Convolutional Networks". 10.

(158) Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. "SchNet – A Deep Learning Architecture for Molecules and Materials". *The Journal of Chemical Physics* **2018**, *148*, 241722.

(159) Chodera, J. D.; Noé, F. "Probability Distributions of Molecular Observables Computed from Markov Models. II. Uncertainties in Observables and Their Time-Evolution". *The Journal of Chemical Physics* **2010**, *133*, 105102.

(160) Paul, A.; Samantray, S.; Anteghini, M.; Strodel, B. "Thermodynamics and Kinetics of the Amyloid-β Peptide Revealed by Markov State Models Based on MD Data in Agreement with Experiment". *bioRxiv* **2020**, 2020.07.27.223487.

(161) Dill, K. A.; Chan, H. S. "From Levinthal to Pathways to Funnels". *Nature Structural Biology* **1997**, *4*, 10–19.

(162) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. "GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers". *SoftwareX* **2015**, *1–2*, 19–25.

(163) Bussi, G.; Donadio, D.; Parrinello, M. "Canonical Sampling through Velocity-Rescaling". *The Journal of Chemical Physics* **2007**, *126*, 014101.

(164) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. "Molecular Dynamics with Coupling to an External Bath". *The Journal of Chemical Physics* **1984**, *81*, 3684–3690.

(165) MacKerell, A. D. et al. "All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins †". *The Journal of Physical Chemistry B* **1998**, *102*, 3586–3616.

(166) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. "Comparison of Simple Potential Functions for Simulating Liquid Water". *The Journal of Chemical Physics* **1983**, *79*, 926–935.

(167) Hess, B. "P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation". *Journal of Chemical Theory and Computation* **2008**, *4*, 116–122.

(168) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; Mackerell, A. D. "CHARMM General Force Field: A Force Field for Drug-like Molecules Compatible with the CHARMM All-Atom Additive Biological Force Fields". *Journal of Computational Chemistry* **2009**, NA–NA.

(169) Chollet, F. "Keras". **2015**.

(170) Abadi, M. et al. In 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16), 2016, pp 265–283.

(171) Head, T. et al. *Scikit-Optimize/Scikit-Optimize: V0.5Rc1*, Zenodo, 2018.

(172) Keller, B. G.; Kobitski, A.; Jäschke, A.; Nienhaus, G. U.; Noé, F. "Complex RNA Folding Kinetics Revealed by Single-Molecule FRET and Hidden Markov Models". *Journal of the American Chemical Society* **2014**, *136*, 4534–4543.

(173) PLUMED consortium "Promoting Transparency and Reproducibility in Enhanced Molecular Simulations". *Nature Methods* **2019**, *16*, 670–673.

(174) Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. "PLUMED 2: New Feathers for an Old Bird". *Computer Physics Communications* **2014**, *185*, 604–613.

(175) Noé, F.; Clementi, C. "Kinetic Distance and Kinetic Maps from Molecular Dynamics Simulation". *Journal of Chemical Theory and Computation* **2015**, *11*, 5002–5011.

(176) Beauchamp, K. A.; Bowman, G. R.; Lane, T. J.; Maibaum, L.; Haque, I. S.; Pande, V. S. "MSMBuilder2: Modeling Conformational Dynamics on the Picosecond to Millisecond Scale". *Journal of Chemical Theory and Computation* **2011**, *7*, 3412–3419.

(177) Trendelkamp-Schroer, B.; Wu, H.; Paul, F.; Noé, F. "Estimation and Uncertainty of Reversible Markov Models". *The Journal of Chemical Physics* **2015**, *143*, 174101.

(178) Cummings, J.; Lee, G.; Zhong, K.; Fonseca, J.; Taghva, K. "Alzheimer's Disease Drug Development Pipeline: 2021". *Alzheimer's & Dementia: Translational Research & Clinical Interventions* **2021**, *7*, e12179.

(179) Glenner, G. G.; Wong, C. W. "Alzheimer's Disease and Down's Syndrome: Sharing of a Unique Cerebrovascular Amyloid Fibril Protein". *Biochemical and Biophysical Research Communications* **1984**, *122*, 1131–1135.

(180) Ball, K. A.; Phillips, A. H.; Nerenberg, P. S.; Fawzi, N. L.; Wemmer, D. E.; Head-Gordon, T. "Homogeneous and Heterogeneous Tertiary Structure Ensembles of Amyloid-β Peptides". *Biochemistry* **2011**, *50*, 7612–7628.

(181) Paul, A.; Samantray, S.; Anteghini, M.; Khaled, M.; Strodel, B. "Thermodynamics and Kinetics of the Amyloid-β Peptide Revealed by Markov State Models Based on MD Data in Agreement with Experiment". *Chemical Science* **2021**, *12*, 6652–6669.

(182) Habchi, J.; Arosio, P.; Perni, M.; Costa, A. R.; Yagi-Utsumi, M.; Joshi, P.; Chia, S.; Cohen, S. I. A.; Müller, M. B. D.; Linse, S.; Nollen, E. A. A.; Dobson, C. M.; Knowles, T. P. J.; Vendruscolo, M. "An Anticancer Drug Suppresses the Primary Nucleation Reaction That Initiates the Production of the Toxic Aβ42 Aggregates Linked with Alzheimer's Disease". *Science Advances* **2016**, *2*, e1501244.

(183) Lieblein, T.; Zangl, R.; Martin, J.; Hoffmann, J.; Hutchison, M. J.; Stark, T.; Stirnal, E.; Schrader, T.; Schwalbe, H.; Morgner, N. "Structural Rearrangement of Amyloid-β upon Inhibitor Binding Suppresses Formation of Alzheimer's Disease Related Oligomers". *eLife* **2020**, *9*, e59306.

(184) Robustelli, P.; Ibanez-de-Opakua, A.; Campbell-Bezat, C.; Giordanetto, F.; Becker, S.; Zweckstetter, M.; Pan, A. C.; Shaw, D. E. "Molecular Basis of Small-Molecule Binding to α-Synuclein". **2021**, 2021.01.22.426549.

(185) Rabiner, L. R.; Juang, B. H. "An Introduction to Hidden Markov Models". *IEEE ASSp Magazine* **1986**, 13.

(186) Suárez, E.; Wiewiora, R. P.; Wehmeyer, C.; Noé, F.; Chodera, J. D.; Zuckerman, D. M. "What Markov State Models Can and Cannot Do: Correlation versus Path-Based Observables in Protein-Folding Models". *Journal of Chemical Theory and Computation* **2021**, *17*, 3119–3133.

(187) Heller, G. T.; Bonomi, M.; Vendruscolo, M. "Structural Ensemble Modulation upon Small-Molecule Binding to Disordered Proteins". *Journal of Molecular Biology* **2018**, *430*, 2288–2292.

(188) Lee, H.; Mok, K. H.; Muhandiram, R.; Park, K.-H.; Suk, J.-E.; Kim, D.-H.; Chang, J.; Sung, Y. C.; Choi, K. Y.; Han, K.-H. "Local Structural Elements in the Mostly Unstructured Transcriptional Activation Domain of Human P53 *". *Journal of Biological Chemistry* **2000**, *275*, 29426–29432.

(189) Vernon, R. M.; Chong, P. A.; Tsang, B.; Kim, T. H.; Bah, A.; Farber, P.; Lin, H.; Forman-Kay, J. D. "Pi-Pi Contacts Are an Overlooked Protein Feature Relevant to Phase Separation". *eLife* **2018**, *7*, ed. by Shan, Y., e31486.

(190) Paton, R. S.; Goodman, J. M. "Hydrogen Bonding and π-Stacking: How Reliable Are Force Fields? A Critical Evaluation of Force Field Descriptions of Nonbonded Interactions". *Journal of Chemical Information and Modeling* **2009**, *49*, 944–955.

(191) Mayne, C. G.; Saam, J.; Schulten, K.; Tajkhorshid, E.; Gumbart, J. C. "Rapid Parameterization of Small Molecules Using the Force Field Toolkit". *Journal of Computational Chemistry* **2013**, *34*, 2757–2770.

(192) Frisch, M. J. et al. *Gaussian 16 Revision A.03*, 2016.

(193) Kabsch, W.; Sander, C. "Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features". *Biopolymers* **1983**, *22*, 2577–2637.

(194) Plosker, G. L.; Figgitt, D. P. "Rituximab". *Drugs* **2003**, *63*, 803–843.

(195) Souriau, C.; Hudson, P. J. "Recombinant Antibodies for Cancer Diagnosis and Therapy". *Expert Opinion on Biological Therapy* **2003**, *3*, 305–318.

(196) Bang, L. M.; Keating, G. M. "Adalimumab". *BioDrugs* **2004**, *18*, 121–139.

(197) Andryukov, B. G. "Six Decades of Lateral Flow Immunoassay: From Determining Metabolic Markers to Diagnosing COVID-19". *AIMS Microbiology* **2020**, *6*, 280–304.

(198) Lonberg, N. et al. "Antigen-Specific Human Antibodies from Mice Comprising Four Distinct Genetic Modifications". *Nature* **1994**, *368*, 856–859.

(199) Jakobovits, A.; Amado, R. G.; Yang, X.; Roskos, L.; Schwab, G. "From XenoMouse Technology to Panitumumab, the First Fully Human Antibody Product from Transgenic Mice". *Nature Biotechnology* **2007**, *25*, 1134–1143.

(200) Winter, G.; Griffiths, A. D.; Hawkins, R. E.; Hoogenboom, H. R. "Making Antibodies by Phage Display Technology". 23.

(201) Sormanni, P.; A. Aprile, F.; Vendruscolo, M. "Third Generation Antibody Discovery Methods: In Silico Rational Design". *Chemical Society Reviews* **2018**, *47*, 9137–9157.

(202) Sormanni, P.; Aprile, F. A.; Vendruscolo, M. "Rational Design of Antibodies Targeting Specific Epitopes within Intrinsically Disordered Proteins". *Proceedings of the National Academy of Sciences* **2015**, *112*, 9902–9907.

(203) Aprile, F. A.; Sormanni, P.; Perni, M.; Arosio, P.; Linse, S.; Knowles, T. P. J.; Dobson, C. M.; Vendruscolo, M. "Selective Targeting of Primary and Secondary Nucleation Pathways in Aβ42 Aggregation Using a Rational Antibody Scanning Method". *Science Advances* **2017**, *3*, e1700488.

(204) Rangel, M. A.; Bedwell, A.; Costanzi, E.; Ricagno, S.; Frydman, J.; Vendruscolo, M.; Sormanni, P. "Fragment-Based Computational Design of Antibodies Targeting Structured Epitopes". **2021**, 2021.03.02.433360.

(205) Daura, X.; Gademann, K.; Jaun, B.; Seebach, D.; van Gunsteren, W. F.; Mark, A. E. "Peptide Folding: When Simulation Meets Experiment". *Angewandte Chemie International Edition* **1999**, *38*, 236–240.

(206) Jeliazkov, J. R.; Sljoka, A.; Kuroda, D.; Tsuchimura, N.; Katoh, N.; Tsumoto, K.; Gray, J. J. "Repertoire Analysis of Antibody CDR-H3 Loops Suggests Affinity Maturation Does Not Typically Result in Rigidification". *Frontiers in Immunology* **2018**, *9*, 413.

(207) Ovchinnikov, V.; Louveau, J. E.; Barton, J. P.; Karplus, M.; Chakraborty, A. K. "Role of Framework Mutations and Antibody Flexibility in the Evolution of Broadly Neutralizing Antibodies". *eLife* **2018**, *7*, ed. by Walczak, A. M., e33038.

(208) Bhat, T. N.; Bentley, G. A.; Boulot, G.; Greene, M. I.; Tello, D.; Dall'Acqua, W.; Souchon, H.; Schwarz, F. P.; Mariuzza, R. A.; Poljak, R. J. "Bound Water Molecules and Conformational Stabilization Help Mediate an Antigen-Antibody Association." *Proceedings of the National Academy of Sciences* **1994**, *91*, 1089–1093.

(209) Shiroishi, M.; Yokota, A.; Tsumoto, K.; Kondo, H.; Nishimiya, Y.; Horii, K.; Matsushima, M.; Ogasahara, K.; Yutani, K.; Kumagai, I. "Structural Evidence for Entropic Contribution of Salt Bridge Formation to a Protein Antigen-Antibody Interaction: THE CASE OF HEN LYSOZYME-HyHEL-10 Fv COMPLEX *". *Journal of Biological Chemistry* **2001**, *276*, 23042–23050.

(210) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L.; Grubmüller, H.; MacKerell, A. D. "CHARMM36m: An Improved Force Field for Folded and Intrinsically Disordered Proteins". *Nature Methods* **2017**, *14*, 71–73.

(211) Webb, B.; Sali, A. "Comparative Protein Structure Modeling Using MODELLER". *Current Protocols in Bioinformatics* **2016**, *54*, 5.6.1–5.6.37.

(212) Parrinello, M.; Rahman, A. "Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method". *Journal of Applied Physics* **1981**, *52*, 7182–7190.

(213) Barducci, A.; Bussi, G.; Parrinello, M. "Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method". *Physical Review Letters* **2008**, *100*, 020603.

(214) Raiteri, P.; Laio, A.; Gervasio, F. L.; Micheletti, C.; Parrinello, M. "Efficient Reconstruction of Complex Free Energy Landscapes by Multiple Walkers Metadynamics". *The Journal of Physical Chemistry B* **2006**, *110*, 3533–3539.

(215) Torrie, G.; Valleau, J. "Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling". *Journal of Computational Physics* **1977**, *23*, 187–199.

(216) Bonomi, M.; Pellarin, R.; Vendruscolo, M. "Simultaneous Determination of Protein Structure and Dynamics Using Cryo-Electron Microscopy". *Biophysical Journal* **2018**, *114*, 1604–1613.

(217) Camilloni, C.; Cavalli, A.; Vendruscolo, M. "Assessment of the Use of NMR Chemical Shifts as Replica-Averaged Structural Restraints in Molecular Dynamics Simulations to Characterize the Dynamics of Proteins". *The Journal of Physical Chemistry B* **2013**, *117*, 1838–1843.

(218) Camilloni, C.; Vendruscolo, M. "Statistical Mechanics of the Denatured State of a Protein Using Replica-Averaged Metadynamics". *Journal of the American Chemical Society* **2014**, *136*, 8982–8991.

(219) Zhang, S.; Tong, H.; Xu, J.; Maciejewski, R. "Graph Convolutional Networks: A Comprehensive Review". *Computational Social Networks* **2019**, *6*, 11.

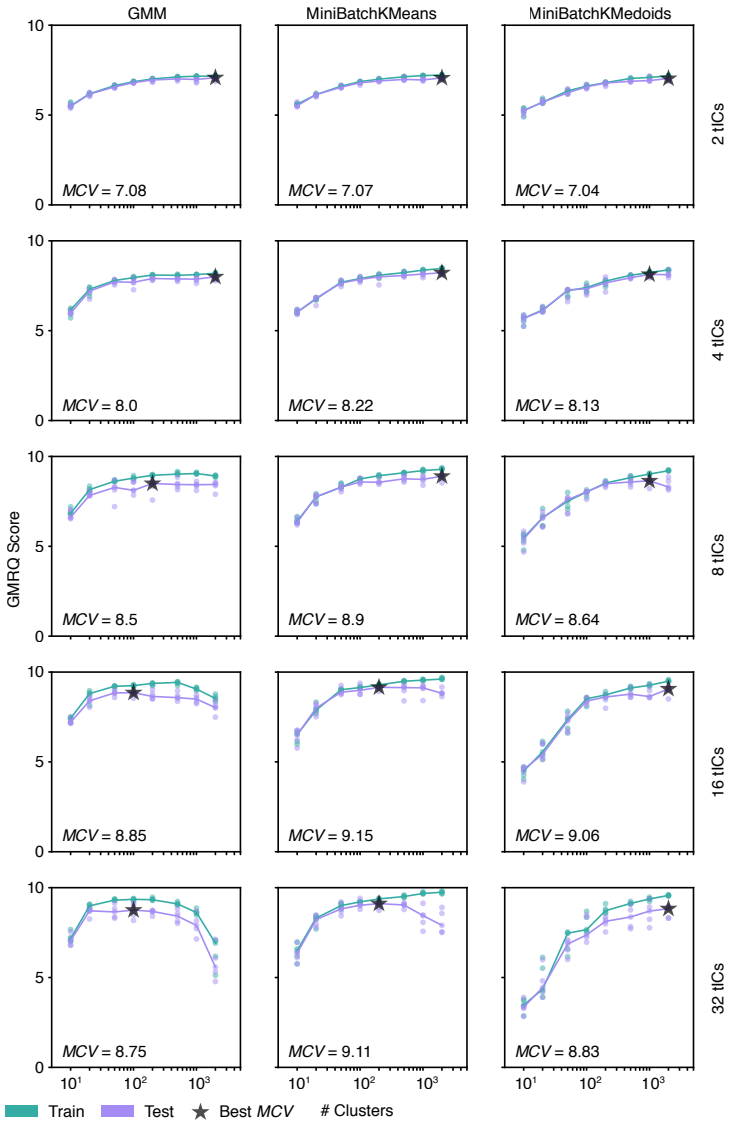
(220) Musil, F.; Grisafi, A.; Bartók, A. P.; Ortner, C.; Csányi, G.; Ceriotti, M. "Physics-Inspired Structural Representations for Molecules and Materials". *Chemical Reviews* **2021**, *121*, 9759–9815.

(221) Iconaru, L. I.; Ban, D.; Bharatham, K.; Ramanathan, A.; Zhang, W.; Shelat, A. A.; Zuo, J.; Kriwacki, R. W. "Discovery of Small Molecules That Inhibit the Disordered Protein, p27Kip1". *Scientific Reports* **2015**, *5*, 15686.

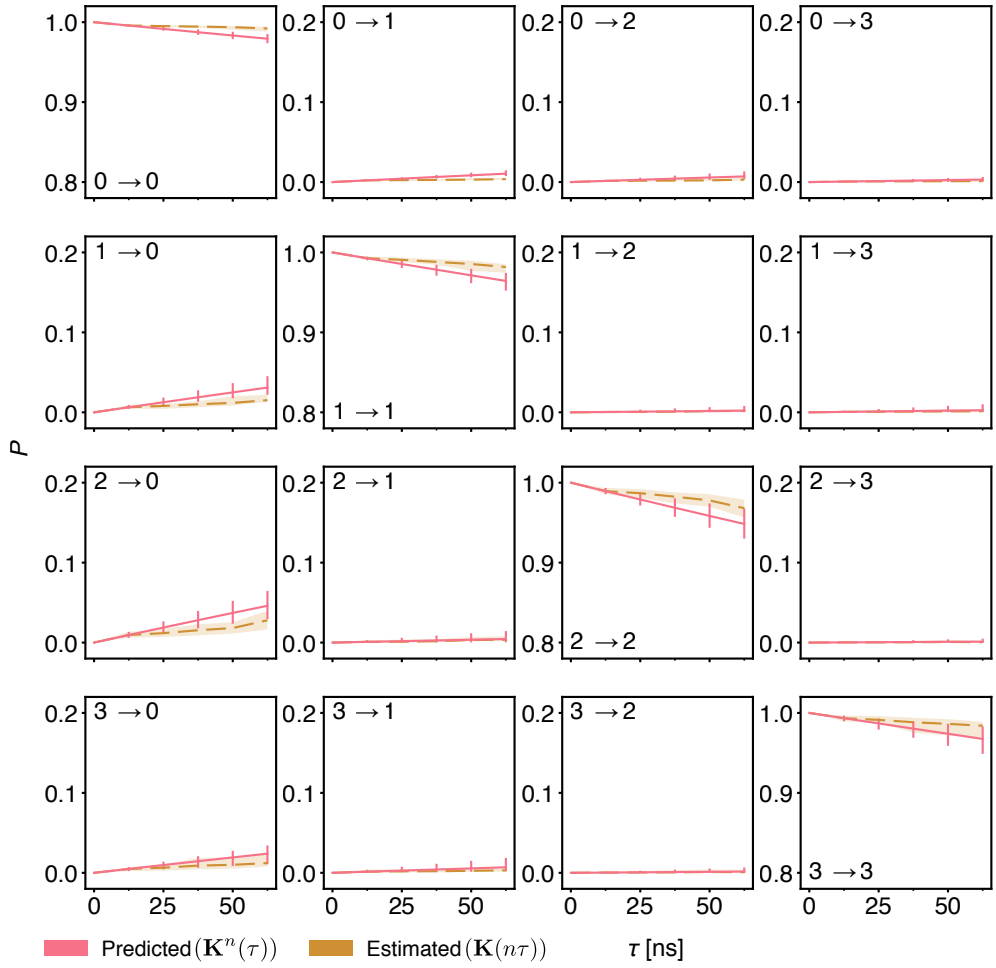
(222) Tóth, G. et al. "Targeting the Intrinsically Disordered Structural Ensemble of α-Synuclein by Small Molecules as a Potential Therapeutic Strategy for Parkinson's Disease". *PLoS ONE* **2014**, *9*, ed. by Cookson, M. R., e87133.

7 Appendix

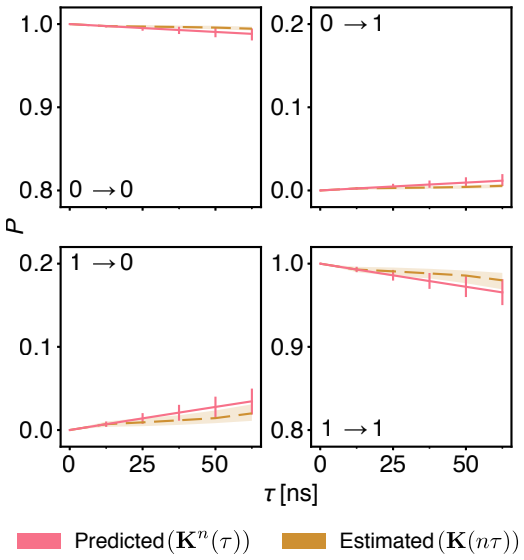
7.1 Supporting information for chapter 2



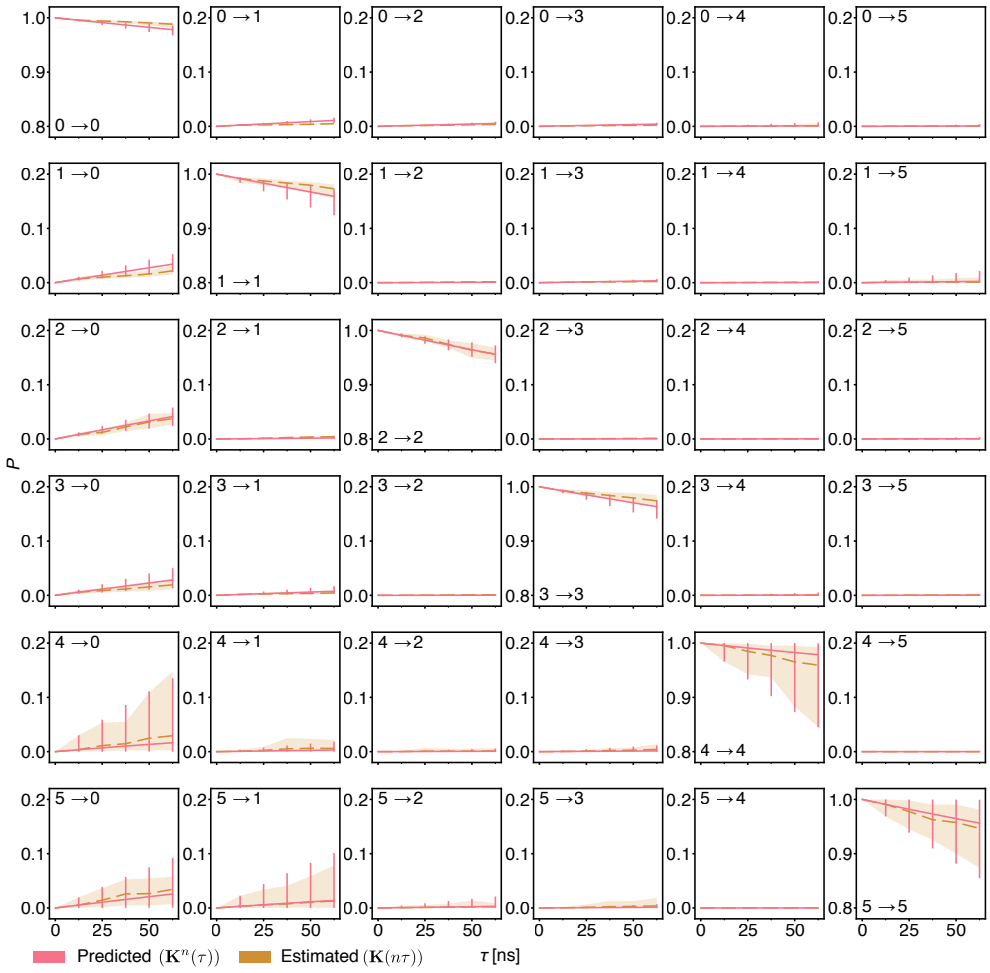
Supplementary Figure 7.1: Hyperparameter scan for a conventional discrete-state Markov state model. Clustering algorithms evaluated include Gaussian mixture models, minibatch k-Means, and minibatch k-Medoids using between 10 and 2000 microstates and between 2 and 32 input dimensions. The mean cross-validation score (MCV) is shown in the bottom left.



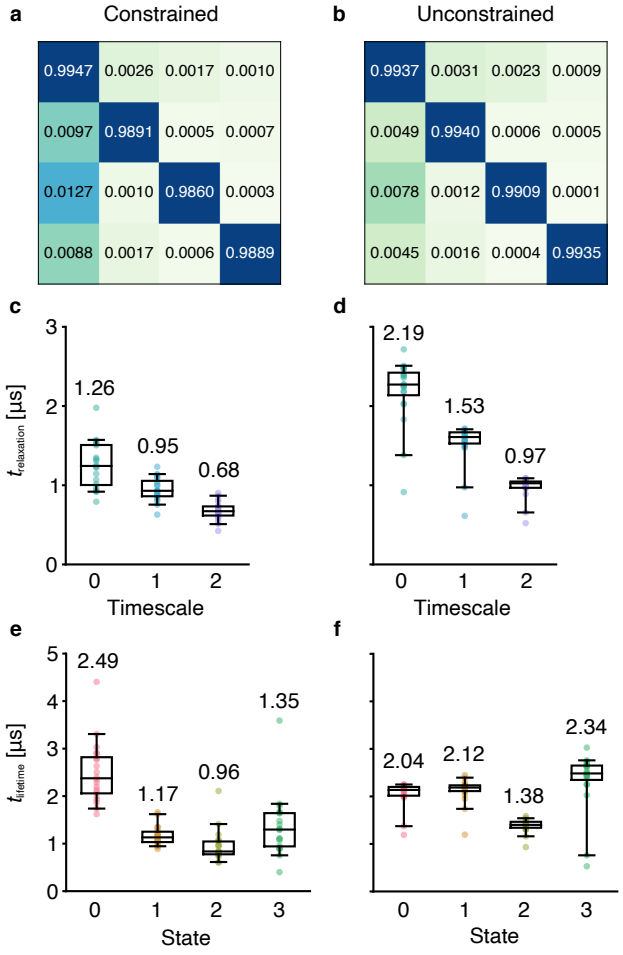
Supplementary Figure 7.2: Chapman-Kolmogorov test for the 4-state model. Each panel indicates the transition probability for one matrix entry for successive applications and estimations of the Koopman matrix. Shaded areas and error bars indicate 95th percentiles of the mean over all 20 models.



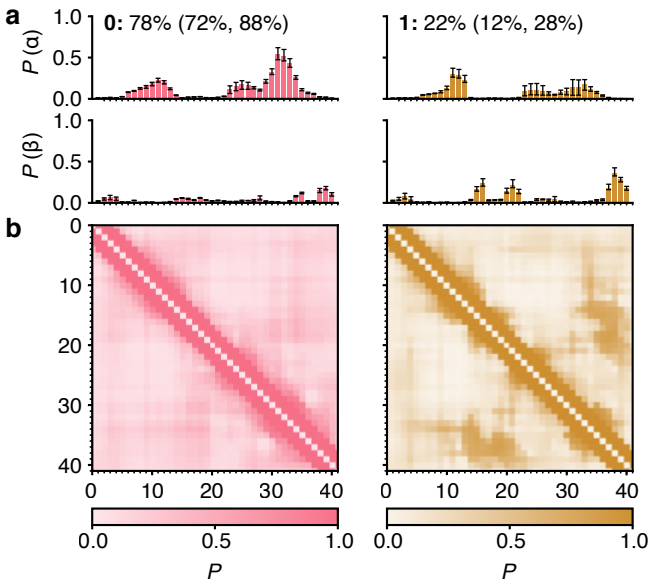
Supplementary Figure 7.3: Chapman-Kolmogorov test for the 2-state model. Each panel indicates the transition probability for one matrix entry for successive applications and estimations of the Koopman matrix. Shaded areas and error bars indicate 95th percentiles of the mean over all 20 models.



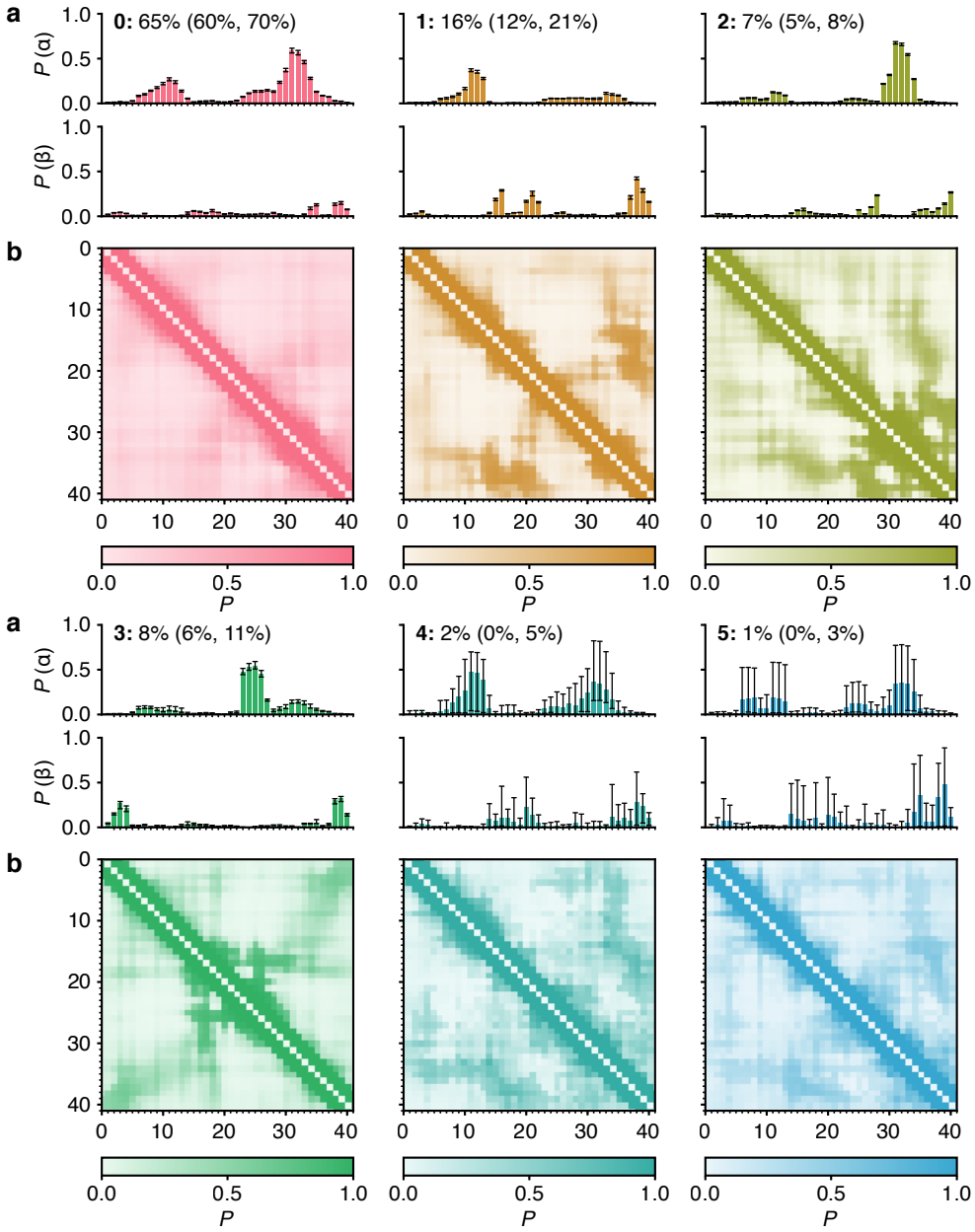
Supplementary Figure 7.4: Chapman-Kolmogorov test for the 6-state model. Each panel indicates the transition probability for one matrix entry for successive applications and estimations of the Koopman matrix. Shaded areas and error bars indicate 95th percentiles of the mean over all 20 models.



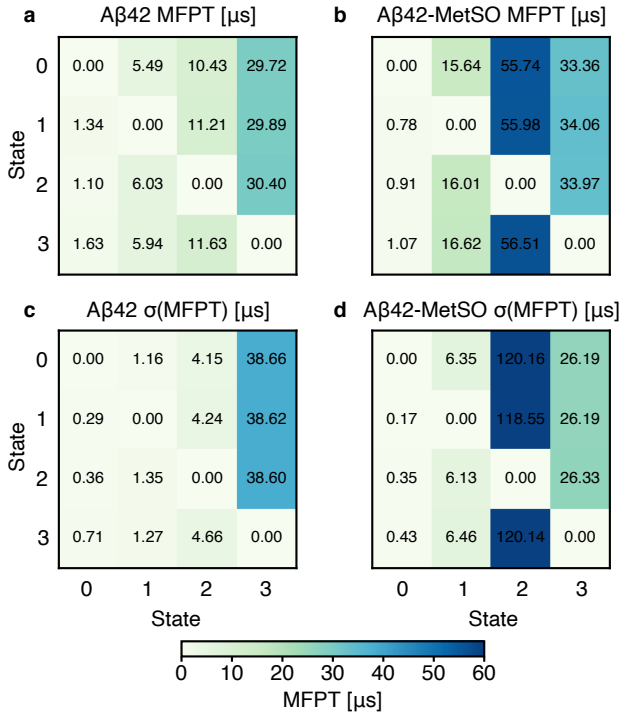
Supplementary Figure 7.5: Comparison of the 4-state model constrained to feature reversibility and positive transition matrix elements and the unconstrained 4-state model. **(a- b)** Mean transition matrix elements (state transition probabilities). **(c-d)** Relaxation timescales of the constrained and unconstrained 4-state models. **(e-f)** State lifetimes for the constrained and unconstrained models. The whiskers, boxes and horizontal lines indicate 95th percentiles, quartiles, and the median values over all 20 models, respectively, the labels show the mean model values.



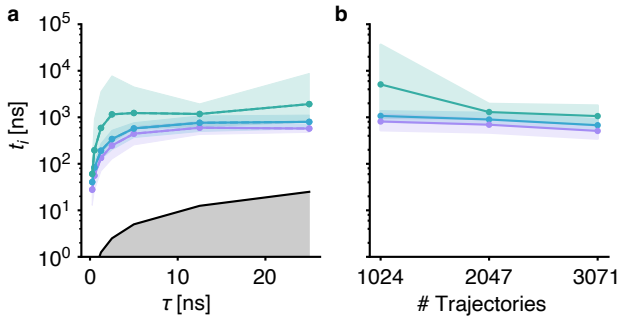
Supplementary Figure 7.6: Structural properties of Aβ42 in the two-state model. **a** Populations of a-helical and b-sheet content per residue as calculated using DSSP[193]. The equilibrium percentage of each state is given above, with the 95th percentile in parentheses. **b** Contact probability maps with a cut-off of 0.8 nm. Error bars indicate 95th percentiles of the bootstrap sample of the mean over all 20 models.



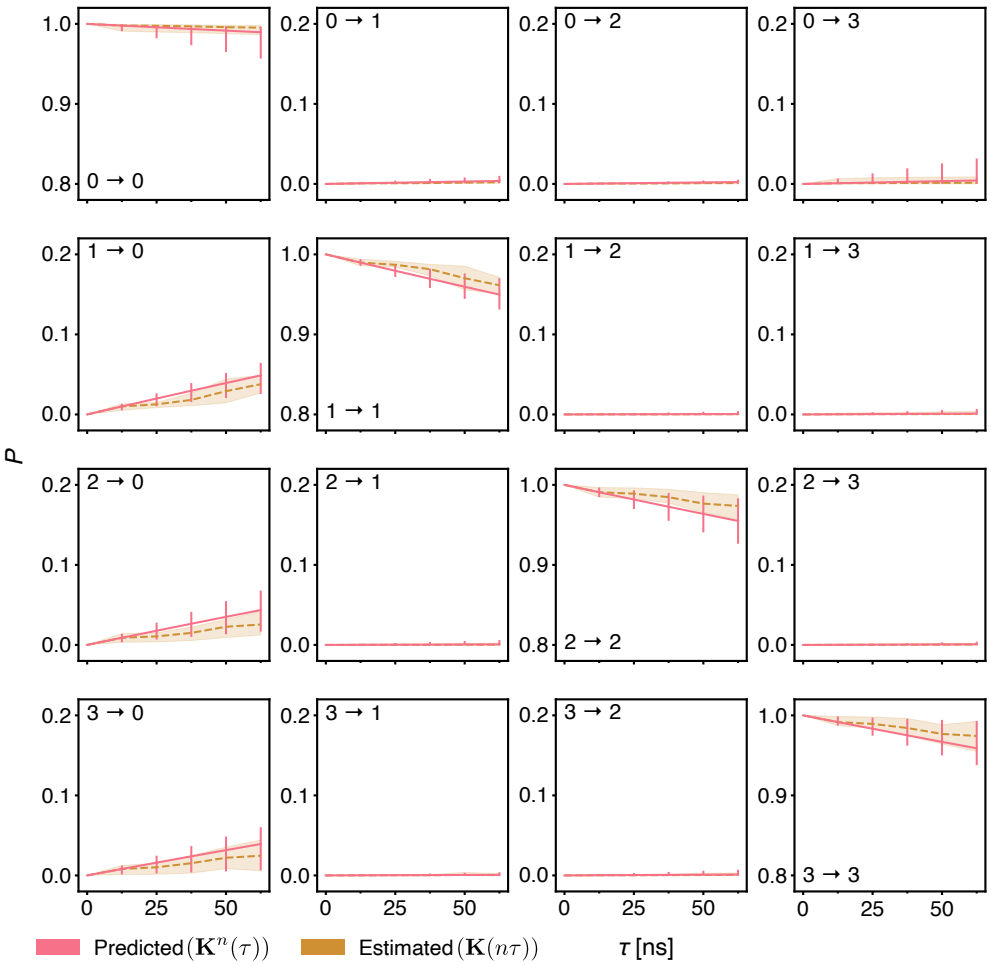
Supplementary Figure 7.7: Structural properties of Aβ42 in the six-state model. **a** Populations of a-helical and b-sheet content per residue as calculated using DSSP[193]. The equilibrium percentage of each state is given above, with the 95th percentile in parentheses. **b** Contact probability maps with a cut-off of 0.8 nm. Error bars indicate 95th percentiles of the bootstrap sample of the mean over all 20 models.



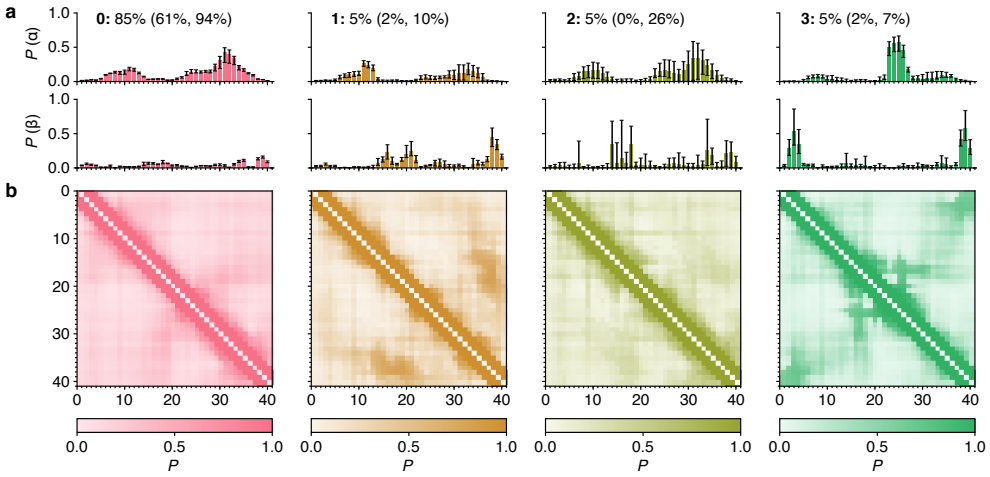
Supplementary Figure 7.8: Full mean first-passage times for **(a)** Aβ42 and **(b)** Aβ42-MetSO in μs. **c-d** Standard deviations for the mean first-passage times of both models in μs, across all 20 models.



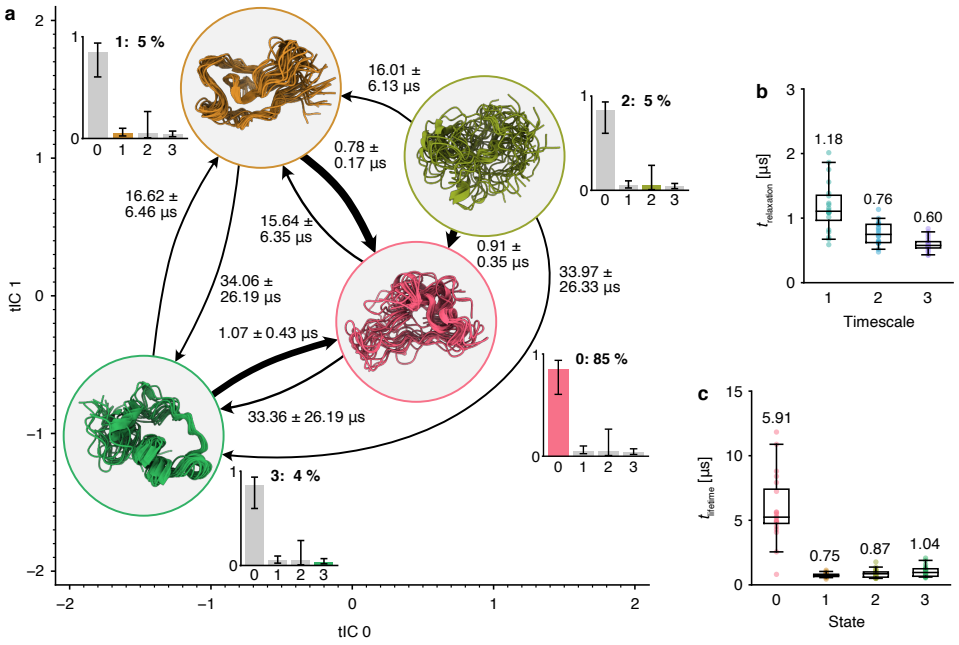
Supplementary Figure 7.9: Computational validation of the Aβ42-MetSO kinetic ensemble. **a** Dependence of the three longest relaxation timescales (green, cyan and purple, respectively) on the lag time τ. The grey shading indicates the timescales for which the Koopman model can no longer resolve the relaxation timescales. **b** Dependence of the relaxation timescales on the number of trajectories used to build the kinetic ensemble as a 4-state model. Shaded areas indicate 95th percentiles of the bootstrap sample of the mean over all 20 models.



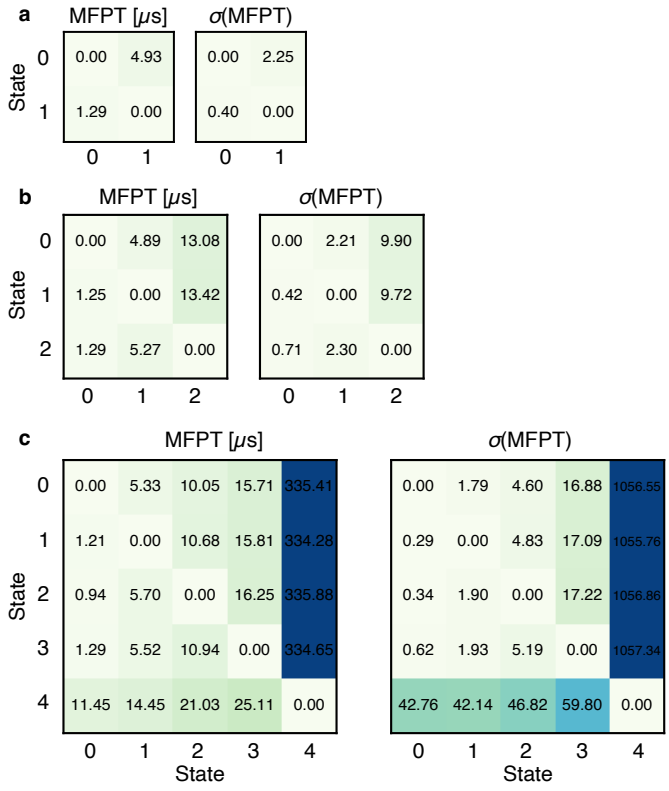
Supplementary Figure 7.10: Chapman-Kolmogorov test for the 4-state model of Aβ42-MetSO. Each panel indicates the transition probability for one matrix entry for successive applications and estimations of the Koopman matrix. Shaded areas and error bars indicate 95th percentiles of the mean over all 20 models.



Supplementary Figure 7.11: Structural properties of Aβ42-MetSO in the four-state model. **a** Populations of a-helical and β-sheet content per residue as calculated using DSSP1. The equilibrium percentage of each state is given above, with the 95th percentile in parentheses. **b** Contact probability maps with a cut-off of 0.8 nm. Error bars indicate 95th percentiles of the bootstrap sample of the mean over all 20 models.



Supplementary Figure 7.12: Populations and mean first-passage times in the kinetic ensemble of Aβ42-MetSO. **a** Mean first-passage times and their standard deviations between states in the kinetic ensemble; thicker arrows correspond to faster transitions. The state location is projected on to the two slowest time-independent coordinates (tICs) and the structures shown are 20 high-weight conformations from all models aligned on the most prominent secondary structure motifs (see Supplementary Figure 7.11 a). Transitions with mean first-passage times slower than 40 µs are not shown (Supplementary Figure 7.8 b, d). **b** Slowest relaxation timescales of the 4-state model. **c** Mean lifetime of each state in the 4-state model. The whiskers, boxes and horizontal lines indicate 95th percentiles, quartiles, and the median values over all 20 models, respectively, the labels show the mean model values.



Supplementary Figure 7.13: Full mean first-passage times for Aβ42 with **(a)** 2 states, **(b)** 3 states and **(c)** 5 states in μs. Standard deviations are shown for the mean first-passage times in μs, across all 20 models.